1. Úvod

2. Morfologie

2.1 Úvod do morfologie
Jednou z velmi důležitých oblastí počítačového zpracování přirozeného jazyka je {\newdef morfologie}. Tímto názvem označujeme nauku o tvoření tvarů a o jejich významech. Hlavním předmětem zkoumání je v morfologii "ohýbání" slov, tj. jejich skloňování a časování.
Název je odvozen od termínu **morfém**, který označuje nejmenší jednotku jazyka, která je nositelem významu. Tato jednotka se obvykle skládá z jednoho nebo více znaků. Morfémy dělíme na **lexikální** (nesoucí význam slova jako takového) a **gramatické** (určující gramatickou roli příslušného slovního tvaru). Ve tvaru *zahradou* je například přítomen lexikální morfém *zahrad-* (který by mohl být dále analyzován na předponový (prefixální) morfém *za-* a základový morfém *-hrad-*) a gramatický morfém *-ou*, který je souborem tří elementárních jednotek (sémat) pádu, čísla a rodu, v našem případě jde po řadě o 7.pád, jednotné číslo (singulár) a ženský rod (femininum). Všechny tyto tři významy (sémata) jsou spojeny v jediném gramatickém morfému.
Formální morfologie studuje typy skloňování (deklinace) a časování (konjugace) se všemi pravidelnostmi a výjimkami, které v jazyce existují. Čeština, podobně jako některé další slovanské jazyky, je známa složitostí své morfologie. Ta je komplikována jak vysokým stupněm víceznačnosti gramatických morfémů tak i množstvím tvaroslovných dublet (stejných slovních tvarů odvozených od různých slovních základů). Víceznačnost gramatických morfémů můžeme ilustrovat například pomocí slovního tvaru *ženy*, kde gramatický morfém *-y* má funkci čtyř morfémů, a sice 2.pád čísla jednotného rodu ženského, 1.pád množného čísla rodu ženského, 4.pád množného čísla rodu ženského a 5.pád množného čísla rodu ženského. Nejvíce víceznačným morfémem v češtině je gramatický morfém *-í* u měkkých přídavných jmen (podle vzoru *jarní*), který má celkem funkci 27 různých morfémů.
Tvaroslovnými dublety jsou například slovní tvary *ženu* (tvar dvou lexikálních morfémů – *žena* a *hnát*), *tři* (číslovka a tvar slovesa *třít*), *už* (příslovce a rozkazovací způsob slovesa *úžit*) apod. Ke složitosti české morfologie na druhé straně nemalou měrou přispívá to, že týž gramatický morfém u téhož lexikálního morfému může mít několik podob (srov. např. *páni/pánové*, *na hradě/o hradu*, *o balíčcích/o balíčkách*). Vedle toho dochází často k (pravidelným) změnám lexikálního morfému pomocí jeho **alternací** (změn hlásky uvnitř kmene), které pak vytvářejí soubor jeho **alomorfů** (variant kmene odvozených od stejného základu). Jedná se např. o tvary *matka/matce/mat(e)k*, popř. i *matčin*, pokud toto slovo nepokládáme již za jinou lexikální jednotku.
I přes uvedené problémy však můžeme oblast morfologie považovat z hlediska počítačového zpracování přirozeného jazyka za oblast relativně dobře popsanou a zpracovanou nejen na teoretické úrovni, ale i v řadě aplikací. Některé zajímavé aplikace založené na uplatnění znalostí o (české) morfologii představujeme v následujících kapitolách.

2.2 Základní pojmy a notace

2.3 Plná morfologie

2.4 Lematizace

2.5 Algoritmy morfologické analýzy

3. Aplikace morfologie

3.1 Úvod
Průvodním jevem pokroku v oblasti vývoje počítačů je vzrůstající množství textů, které jsou k dispozici v elektronické podobě, ať již na Internetu nebo lokálně na disku či na CD přímo v uživatelově počítači. Snaha umožnit jednoduché, rychlé a spolehlivé vyhledávání v elektronických textech je proto zcela přirozená.

Názor na použití lingvisticky motivovaných vyhledávacích metod se v praxi velmi liší v závislosti na jazyce, ve kterém jsou texty napsány. Na první pohled je zřejmé, že pro jazyky, které jenom minimálně používají skloňování a časování (flexi), není třeba uplatňovat komplikované metody vyhledávání jednotlivých tvarů slov. Ve většině případů totiž při vyhledávání vystačíme se standardními zástupnými znaky, otazníkem (zastupujícím jakýkoli jednotlivý znak) a hvězdičkou (zastupující libovolnou posloupnost znaků neobsahující mezeru).
Chceme-li například v angličtině vyhledat všechny výskyty jednotného a množného čísla podstatného jména *table* (*tabulka, stůl*), můžeme tento požadavek snadno zapsat ve formě řetězce "table?". Vyhledávací program potom snadno nalezne oba přípustné tvary *table* i *tables*. Kromě nich ovšem navíc nalezne i slovní tvar *tablet*, který také vyhovuje uvedenému vzorku. Tuto nesnáz ovšem je možné snadno odstranit malým rozšířením syntaxe pro zápis vzorku pro vyhledávání - zapíšeme-li požadovaný vzorek např. v jazyce FLEX, můžeme mnohem přesněji zadat, jaké tvary chceme vyhledávat. V našem příkladu bychom dosáhli požadovaného výsledku pomocí řetězce "table(s)?", kde má znak ?poněkud jinou funkci než v předchozím případu. Znamená, že hledaný vzorek smí být ukončen maximálně výskytem písmene *s* na konci hledaného slova. Tento zápis v podstatě připouští pouze dva tvary hledaného slova, a to *table* a *tables*, což je přesně to, co uživatel potřebuje.
V jazycích s bohatším skloňováním, než má angličtina, je řešení mnohem složitější. Prvním problémem je velikost množiny potřebných koncových skupin hlásek. Pro češtinu bychom například museli do vzorku pro všechny tvary jednotného a množného čísla slova *tabulka* zadat:
"tabul(ka|ky|ce|ku|ko|kou|ek|kám|kách|kami)",
což již pro běžného uživatele může znamenat značný problém. Nebezpečí vynechání správného nebo naopak uvedení nesprávného zakončení je příliš velké, nehledě na to, že samotné zapsání vzorku bude trvat relativně dlouhou dobu.
Druhým problémem je to, že zdaleka ne všechny změny slovních tvarů při skloňování nebo časování se dají vyjádřit pouze pomocí předpon, přípon a koncovek. V češtině (podobně jako v jiných jazycích) existuje celá řada slov, jejichž kmen se při ohýbání mění - např. *stůl/stolu*,*matka/matce/matek*, *švec/ševce* apod. Je zřejmé, že i v tomto případě bychom mohli problém vyřešit podobně, tedy pomocí uvedení kompletního seznamu hledaných slovních tvarů, nároky na schopnosti a čas uživatele by však byly ještě větší než v předchozím případě. Z tohoto důvodu je výhodné doplnit vyhledávací modul o modul jazykový, který má za úkol uživatelem zadaný vstup přeložit do takové formy, která zajistí nalezení všech slovních tvarů odvozených od jednoho základního tvaru.
Tento úkol je v podstatě možné vyřešit dvěma způsoby. Prvním z nich je doplnění textu o speciální index, obsahující základní tvary všech významových slov z daného textu (tedy všech slov, která má vůbec smysl vyhledávat). Každý takový základní tvar je potom doplněn pomocí výše zmíněného jazykového modulu o odkazy vedoucí ke všem výskytům všech tvarů odvozených od příslušného základního tvaru. Vlastní prohledávání potom již neprobíhá na celém textu, ale právě jen na onom indexu, což ve svém důsledku znamená, že celý vyhledávací proces je nezanedbatelně rychlejší než při prohledávání celého textu. Vzhledem k tomu, že se vytvoření indexu provádí pouze jednou, nevadí ani eventuální značná časová náročnost tohoto procesu . Jediným problémem této metody je přidávání dalších textů k množině již zpracovaných textů - v takovém případě je nutné vždy znovu doplnit soubor s indexy jednak o nová slova, která mohou být v nových textech obsažena, jednak o nové výskyty již zařazených slov.
Druhý způsob nepředpokládá žádné předchozí indexování textu. Uživatel zadá systému dotaz v podobě jednoho nebo více slov v základním tvaru a systém sám (opět pomocí speciálního jazykového modulu) vytvoří příslušné slovní tvary, které potom vyhledává v textu. Je zřejmé, že i v tomto případě je možné použít speciálního indexu, urychlujícího hledání. Bude ovšem mnohem větší než u prvního způsobu, protože tentokrát bude obsahovat veškeré slovní tvary (významových slov) a odkazy na místa jejich výskytu v textu.
V obou výše uvedených případech je standardní metodou pro určení základního tvaru (lemmatu) i pro vygenerování tvarů odvozených využití morfologické analýzy (lematizace) nebo morfologického generátoru, založených na rozsáhlém slovníku příslušného jazyka. Jediným slabším místem této metody je práce se slovy, která nejsou ve slovníku obsažena. I když slovník pokrývá téměř úplně slovní zásobu nějakého jazyka, vždy existují úzce odborné výrazy, nově vytvářená slova a termíny apod., které ve slovníku nejsou. Vzhledem k tomu, že se přirozený jazyk neustále vyvíjí, je možné

souhlasit s tvrzením, že žádný slovník nikdy nebude zcela pokrývat veškeré výrazy nějakého přirozeného jazyka. Z tohoto důvodu je rozumné metody založené na slovníku kombinovat s pomocnými algoritmy, schopnými nějakým způsobem zpracovat i neznámá (tedy ve slovníku neobsažená) slova. Jednou z použitelných metod je i metoda ASIMUT, popsaná v jednom z následujících oddílů.

3.2 Kontrola překlepů

3.3 MOZAIKA

Na začátku sedmdesátých let minulého století byla lingvistická skupina na MFF UK postavena před úkol navrhnout metody, které by prokázaly užitečnost zapojení počítačové lingvistiky do řešení praktických problémů v oblasti informatiky. Jedním z těchto problémů bylo automatické indexování nezkrácených dokumentů. K tomuto účelu byla Z.Kirschnerem a ostatními členy týmu vyvinuta metoda, nazvaná MOZAIKA (podle anglického akronymu MOSAIC - Morphemics Oriented System of Automatic Indexing and Condensation). Jak plný anglický název napovídá, metoda byla použitelná nejen k automatickému indexování, ale i ke kondenzaci odborných (technických) dokumentů z nejrůznějších oblastí.
Metoda MOZAIKA byla na konci sedmdesátých let několikrát implementována v různých programovacích jazycích (Systémy Q, PL1) a pro různé tématické oblasti. Použitelnost metody pro další slovanské jazyky byla testována na slovenštině a ruštině. Podobně jako metoda ASIMUT, představená v následujícím oddíle, měla i MOZAIKA hlavní výhodu v tom, že nebyla založena na rozsáhlých slovnících, pouze využívala morfologických vlastností slovanských jazyků a speciálních algoritmů k tomu, aby tyto (v tehdejší době nedostupné) slovníky nahradila.

3.3.1. Vstup a výstup

Vstupními texty byly nijak neupravené odborné dokumenty - články, studie, zprávy, knihy. Metoda pracovala s vyšší úspěšností, pokud byla předem známa jejich tématická oblast. Standardně byla metoda určena především k práci s dokumenty plné délky, v zásadě však bylo možné ji použít i na dokumenty nějakým způsobem předzpracované (na výtahy z článků, abstrakty apod.). U zkrácených dokumentů však některé implementované algoritmy postrádaly smysl, a proto i výsledky byly na těchto dokumentech horší. Ve vstupních textech byla zachována diakritika (to v době vzniku metody nebylo zdaleka samozřejmé), členění do odstavců, interpunkce, velká písmena a struktura textu (jeho rozdělení na abstrakt, úvod, seznam klíčových slov, závěr, komentáře pod ilustracemi, poznámky pod čarou apod.). To, že vstupní texty nevyžadovaly žádné ruční předzpracování, bylo nespornou výhodou MOZAIKY. Jakákoli účast člověka na přípravě textů k automatickému zpracování totiž s sebou kromě zvýšených nákladů nevyhnutelně přináší i riziko subjektivního zkreslení.
Výstupem metody byla množina jednoduchých nebo složených termínů, které určitým způsobem reprezentovaly základní téma zpracovávaných textů. Výběr těchto termínů byl založen na lingvistických kritériích, mezi nimiž hrály klíčovou roli sémantické vlastnosti koncovek podstatných a přídavných jmen.

3.3.2. Základní myšlenka

Standardní metody automatického indexování většinou využívají slovníků odborných termínů,patřících do příslušné tematické oblasti. Výrazy uvedené ve slovnících jsou pak vyhledávány v textu a řazeny podle četnosti výskytu, eventuálně i podle místa výskytu. Metoda MOZAIKA naproti tomu vychází z relativně jednoduchých jazykových zákonitostí, které umožňují se obejít bez odborných slovníků, a tak dosáhnout i větší univerzálnosti, protože celá řada jazykových pravidel platí obecně, bez ohledu na tematickou oblast.
MOZAIKA konkrétně využívá skutečnosti, že v češtině, podobně jako v celé řadě dalších přirozených jazyků, existují určité koncovky, které jsou spjaty se sémantikou daného slova. V angličtině se například koncovky *-er* nebo *-or* obvykle vyskytují u slov, označujících konatele nějaké činnosti, zatímco koncovka *-tion* obvykle signalizuje, že se jedná o nějakou činnost. Koncovky *-ity* a *-ness* nalezneme u slov označujících vlastnosti. V češtině je množina takto významných koncovek ještě bohatší. Metoda MOZAIKA především spoléhá na produktivní koncovky, které nalézáme u

terminologických elementů, ať již cizího nebo domácího původu. Mezi těmito koncovkami metoda používá zejména koncovky spjaté s názvy:
nástrojů nebo **přístrojů**, např. *-ič*, *-ač*, *-čka*, *-ér*, *-or*, *-dlo*, *-metr*, *-graf*, *-fon*, *-skop*,
procesů nebo **činností** *-ace*, *-kce*, *-áž*, *-ní*, *-za*,
vlastností *-ost*, *-ita*, *-nce*,
a s velmi typickými koncovkami přídavných jmen označujících **výsledky procesů** *-aný*, *-ený* nebo **účel** *-ací*, *-ecí*.
Poněkud méně sémanticky jednoznačné, přesto však použitelné, jsou koncovky cizího původu, např. *-ium*, *-oda*, *-ika*, *-ura*, *-éra*, *-smus* nebo zdomácnělých přídavných jmen *-tivní*, *-sívní*, *-ční*, *-zní*, *-ický* apod.
Praktické zkušenosti z implementace systému ukázaly, že pro pokrytí tematické oblasti integrovaných obvodů stačilo přibližně 800 koncových segmentů, sborník [Kirschner 83] uvádí odhad, že k plnému pokrytí současné technické terminologie by stačilo přibližně 2000 koncových segmentů. Je tedy zřejmé, že nahrazení plného slovníku odborných termínů seznamem koncových segmentů přináší kromě jiných výhod i obrovskou úsporu místa, což zejména v době vzniku této metody bylo velmi podstatné.

3.3.3. Algoritmus

Prvním krokem zpracování vstupního textu byla lematizace. Tento proces obecně znamená určení základního tvaru, ze kterého vznikl příslušný tvar nalezený v textu. V případě metody MOZAIKA byla lematizace kombinována i s morfologickou analýzou, neboť kromě určení základního tvaru rozpoznala i morfologické charakteristiky příslušného slovního tvaru, např. jeho rod, číslo, pád apod. Tyto informace jsou zachovány pro využití v následných fázích zpracování.
Nalezená lemata jsou podrobena výběru na základě sady omezení na kmen slova. V mnoha případech se totiž stává, že se některý koncový segment, nesoucí určitou sémantickou informaci, může spojovat se slovními základy (kmeny), které do dané tematické oblasti nepatří. Takové kmeny nemohou být v následujících krocích algoritmu brány v úvahu. Jako příklad nám poslouží koncový segment *-dlo*, který v češtině označuje různé prostředky, zejména nástroje či reagenty, jako např. *hradlo*, *rozpouštědlo*, *ředidlo*, *bidlo*, *kadidlo*, *páčidlo* apod., z nichž pouze první tři se pravděpodobně mohou vyskytnout v textech z oblasti elektrotechniky. Ostatní kmeny není vhodné brát v úvahu.
Kromě omezení na určité kmeny existují i omezení na určité slovní tvary. Některá slova se v odborné terminologii například mohou vyskytovat pouze v některých pádech, jiná zase v určitém čísle.
Dalšími omezeními jsou pravidla, určující jednu nebo více hlásek, které mohou nebo naopak nesmějí předcházet příslušnému koncovému segmentu. Sborník [Kirschner 83] uvádí jako příklad koncový segment *-otor*,jemuž nesmí ve slovním základu bezprostředně předcházet znak *m*.
Také je možné formulovat omezení určitých kombinací znaků, které se mohou objevit ve slovním základu. To je nutné zejména v případech velmi často se vyskytujících koncových segmentů, které jsou víceznačné. Jedná se například o segmenty *-ení*, *-ace*, *-aný*, *-ený*, *-ací*, *-ční* apod. Určení přípustného vnitřního segmentu kmene například pouze na *-pař-* umožní jako významné termíny přijmout pouze slova *odpaření*, *odpařování*, *odpařovaný*, *odpařovací*, *napařování*, *napařený*, *napařovací* apod., nikoli však *odblokování*, *odblokovaný*, *odblokovací* apod.
V případě, že je třeba vyřadit ze zpracování pouze několik výjimek, kvůli kterým je nepraktické formulovat omezovací pravidla, používá metoda tzv. negativní slovník. Jeho použití je však opravdu výjimečné, v žádné implementované verzi jeho velikost nepřekročila 60 slov.
Dalším krokem zpracování je zjednodušená syntaktická analýza, soustřeďující se zejména na analýzu jmenných skupin. Umožňuje jednotlivá vybraná slova spojit do delších termínů. To je pro úspěch metody velmi důležité, protože delší termíny lépe charakterizují obsah textu. Je zřejmé, že termín *zesilovač* obsah textu charakterizuje mnohem méně než termín *operační zesilovač TESLA KC 415*.
Modul zjednodušené syntaktické analýzy využívá údajů zjištěných morfologickou analýzou a jednoduchých pravidel.
Následuje výpočet vah jednotlivých termínů. Jeho základem je přiřazení vah jednotlivým slovním tvarům již ve fázi morfologické analýzy, teprve po fázi syntaktické analýzy je však možné spočítat váhy celých složených termínů. Během morfologické analýzy se přiřazují slovům váhy na základě:

- Pozice výrazu v textu (nadpis, podtitul, shrnutí, komentář, první a poslední odstavec, první a poslední věta v odstavci apod. Váha je exponenciální, slovo vyskytující se na nejméně významné pozici v textu obdrží váhu 1, na pozici jen o jeden stupeň významnější 2 a každý další stupeň je vždy dvojnásobkem hodnoty váhy stupně předchozího.
- Počtu slov, ze kterých se termín skládá. Delší termíny mají vyšší váhu.
- Vztahu k jiným výrazům. Jedná se zejména o vztah inkluze, kdy jeden termín je zcela obsažen v jiném termínu. V takovém případě se váha obou těchto termínů určitým způsobem zvyšuje. To je motivováno faktem, že v plynulém českém textu se jen velmi zřídka opakují delší termíny. Místo toho se používá spíše zkrácených odkazů, někdy doplněných zájmeny. Představme si třeba text:
  Jedním z nejvýkonnějších u nás vyráběných přístrojů je operační zesilovač TESLA KC 415. Tento zesilovač ...
  Zkrácený termín *zesilovač* ve druhé větě evidentně zastupuje celý název, proto je zcela přirozené, aby se váhy obou termínů navzájem ovlivnily (zvýšily).

Posledním krokem je normalizace získaných vah vzhledem k délce dokumentu. Ta je nutná, aby bylo možné porovnávat váhy termínů v jednotlivých odlišně dlouhých dokumentech a na základě toho se rozhodnout,který z dokumentů je pravděpodobně relevantnější pro daný účel.

3.4. ASIMUT

Akronym ASIMUT znamená *Automatická selekce informací metodou úplného textu*. Metoda vyhledávání informací v úplném textu (tedy v textu, který není nijak předběžně upravován pro účely snazšího vyhledávání) byla pro angličtinu vyvinuta Hortym (v [Kehl 61]).Jejím úkolem není vyhledat pouze informace o tom, ve kterém dokumentu,na které stránce či ve kterém odstavci se hledané slovo nebo slovní spojení nachází, ale má uvést i přesnou citaci kontextu. V tomto smyslu se nejedná o vyhledávání relevantních dokumentů, ale o získávání konkrétních úseků textu, tedy spíše o vyhledávání informací.
Předchůdcem systému ASIMUT (viz. [Králíková, Panevová 90]) byl systém SIUT (srov. [Cejpek 82], [Kirschner 82] a [Cejpek 88]). Hlavní výhodou obou systémů byla automatická metoda generování slovních tvarů nutných pro úspěšné vyhledávání bez použití rozsáhlého slovníku,pokrývajícího podstatnou část významových slov jazyka. Také dotazovací jazyk použitý v systému byl na svou dobu zajímavý.
Systém ASIMUT se skládal ze dvou základních částí, z **dotazovacího modulu** a z **modulu jazykového**. Text v textové databázi byl předem upraven - jednotlivé dokumenty byly rozděleny na sekce, části,odstavce věty a slova v souladu s tím, jaké členění bylo vhodné vzhledem k povaze uložených textů a vzhledem k očekávanému typu dotazů na ně. Toto rozdělení umožňovalo sjednotit formát odkazů na konkrétní slovo do tvaru šestice čísel. Tato čísla pak bylo možné nahradit titulky příslušných úseků textu (zejména na třech nejvyšších úrovních).
Důležitou součástí systému byl také tzv. **negativní slovník**,zahrnující jednak slova, která se v textech vyskytují příliš často a která by zbytečně zvyšovala počty nalezených výskytů, jednak slova,která nemají pro uživatele žádnou informační hodnotu. Jednalo se např. o spojky, zájmena, některá příslovce apod. Na první pohled může být překvapivé, že v negativním slovníku chyběly předložky, které také samy o sobě nenesou žádnou informační hodnotu. Je však třeba vzít v úvahu, že předložky velmi často bývají součástí některých víceslovných termínů nebo pojmů, jako např. *ochrana před úrazem*, *nárok na náhradu škody* apod. Nutnost začlenění negativního slovníku do systému byla v zahraničí prokázána experimentálně, např. pro angličtinu došlo ke skutečně podstatnému zkrácení doby odezvy systému v závislosti na velikosti negativního slovníku.
Při vkládání nového dokumentu do databáze bylo vždy nutné vytvořit tzv. **konkordanci**. Při tomto procesu byla všem slovním tvarům nezařazeným do negativního slovníku přiřazena adresa a frekvence výskytu, používaná pro účely urychlení hledání. Slova z negativního slovníku obdržela pouze adresu, která sloužila zejména pro správné určení vzdálenosti mezi jednotlivými významovými slovy v textu. Samotné vyhledávání potom probíhalo na konkordanci. Ta sloužila více méně jako jakýsi index, tedy se v podstatě jednalo o první z výše uvedených metod vyhledávání.

3.4.1 Dotazovací jazyk

Důležitou součástí celého systému byl dotazovací jazyk. Uživatel mohl pomocí tohoto jazyka zadat, která slova z hledaného termínu se mají vyskloňovat, a také mohl určit jejich vzájemnou vzdálenost. K tomuto účelu sloužily tzv. **distanční operátory**. Byly celkem čtyři,-1-, -2-, -3- a -4-:

-1-: Obě slova, mezi kterými byl v dotazu umístěn tento distanční operátor, musela být v prohledávaném textu umístěna bezprostředně vedle sebe.

-2-:Rozdíl mezi adresami obou slov nesměl v konkrétním úseku prohledávaného textu být větší než 3. Jinými slovy, mezi oběma slovy mohla v textu ležet maximálně dvě jiná slova.

-3-: Obě slova se musela nacházet ve stejné větě.

-4-: Obě slova se musela vyskytovat ve stejném odstavci.

Ve všech případech bylo pořadí zadaných slov libovolné, nemuselo se nutně shodovat s pořadím zadaným v dotazu. Pokud mezi dvěma slovy nebyl uveden žádný distanční operátor, bral systém jako standardní operátor -2-, tedy připouštěl mezi těmito slovy vzdálenost maximálně tří slov.

Kromě distančních operátorů bylo možné použít v dotazu ještě dva další speciální znaky. Jedním z nich byla *čárka*, znamenající disjunkci,druhým byl *vykřičník*. Pokud byl vykřičník umístěn bezprostředně za základní tvar nějakého slova v dotazu, znamenalo to pro jazykový modul pokyn toto slovo vyskloňovat. Systém nekontroloval mluvnickou shodu mezi jednotlivými slovními tvary, ve skutečnosti hledal pouze souvýskyt libovolných kombinací slovních tvarů dodaných jazykovým modulem, což samozřejmě mohlo v některých případech vést k nesprávným odpovědím.

Následující příklady dotazů byly uvedeny v [Králíková, Panevová 90] a pocházejí z verze systému ASIMUT, která byla implementována pro vyhledávání v právních textech, týkajících se stavebnictví:

Dotaz : *vzdálenost!, odstup! -3- rodinný! domek!*

Odpověď 1: Vyhl. 83/1976, Část IV., Díl I.,Oddíl II., \S 44, Věta 1

Vytváří-li *rodinné domky* mezi sebou volný prostor, musí *vzdálenost* mezi nimi být nejméně 10 m.

Odpověď 2: Vyhl. 83/1976, Část IV., Díl I., Oddíl II., \S 44, Věta 3

Vzdálenost rodinných domků vytvářejících mezi sebou volný prostor nesmí být od hranic pozemku menší než 3 m.

Odpověď 3: Vyhl. 83/1976, Část IV., Díl I., Oddíl II., \S 41, Věta 2

Jsou-li v některé z protilehlých částí stěn sousedících staveb pro bydlení okna obytných místností, nesmí být *odstup* staveb menší než výška vyšší stěny, s výjimkou staveb *rodinných domků* podle \S 44.

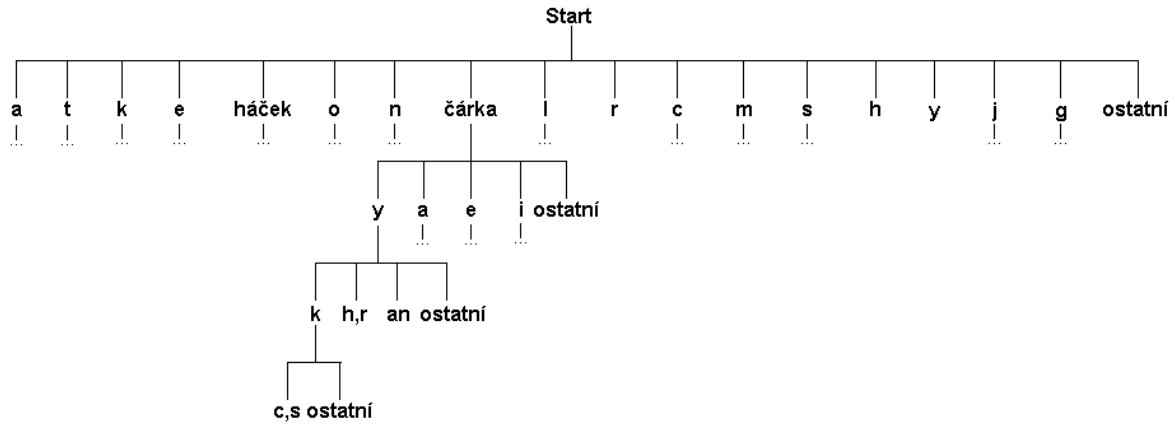
3.4.2. Jazykový modul

Nejzajímavější částí systému ASIMUT byl jeho jazykový modul. Ten dokázal v základní verzi skloňovat libovolná česká podstatná a přídavná jména, zadaná v základním tvaru. Sborník [Králíková, Panevová 90] popisuje i variantu jazykového modulu pro časování sloves, ta však nikdy nebyla implementována. Zcela zásadním problémem časování sloves pouze na základě jejich infinitivu je mnohem větší víceznačnost slovesných tvarů, než je tomu u podstatných nebo přídavných jmen.

Jazykový modul prošel několika vývojovými stadii, při kterých byla postupně minimalizována role uživatele tak, že v poslední verzi od něho systém nepožadoval zadání žádných doplňkových informací k základnímu tvaru. V dřívějších verzích jazykový modul požadoval zadání slovního druhu a u podstatných jmen také rodu. Navíc uživatel musel dodat ještě další informace v případě, že u daného slova během skloňování docházelo ke změnám kmene. Toto řešení bylo nepraktické,proto autoři zcela změnili přístup k problému. Namísto toho, aby podkladem pro skloňování byly jak údaje o kmeni daného slova (slovní druh, rod), tak i o koncovém segmentu (ne vždy se jednalo o koncovku), rozhodli se autoři vycházet pouze z vlastností jednotlivých koncových segmentů základních tvarů. Hlavním podkladem pro tuto metodu byl **retrográdní slovník** Slavíčkové [Slavíčková 75]. Retrográdní slovník se od běžného slovníku liší řazením hesel podle abecedy. Zatímco běžný slovník je seřazen podle abecedy nejprve podle prvního znaku, dále podle druhého, třetího, ev. dalších, je retrográdní slovník seřazen opačně, tedy nejprve podle posledního znaku, dále podle předposledního atd. Tento způsob řazení měl pro přípravu algoritmu jazykového modulu zcela klíčový význam. V retrográdním slovníku jsou totiž jednotlivé základní tvary se stejným koncovým segmentem umístěny bezprostředně vedle sebe a je tedy možné relativně snadno nejen zjistit, jaký koncový segment

jednoznačně určuje vzor pro skloňování, ale i najít všechny výjimky, které se chovají jinak než ostatní slova se stejným koncovým segmentem základního tvaru. V době, kdy prakticky neexistovaly slovníky v elektronické podobě (u nichž je vytvoření retrográdní verze jednoduchým programátorským cvičením), byla existence slovníku [Slavíčková 75] pro vylepšení jazykového modulu systému ASIMUT skutečně klíčová.
Publikace [Králíková, Panevová 90] uvádí jako příklad jednoho z nejkratších koncových segmentů, které stačí k jednoznačnému určení slovního druhu a vzoru pro skloňování, segment *-ý*. Všech přibližně 14 000slov s tímto koncovým segmentem v retrográdním slovníku [Slavíčková 75] se skloňuje stejně (většina z nich jsou tvrdá přídavná jména skloňující se podle vzoru *mladý*). Existuje pouze jediná výjimka,podstatné jméno *úterý*. Na druhou stranu existují i koncové segmenty, u nichž ani nejdelší možný koncový segment nedává spolehlivou informaci o skloňování. Jedná se například o segmenty *-ák*, *-ík* a *-or*, které zakončují jak životná tak neživotná podstatná jména rodu mužského, např. *sedlák/bodlák*, *trávník/právník* či *operátor/generátor*. V takovém případě systém vygeneruje tvary skloňování v obou rodech, tedy kromě jiných i tvary *trávníkovi/právníkách*.
Tam, kde je to nutné, systém dokáže i odvodit potřebné odvozené kmeny. Systém pracuje s těmito variantami kmenů:
E: Vložení *-e-* do kmene, např. *patro/pater*, *heslo/hesel* apod.
V: Vynechání *-e-*, např. *počet/počtu*, *kotel/kotle* apod.
K: Zkrácení kmenové samohlásky, např. *síla/silou*, *vůz/vozu* apod.
D: Prodloužení kmenové samohlásky, např. *zdravý/zdráv*, *starý/stár* apod.
M: Měkčení kmenové souhlásky, např. *pýcha/pýše*, *matka/matce* apod.
P: Přivlastňovací měkčení přídavných jmen, např. *matka/matčin* .

Jak již bylo uvedeno, jazykový modul systému ASIMUT nepracoval se žádným slovníkem základních tvarů. Rozsáhlý slovník v něm byl nahrazen seznamy výjimek pro jednotlivé koncové segmenty. Tyto seznamy měly obvykle pouze několik položek, nejdelší z nich obsahovaly řádově desítky výjimek.



Obr. 1 Částečná struktura algoritmu odtrhávání koncových segmentů v jazykovém modulu

Algoritmus jazykového modulu byl velmi jednoduchý - procházel vstupní slovo od posledního znaku tak dlouho, dokud nebylo možné jednoznačně určit, jaký vzor pro skloňování je třeba danému slovu přiřadit. Částečná struktura tohoto algoritmu je znázorněna na obrázku 1, tři tečky znamenají, že proces odtrhávání koncových segmentů v dané větvi ještě není ukončen.
Před přidáním pádových koncovek k příslušným základním tvarům (každá koncovka v seznamu obsahovala jeden z výše uvedených kódů základního tvaru, ke kterému má být připojena) systém v některých případech ještě procházel seznamem výjimek. Celkem se jednalo o řádově několik stovek výjimek. Znaky s diakritickými znaménky byly reprezentovány vždy dvojicí znaků - příslušným znakem bez diakritického znaménka a číselným kódem:
2 místo čárky, 3 místo háčku a 7 místo kroužku nad *u*.
Slovo *králíčků* tak bylo například reprezentováno řetězcem *kra2li2c3ku7*. Tento způsob kódování češtiny nejen umožňoval jednotné zacházení s českou diakritikou v době, kdy prakticky neexistoval

žádný standardní způsob kódování češtiny, ale měl také vliv na samotný algoritmus. Bylo totiž možné snadno v průběhu výpočtu formulovat dotazy typu *"Je na konci slova dlouhá samohláska?"*.
Pokusme se ilustrovat práci jazykového modulu pomocí několika příkladů. Vezměme např. základní tvar slova *kojenec* a pokusme se projít algoritmem popsaným v [Králíková, Panevová 90]:
Systém nejprve testuje poslední znak základního tvaru, v tomto případě znak *c*. Vytvoří základní kmen (Z), který je v tomto případě totožný se vstupním tvarem (*kojenec*). V zápětí systém vytvoří i odvozený kmen (V) neobsahující samohlásku *e*: *kojenc*. Poté projde seznamem výjimek (ten v tomto případě obsahuje slova *konec*, *odstavec*, *Třinec*, *válec* a *živec*, která se skloňují podle vzoru *stroj*, a slova *obec*, *plec* a *klec*, jež mají vzor *píseň*). Posledním krokem je připsání koncovek vzoru *muž* k příslušným kmenům. Výsledkem je množina následujících slovních tvarů:
kojenec, *kojence*, *kojenci*, *kojencovi*, *kojencem*, *kojenců*, *kojencům*, *kojencích* a *kojencové*. Jak je vidět na tomto příkladu, systém se nevyhnul vytvoření nesprávných tvarů *kojencovi* a *kojencové*. To však v zásadě nevadí, neboť tyto tvary se v textech pravděpodobně nebudou vyskytovat.
Dalším základním tvarem, jehož tvary budeme odvozovat,je slovo *dům*. Toto slovo nám umožní ukázat, jakým způsobem se systém vyrovná s relativně náročnou změnou kmenové samohlásky. Stejně jako v předchozím případě systém nejprve zjišťuje,jakou hláskou končí základní tvar slova. Po zjištění, že se jedná o hlásku *m*, testuje, zda předcházející hláskou je *u*. V případě slova *dům* tomu tak není, protože tento slovní tvar je zakódován jako *du7m* a předposledním znakem proto není *u*, ale *7*. Následujícím krokem je kontrola seznamu výjimek. V něm systém nalezne právě slovo *dům* a vytvoří změnou kmenové samohlásky k základnímu tvaru i tvar, označený symbolem K: *dom*. Po průchodu seznamem koncovek pak systém vygeneruje následující slovní tvary:
dům, *doma*, *domu*, *dome*, *domě*, *domem*, *domy*, *domé*, *domů*, *domům*, *domech*, *domích*, *domi*
V tomto případě systém vydá rovněž několik chybných tvarů, pouze jeden z nich (*doma*) však teoreticky může působit problémy při vyhledávání, neboť se zároveň jedná o příslovce. Určitým řešením pro podobné případy je zařazení příslušného problematického slova do negativního slovníku, pokud je to možné.
Abychom ukázali, že v některých případech je systém schopen vygenerovat požadované tvary naprosto přesně, uvedeme ještě jeden příklad, slovo *žádost*. Standardním prvním krokem je test posledního znaku. Po něm následuje kontrola, zda před koncovým *t* předchází skupina hlásek *os*. Odpověď je kladná, proto následuje průchod seznamem výjimek. V něm se nacházejí podstatná jména rodu mužského, skloňovaná podle vzoru *aspekt*, např. *chvost*, *dorost*, *nerost*, *porost* apod. Slovo *žádost* mezi výjimkami není, proto systém k základnímu tvaru připojí koncovky vzoru *kost*. Tím dostaneme tyto tvary:
žádost, *žádosti*, *žádostí*, *žádostem*, *žádostech*, *žádostmi*

3.5. ASPI

3.6. ORACLE Context

4. Syntaxe

4.1. Základní pojmy
Martin

4.2. Systémy Q
Alainem Colmerauerem vyvinuté Systémy Q, nazývané často také Q jazyk, reprezentují formalismus pro transformaci grafů, jejichž hrany jsou ohodnocené stromy. Jedná se o interpretovaný deklarativní vyšší programovací jazyk (jímž se inspirovali tvůrci Prologu, jehož byl A. Colmerauer spolutvůrcem), pomocí něhož lze popisovat obecné gramatiky pomocí systémů přepisovacích pravidel provádějících transformaci na podřetězcích stromových grafů v lineární reprezentaci. Takové systémy, nezávislé sady pravidel a gramatik, mohou tvořit sekvenci neboli systém samostatných systémů, které se potom nazývají podsystémy.

V rámci každého podsystému se berou v úvahu všechny možné kombinace aplikací pravidel na vstupní řetězec či jeho část, přičemž, neformálně řečeno, “přežijí” pouze ty řetězce, které tvoří

cestu z počátečního do koncového vrcholu. V tzv. fázi čištění se odstraňují hrany neležící v řezětcích splňujících tuto podmínku. V každé fázi slouží výstup předchozího podsystému jako vstup podsystému bezprostředně následujícího.

4.2.1.Syntaxe

Systémy Q používají dva druhy znaků: jednak písmena a číslice reprezentované standardní sadou alfanumerických znaků (26 písmen a 10 číslic), jednak znaky speciální, jež zahrnují znaky signifikantní (+ – * / ( ) $ = " . ,), a znaky nesignifikantní, reprezentované výběrem ostatních znaků z repertoáru konkrétního počítače. Bílé znaky jsou ignorovány. Mezery mezy slovy (přesněji hrany mezi vrcholy řetězce) se reprezentují znakem +.

V jazyce se používají tři typy objektů: atomy, stromy a seznamy stromů. Atom je řetězec znaků, z nichž první může být libovolný a ostatní mohou být pouze alfanumerické; tedy $PAT, *3HS5L, CIRCUIT, –1, * jsou dobře formulované atomy, zatímco 3/, ;;, SO-CALLED akceptovány nebudou. Stromem se rozumí souvislý orientovaný graf bez cyklů, přičemž do jednoho z jeho vrcholů (tzv. kořen) nevstupuje žádná hrana. V Systémech Q se stromy reprezentují lineárním zápisem pomocí závorek (strom na obrázku je pak reprezentován zápisem A(B,C(D))). Vrcholy mající stejného předka jsou odděleny čárkou. Atom reprezentuje nejmenší možný strom, neboť každý strom musí mít alespoň jeden vrchol.

B

Seznam reprezentuje n-tici stromů (nebo atomů) oddělenou čárkami v lineární reprezentaci, t.j. množinu sousedících vrcholů v grafu, z nichž každý je zapsán včetně případných synů. Seznam může být rovněž prázdný či jednoprvkový, přičemž samostatný strom nebo atom jsou považovány za jednoprvkový seznam.

V pravidlech mohou být výše popsané objekty zastoupeny proměnnými. Používají se tři typy proměnných:

proměnné pro atomy — značí se písmeny ze začátku abecedy: a – f
proměnné pro stromy — značí se písmeny ze středu abededy: i – n
proměnné pro seznamy — značí se písmeny z konce abecedy: u – z.

Za písmenem označujícím typ proměnné následuje znak *. Je-li potřeba více proměnných, lze přidat k písmenům indexy 1–9. Proměnné se tedy zapisují takto: A*, E*, A*1, E*5, U*3.

Pravidla mají tvar x == y / n. , což lze interpretovat následovně: je-li podmínka n splněna, přepiš výraz x výrazem y. Pravidla jsou nezávislá v tom smyslu, že proměnné vyskytující se v jednom pravidle nemají žádný vliv na pravidla jiná. Jakákoli proměnná vyskytující se v pravé části pravidla nebo podmínce se musí vyskytovat v levé části. Libovolná proměnná nebo konstanta může v pravé části chybět (což znamená její vymazání) a lze přidat libovolné množství konstant.

Levé části pravidel jsou porovnány s podřetězci vstupního řetězce a pokud dojde ke shodě, je odpovídající podřetězec nahrazen strukturou z pravé části pravidla. To se děje se všemi pravidly a ve všech možných pořadích použití, dokud nějaká shoda existuje. Tzv. čistící procedura potom “promaže” výsledné struktury. Automatické kombinování výpočetních variant může v některých případech působit problémy: nepřesnými formulacemi jednotlivých pravidel nebo jejich libovolnou kombinací mohou vznikat nekonečné smyčky. Pokud např. nějaké pravidlo produkuje kromě jiného také svou nezměněnou levou část nebo pokud se toto děje byť jen v jednom možném pořadí aplikace více určitých pravidel, dojde nevyhnutelně k zacyklení. a == a + b . nebo a == c . c == a + b . jsou jednoduché a názorné příklady takových chybných pravidel.

Podmínky určují omezení platnosti pravidel přímou nebo nepřímou specifikací rozsahu proměnných: buď explicitním vyjmenováním množiny objektů, kterých může proměnná nabývat, nebo stanovením vztahu mezi proměnnými. Podmínka následuje za pravou částí pravidla a je oddělena lomítkem. V podmínce lze použít dvě skupiny operátorů: běžné operátory =, ≠, NOT, AND, OR značené v kódu =, ", –NON–, –ET–, –OU–, a pár speciálních operátorů –DANS– a –HORS–.

Operátor –DANS– vyjadřuje vztah inkluze (u –DANS– v znamená, že seznam v obsahuje seznam u) a tato relace je asymetrická. Jeho protějšek –HORS– znamená prázdný průnik dvou seznamů: u –HORS– v znamená, že seznamy u, v nemají žádný společný prvek. Tento operátor je symetrický. Je-li jeden z parametrů (nebo oba) –NUL– (prázdný seznam), je operátor –HORS– vyhodnocen jako nepravdivý.

Priority operátorů jsou následující:

1. =, '', –DANS–, –HORS– 2. –NON– 3. –ET– 4. –OU–

V případě potřeby lze ke změně priorit použít závorky, které se zapisují symboly (. a .) (tečka v zápisu je důležitá, protože samotné závorky lze považovat za atom).

Příklad podmínky : u –HORS– v –ET– (. *PL –DANS– w –OU– *SG –HORS– x .) .

Konec podsystému a začátek následujícího se signalizuje zvláštním symbolem –REQ–. Speciálním a užitečným operátorem je tzv. dezintegrátor. Značí se symbolem $$ a reprezentuje seznam jednotlivých znaků příslušného atomu, tedy např. $$AUTOMATON reprezentuje seznam atomů A,U,T,O,M,A,T,O,N. Tímto způsobem lze atomy rozkládat a opět rekonstruovat. Pokud se část pravidla objevuje v nezměněném tvaru, není třeba ji znovu rozepisovat, ale je možné použít operátor ––. To se týká i podmínek. Je-li podmínka předcházejícího pravidla obsažena v podmínce zpracovávaného pravidla, lze ji celou nahradit tímto operátorem. Samozřejmě musí odpovídat označení proměnných a veškeré použité proměnné se musí vyskytovat v levé části pravidla.

Komentář je uvozen dvojznakem ** a ukončen tečkou.

Proceduru aplikace pravidel lze reprezentovat speciálním grafem s jedním počátečním a jedním koncovým vrcholem bez cyklů. Nazývá se řetězec nebo řetězcový graf a obsahuje řetězec uzlů pospojovaných hranami, přičemž existuje cesta z počátečního do koncového vrcholu. Hrany jsou ohodnoceny stromy a vrcholy představují hranice mezi stromy. Vrcholy lze rozdělit na dva typy: bez větvení a takové, z kterých je možné pokračovat více způsoby. První typ se označuje znakem +, druhý symbolem –n–, kde n je číslo vrcholu. Při aplikaci pravidel jsou do grafu přidávány řetězce hran, dokud nejsou vyčerpány všechny možnosti aplikace pravidel. Je-li vstupní řetězec správně formulován a jsou-li pravidla korektní a adekvátní, je výsledkem jedna struktura spojující počáteční a koncový vrchol. Dvě či více různých výsledných struktur signalizují nejednoznačnost vstupního řetězce, více stejných struktur ve výsledku poukazuje na redundanci v systému pravidel.

4.3 Tree adjoining grammars

Tree adjoining grammars (TAG) jsou formalismem popsaným Joshim, Levym a Takahashim v polovině sedmdesátých let. Později bylo navrženo několik modifikací těchto gramatik, jako např. lexikalizované TAG (LTAG) nebo TAG s omezeními (FTAG). Specifickým rysem tohoto formalismu je, že nepracuje s řetězci, nýbrž se stromy. Formální síla je větší než u bezkontextových gramatik, některé z nich rozpoznávají jazyky kontextové.

Základními složkami TAG jsou elementární stromy, které se dělí na iniciální a pomocné. Na stromech odvozených z elementárních jsou definovány operace substituce a připojení, s jejichž pomocí lze stromy kombinovat a vytvářet tak složitější struktury. Jazykem gramatiky je množina všech stromů, které lze odvodit zmíněnými operacemi ze stromů elementárních.

Iniciální strom je takový strom, jehož vnitřní vrcholy jsou ohodnoceny neterminálními symboly a listy jsou ohodnoceny buď terminály, nebo neterminály označenými symbolem substituce (↓).

Pomocný strom je definován jako iniciální s tím rozdílem, že právě jeden jeho list (tzv. připojovací) musí být označen symbolem připojení (*). Tento list musí být ohodnocen stejným neterminálem jako kořen.

Odvozené stromy vznikají z iniciálních a pomocných stromů substitucí a připojením. Substituce stromu α ve stromě α' znamená nahrazení substitučního vrcholu stromu α' stromem α, přičemž musí platit, že substituční vrchol v α' a kořen stromu α jsou ohodnoceny stejným neterminálem. Substitucí mohou odvozené stromy vznikat pouze ze stromů iniciálních a odvozených.

Operace připojení je o něco složitější. Připojení pomocného stromu β k vrcholu *u* stromu γ probíhá takto:

1. Neterminální symbol kořene stromu β musí být shodný s neterminálem ohodnocujícím vrchol *u*.

2. Podstrom *t* stromu γ, jehož kořenem je *u*, je odstraněn a nahrazen stromem β, přičemž *t* nahradí vrchol stromu β označený symbolem připojení.

Rozšířením TAG jsou lexikalizované tree adjoining grammars, ve kterých navíc platí podmínka, že každý elementární strom obsahuje alespoň jeden list ohodnocený terminálním symbolem, jinými slovy, každý elementární strom je asociován s alespoň jedním terminálem.

Dalším rozšířením je definování omezení pro operaci připojení. Je možné například explicitně zakázat připojení k určitému vrcholu nebo povolit připojení pouze některých pomocných stromů. Jinou možností omezení jsou tzv. rovnice rysů.

V FTAG má každý vrchol elementárního stromu přiřazeny dvě proměnné nazývané *top* a *bot*. V případě substituce stromu α za nějaký vrchol *u* stromu α' se nahrazení provede jako dříve popsaná substituce, ale pouze tehdy, pokud hodnoty proměnných *top* a *bot* kořene stromu α jsou shodné s odpovídajícími hodnotami proměnných vrcholu *u*.

Poněkud komplikovanější připojení pomocného stromu β k vrcholu *u* stromu γ se provádí jako u TAG bez omezení za splnění dvou podmínek:

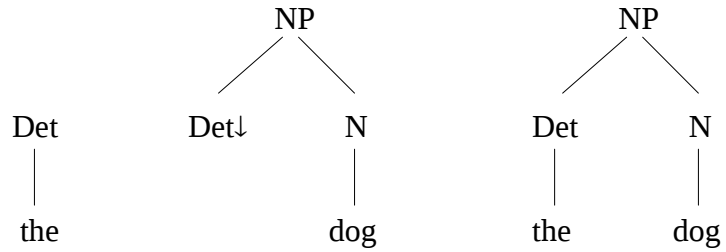
1. Hodnota proměnné *top* vrcholu *u* je shodná s proměnnou *top* kořene stromu β.
2. Hodnota proměnné *bot* vrcholu *u* je shodná s proměnnou *bot* připojovacího uzlu stromu β.

Díky takovéto formulaci podmínek zůstává ohodnocení proměnných *top* a *bot* v odvozeném stromě konzistentní. Odvozený strom obsahující vrchol, jehož hodnoty proměnných *top* a *bot* nelze unifikovat, nepatří do množiny stromů generované gramatikou. Omezení odvozování stromů pomocí podmínek je vhodné např. pro zajištění shody v čísle, pádě apod.

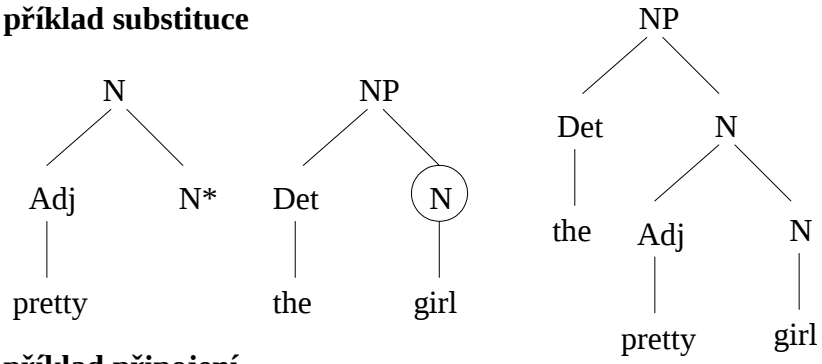
Při zpracování přirozeného jazyka často dochází k převádění zpracovávaných dat z jednoho způsobu zápisu do jiného, ať už se jedná o strojový překlad mezi přirozenými jazyky nebo o převod textu do formální logické reprezentace. K podpoře takových procesů byly vyvinuty synchronní tree adjoining grammars. Synchronní gramatika se skládá ze dvou gramatik, přičemž elementární stromy tvoří páry (v každém páru jsou stromy z různých gramatik) a v každém páru jsou definovány korespondence mezi vrcholy obou stromů. Generování probíhá paralelně u obou gramatik. Mějme pár odvozených stromů α a β:

1. Zvolíme korespondenci vrcholů *u* a *v* z páru stromů α a β (vrchol *u* je z α, vrchol *v* je z β).
2. Zvolíme pár pomocných stromů γ a δ takový, že strom γ lze připojit na vrchol *u* ve stromě α a strom δ na vrchol *v* ve stromě δ.
3. Na obou stromech provedeme operaci připojení.

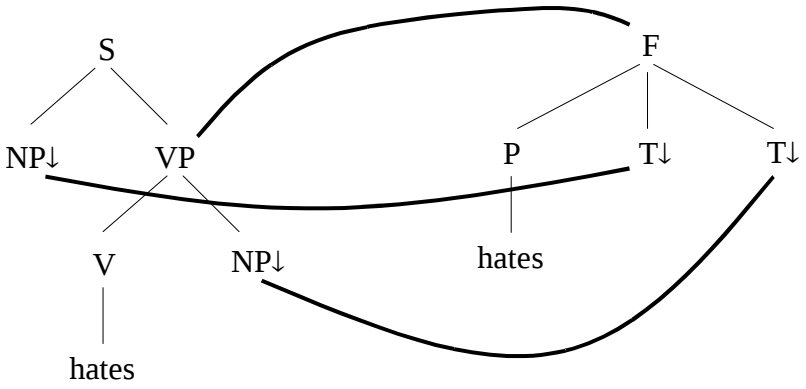
Popsaný formalismus je zaměřen na praktické použití. Na University of Pennsylvania byla v rámci projektu XTAG (viz www.cis.upenn.edu/~xtag) pomocí lexikalizovaných TAG implementována gramatika angličtiny. Kromě gramatiky jsou výsledkem projektu programové nástroje pro tvorbu TAG, analýzu apod.



příklad substituce



příklad připojení



příklad páru synchronní gramatiky
(převod syntaktické struktury na logický predikát)

4.4. Unifikační syntaktické formalismy

Téměř bez výjimky je základním metodologickým východiskem všech formalismů vytvořených za účelem zpracování syntaxe přirozeného jazyka na počítači myšlenka, že struktura věty má být prostředky formalismu popisována staticky - popis syntaktické struktury má být prostřednictvím formalismu formulován jako popis pevných, neměnných vlastností, které lze na větě a její struktuře jakožto na objektech lingvistického pozorování spatřit. Tato zásada přitom platí všeobecně, ať už jde o řešení takových úloh počítačového zpracování jazyka jako je analýza vět, generování vět, nebo o jiné aplikace. To je patrný posun od transformačních popisů let šedesátých a sedmdesátých minulého století, ve kterých k popisu struktury věty nedílně patřil popis vzniku této struktury (úkolem syntaxe vlastně bylo především popsat postup vzniku struktury, nikoliv samotnou strukturu, ta byla jen

jakýmsi „vedlejším produktem“ syntaktického popisu). Tato změna pohledu na metodologii formalizace syntaxe, tj. přechod od popisů tzv. procedurálních (tranformačních) k popisům tzv. deklarativním (netransformačním), ovšem nenastala pouze v aplikační oblasti, ale je do značné míry posuvem i v oblasti čistě teoretické, o čemž svědčí rozšíření řady lingvistických teorií založených na stejné základní myšlence za posledních zhruba dvacet let (Functional Unification Grammar, Kay 1978; Lexical Functional Grammar, Bresnan (ed.) 1982; Generalized Phrase Structure Grammar, Gazdar, Klein, Pullum & Sag 1985, Head-driven Phrase Structure Grammar, Pollard & Sag 1987, 1994; jakožto i mnohé jejich varianty a další, méně rozšířené teorie).

4.4.1. Popis vlastností objektů - jednoduché sestavy rysů

Základem těchto formalismů (a deklarativního přístupu k popisu syntaxe obecně) je tedy popis (statické) syntaktické struktury pomocí informace, kterou o ní na základě pozorování máme. Celková informace se pak skládá z informací dílčích, a základním kamenem takové informace je popis jedné pozorované vlastnosti – v dalším jej budeme nazývat *rys* (angl. *feature*). Formalizace rysu – (popisu) znalosti jedné takové základní vlastnosti - je přitom rozložena do dvou částí: do (popisu) názvu vlastnosti a do (popisu) hodnoty vlastnosti. Název vlastnosti přitom udává typ informace, hodnota tento typ konkretizuje (viz příklady bezprostředně níže). Jednotlivé rysy, tj. jednotlivou informaci o vlastnosti nějakého objektu (nejtypičtěji pro nás samozřejmě objektu syntaktického, jako je slovo nebo konstrukce), potom schematicky zapisujeme jako

<název_vlastnosti> : <hodnota_vlastnosti>

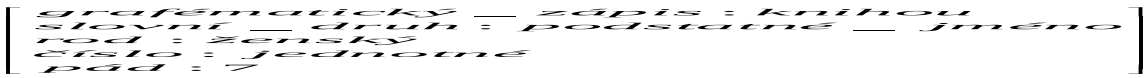
kde jak <název_vlastnosti> tak <hodnota_vlastnosti> jsou metasymboly, které v konkrétním zápise budou nahrazeny konkrétními hodnotami. V dalším textu budeme předpokládat, že <název_vlastnosti> je vždy identifikátor (posloupnost sestávající z písmen, číslic a podtržítek, začínající malým písmenem), <hodnota_vlastnosti> může být obecně vzata z pestřejšího repertoáru, ale v této chvíli předpokládejme, že je zde možný pouze identifikátor nebo číslo (dále v textu potom možnosti postupně rozšíříme).

Příklady rysů - formálně vyjádřených vlastností objektů:

barva	:	žlutá
cena_v_korunách	:	100
slovní_druh	:	zájmeno

(Syntaktický) objekt je v takovémto přístupu popsán (reprezentován) jako souhrn rysů, tj. vlastností, které tento objekt má (přesněji řečeno: souhrn informací, které jsou o vlastnostech tohoto objektu známy). Popis objektu budeme tedy formalizovat jako množinu vlastností, které jsou o tomto objektu známy. Abychom dali najevo, že se nejedná o “obyčejnou” množinu, ale o popis objektu, budeme psát každou vlastnost na nový řádek a celý zápis uzavřeme do velkých hranatých závorek. Podobně jako při zápisu množiny nebude přitom pořadí zápisu jednotlivých vlastností objektu hrát žádnou roli, tj. dva zápisy, které se liší pouze pořadím zápisu rysů, budeme pokládat za identické. Takový zápis budeme nazývat *sestava rysů* (angl. *feature structure*).

Příklad: morfologické vlastnosti slovního tvar “knihou” popíšeme následující sestavou rysů:



Stanovíme si přitom formální omezení, že žádný <*název_vlastnosti*> se v sestavě rysů nesmí vyskytnout více než jednou. Takové omezení odpovídá intuitivní představě, že dvojí zápis stejné vlastnosti je zbytečný, pokud by šlo o zápis, který by přiřazoval stejnému názvu vlastnosti stejnou hodnotu této vlastnosti, a že by vedl ke sporu, pokud by stejnému názvu vlastnosti přiřazoval hodnoty

různé. Zvláště ve druhém případě je toto omezení velmi důležité, neboť automaticky zabraňuje nekonsistenci popisu vlastností objektu.

Příklad: následující zápis morfologické informace o slovním tvaru “knize” není zápisem platné (správně tvořené) sestavy rysů, protože se tomto zápise objevuje dvakrát vlastnost s názvem “pád”, což není přípustné (odporuje to definici sestavy rysů).

$$\begin{bmatrix} grafémický\_zápis : kniže \\ slovní\_druh : podstatné\_jméno \\ pád : 3 \\ rod : ženský \\ pád : 6 \\ číslo : jednotné \end{bmatrix}$$

Vyloučení zápisů obsahujících stejné jméno vlastnosti vícekrát s sebou ovšem přináší problém, jak zachycovat situace, kdy jeden rys může potenciálně mít několik hodnot (např. v případě víceznačnosti popisovaného objektu, viz minulý příklad). Tento problém vyřešíme níže.

4.4.2. Unifikace jednoduchých sestav rysů

Kromě bezprostřední informace o vlastnostech objektů a o objektech jako takových je důležité pracovat i s informací o informaci (rozumí se “s informací o informaci o objektech a jejich vlastnostech”). Nejdůležitější mezi nimi je informace o tom, že dvě informace (tj. dva popisy objektů, dvě sestavy rysů) popisují tentýž objekt. V tom případě je zřejmě rozumné oba příslušné popisy (sestavy rysů) zkombinovat do popisu jediného, který ponese najednou informaci předtím rozdělenou. Takovou operaci budeme nazývat *unifikace* (plným termínem *unifikace sestav rysů*).

Příklad: mějme dvě sestavy rysů

$$\begin{bmatrix} slovní\_druh : sloveso \\ osoba : 3 \\ číslo : množné \end{bmatrix}$$

$$\begin{bmatrix} rod : mužský\_životný \\ číslo : množné \end{bmatrix}$$

a současně informaci, že obě tyto sestavy rysů popisují tentýž objekt. Pokud výše zmíněnou operaci unifikace označíme symbolem $\cup$ (tj. stejným symbolem jakým značíme sjednocení množin, což není podobnost vůbec náhodná), můžeme pro sestavy rysů v tomto příkladě psát následující rovnost (která by čtenáři měla být zřejmá z analogie mezi unifikací sestav rysů a sjednocením množin).

$$\begin{bmatrix} slovní\_druh : sloveso \\ osoba : 3 \\ číslo : množné \end{bmatrix} \cup \begin{bmatrix} rod : mužský\_životný \\ číslo : množné \end{bmatrix} =$$

$$\begin{bmatrix} slovní\_druh : sloveso \\ osoba : 3 \\ rod : mužský\_životný \\ číslo : množné \end{bmatrix}$$

Výsledný objekt může být popisem nějakého slovesného tvaru ve třetí osobě času minulého, čísla množného, rodu mužského životného (např. “přišli, pracovali, …”)

Zajímavá (a relevantní) je otázka, jaký má být výsledek unifikace v případě, že sestavy rysů, které do operace vstupují, nesou informace, jež spolu nejsou kompatibilní, tj. v případě, že by unifikací vznikla “formálně nesprávná sestava rysů”. Triviálním příkladem takové dvojice sestav rysů je pár

$$\begin{bmatrix} číslo : jednotné \end{bmatrix} \quad \text{a} \quad \begin{bmatrix} číslo : množné \end{bmatrix}$$

Je ovšem zřejmé, že příklady nekompatibilní informace mohou být i mnohem rozsáhlejší, v podstatě k nekompatibilitě dvou sestav rysů stačí, když se popis jedné jediné vlastnosti vyskytuje v obou

sestavách, v každé s jinou hodnotou (bez ohledu na počet a kompatibilitu všech ostatních rysů v sestavách)[1].
V takovém případě by na první pohled výsledek operace zřejmě vůbec neměl být definován (podobně jako např. v aritmetice není definován výsledek dělení nulou), což ostatně odpovídá našemu omezení týkajícímu se počtu jmen vlastností obsažených v jedné sestavě rysů (obecně bychom pro zachycení kombinace informace z obou sestav rysů potřebovali, aby v sestavě výsledné bylo jméno vlastnosti, ve které jsou původní sestavy nekompatibilní, dvakrát, pokaždé s jinou hodnotou).
Obecně se však přijímá poněkud odlišný pohled. Ten vychází z toho, že každá sestava rysů obsahuje určité množství informace o objektu, který popisuje. Z tohoto hlediska je výsledkem unifikace dvou nekompatibilních sestav rysů (ať už jejich nekompatibilita spočívá v jakémkoliv rysu či v jakýchkoliv rysech a bez ohledu na počet takových nekompatibilních rysů) sestava speciální, která má tu vlastnost, že obsahuje (intuitivně řečeno) "příliš mnoho" informace – totiž tak mnoho, že je to sestava vnitřně sporná. Taková sestava se značí $\perp$ a definitoricky se pokládá za výsledek unifikace ve všech těchto případech.

Příklad: platí tedy například následující rovnosti

$$\left[\textit{číslo} : \textit{jednotné}\right] \cup \left[\textit{číslo} : \textit{množné}\right] = \begin{bmatrix} \textit{slovní\_druh} : \textit{sloveso} \\ \textit{osoba} : 3 \\ \textit{číslo} : \textit{jednotné} \end{bmatrix} \cup \begin{bmatrix} \textit{rod} : \textit{mužský\_životný} \\ \textit{číslo} : \textit{množné} \end{bmatrix} =$$

$$= \begin{bmatrix} \textit{osoba} : 3 \\ \textit{číslo} : \textit{množné} \end{bmatrix} \cup \begin{bmatrix} \textit{slovní\_druh} : \textit{zájmeno} \\ \textit{osoba} : 2 \\ \textit{číslo} : \textit{množné} \end{bmatrix} = [\textit{barva:černá}] \cup [\textit{barva:bílá}] = \perp$$

Formálně tedy můžeme definovat unifikaci dvou sestav rysů SR1 a SR2 následujícím způsobem:
nechť SR1 obsahuje rysy r11,r12, …, r1m, po řadě s hodnotami h11, h12,…, h1m,
nechť SR2 obsahuje rysy r21,r22, …, r2n, po řadě s hodnotami h21, h22,…, h2n.
Pak SR1 $\cup$ SR2 je definováno jako
a) sestava obsahující rysy r11,r12, …, r1m,r21,r22, …, r2n po odstranění duplicit právě tehdy, když pro všechna i ε {1,...,m}, j ε {1,...,n} platí právě jeden ze vztahů
    a1. rys s názvem r1i není obsažen mezi rysy sestavy SR2
    a2. rys s názvem r2j není obsažen mezi rysy sestavy SR1
    a3. rys r1i je identický (co do názvu i hodnoty) s rysem r2j pro právě jedno j
b) sestava $\perp$ ve ostatních případech (tj. v případech, kdy existují i,j ε N taková, že název rysu r1i je shodný s názvem rysu r2j, ale hodnoty r1i a r2j jsou rozdílné).
Je zřejmé, že tato definice je vlastně jen jednoduchým rozšířením běžné definici sjednocení množin - jediný rozdíl je v tom, že jsou (oproti definici množinového sjednocení) ještě přidány podmínky, které zaručují, že výsledek operace bude platnou strukturou rysů.

4.4.3. Jednoduché unifikační gramatiky

V předchozím textu jsme v podobě sestav rysů a operace unifikace na těchto strukturách vybudovali základní formální aparát pro zápis jednoduchých unifikačních gramatik. Předveďme si nyní na krátkém příkladě, jak unifikační gramatiky vypadají a především to, jaký je jejich vztah k "normálním" bezkontextovým gramatikám.

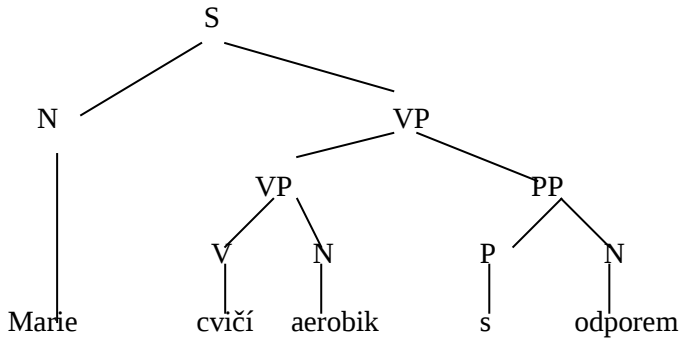
Mějme jednoduchou bezkontextovou gramatiku, jejíž jazyk obsahuje větu "*Marie cvičí aerobik s odporem*". Taková gramatika může vypadat následujícím způsobem:

[1] Z toho by bylo možné odvodit i definici unifikace rysů: dva rysy lze unifikovat právě tehdy, jsou-li identické (tj. mají-li stejný nazev a stejnou hodnotu). Výsledkem takové unifikace je právě tato hodnota. Běžně se však unifikace rysů nedefinuje, protože definice unifikace struktur rysů je postačující.

Pravidla:

S	→	N	VP	N	→	Marie
VP	→	V	NP	V	→	cvičí
VP	→	VP	PP	N	→	aerobik
PP	→	P	N	P	→	s
				N	→	odporem

Taková gramatika přiřadí příkladové větě strukturu



(Je zřejmé, že tato gramatika slouží pouze pro ilustrační účely a rozhodně si nečiní nárok na to, aby byla lingvisticky adekvátním popisem nějaké české věty – kromě jiného proto, že by byla i popisem ne-vět *"Odporem cvičí Marie s aerobik"* atd.)

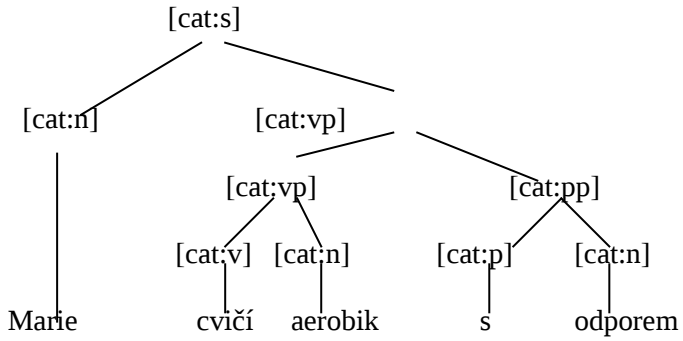
V pravidlech takové gramatiky se vyskytují dva druhy symbolů: terminály, tj. slova jazyka, a neterminály, které budeme pro účely tohoto příkladu pokládat za "jediné skutečné" syntaktické objekty. Veškerá informace, kterou o těchto objektech máme, je ale velmi skromná – víme pouze, o jaký druh objektu se jedná (že jde o větu, podstatné jméno, slovesnou frázi atd.). Přestože je z tohoto pohledu vše velmi triviální, přepišme nyní tuto gramatiku tak, že místo jednoduchých neterminálních symbolů použijeme sestavy rysů vyjadřující příslušnou informaci o kategorii objektu (kategorii – tj. název vlastnosti - budeme zkracovat obvyklým způsobem jako *cat*, hodnotu kategorie – původní neterminál - budeme psát malým písmenem). Dostaneme tak gramatiku s pravidly v následujícím tvaru:

Pravidla:

[*cat:s*]	→	[*cat:n*]	[*cat:vp*]	[*cat:n*]	→	Marie
[*cat:vp*]	→	[*cat:v*]	[*cat:np*]	[*cat:v*]	→	cvičí
[*cat:vp*]	→	[*cat:vp*]	[*cat:pp*]	[*cat:n*]	→	aerobik
[*cat:pp*]	→	[*cat:p*]	[*cat:n*]	[*cat:p*]	→	s
				[*cat:n*]	→	odporem

Taková gramatika sice používá struktury rysů jako neterminálních symbolů, stále však ještě není gramatikou unifikační – operaci unifikace vneseme až do její aplikace, tím, že při přepisování symbolů (rozvíjení neterminálů) nebudeme požadovat striktní identitu přepisovaného a přepisujícího neterminálu (jako je to obvyklé u standardních gramatik), ale pouze stanovíme podmínku, že výsledek unifikace přepisovaného a přepisujícího neterminálu je různý od $\perp$ (tj. že tyto dva neterminály jsou unifikovatelné). Všimněmě si, že takovový požadavek je rozumným rozšířením požadavku striktní formální identity neterminálů, neboť unifikace dvou popisů objektů znamená vlastně znalost faktu, že oba tyto popisy se týkají identického objektu (tj. identita, od které bylo formálně ustoupeno, zůstává zachována na intuitivním pozadí). V souvislosti s nahrazením identity neterminálů jejich unifikovatelností zavedeme navíc konvenci, že do stromu odvození budeme vždy zapisovat tu nejpodrobnější informaci, kterou jsme na daném místě schopni zapsat – konkrétně tedy budeme do uzlů stromu odvození vždy zapisovat výsledek unifikace přepisovaného a přepisujícího neterminálu (z hlediska stromové geometrie unifikaci informace přicházející "shora" a informace přicházející

“zdola”). Strom odvození příslušný příkladové větě "*Marie cvičí aerobik s odporem"* v takovéto gramatice vypadá takto:



V předchozím triviálním příkladě znamenají přitom identita a unifikace totéž: sestavy rysů tvořící neterminály jsou totiž velmi jednoduché. Tento příklad (tj. gramatiku) nyní rozšíříme tak, aby se rozdíl mezi identitou a unifikací stal zřejmým – rozdíl v tom, že identitu již nebude možné pro odvozování či analýzu řetězců pomocí takto rozšířené gramatiky použít, zatímco s unifikací bude vše "fungovat" správně. Použijeme při tom – opět jen pro ilustrační účely, bez valné lingvistické motivace – následujících faktů:

- podstatné jméno v podmětu a sloveso v přísudku věty se musejí shodovat v čísle – buď musí být jak podmět tak přísudek v singuláru, nebo musí být obě v plurálu
- v oznamovacím způsobu přítomného času (a předpokládejme, že naše gramatika platí jen pro věty, jejichž sloveso splňuje tyto charakteristiky) není potřeba vyjadřovat požadavky na shodu podmětu a přísudku ve jmenném rodě (jako je to např nutné v čase minulém – *“Jan přišeL”* vs. *“Marie přišLA”*)
- tvar *“Marie”* je tvarem prvního pádu podstatného jména ženského rodu, avšak informaci o tom, zda jde o singulár nebo plurál, nám tento tvar nepodává (tj. v zápisu informace o slově *"Marie"* musí chybět rys vyjadřující číslo)
- tvar *“cvičí”* je slovesným tvarem, na němž je patrna pouze osoba (třetí), čas (přítomný) a slovesný rod (činný), nikoliv však číslo ("*cvičí"* je shodně tvarem singuláru i plurálu) a jmenný rod (jako by tomu bylo např. u tvaru *“cvičilo”*)

Když znalost těchto faktů promítneme do neterminálů naší příkladové gramatiky, změní se její tvar na gramatiku s následujícími pravidly. Je velmi důležité si všimnout, že se počet pravidel zvětšil, a to proto, že je potřeba, aby se pravidla popisující strukturu slovesné fráze vyskytovala v gramatice dvakrát – tato pravidla totiž slouží – viz obrázek struktury níže - také k “přenosu” informace o shodě mezi podmětem a přísudkem, je tedy třeba je formulovat separátně pro shodu singulárovou a pro shodu plurálovou[2].

$$[cat{:}s] \rightarrow \begin{bmatrix} cat : n \\ num : sg \\ case : nom \end{bmatrix} \quad \begin{bmatrix} cat : vp \\ num : sg \end{bmatrix}$$

$$\begin{bmatrix} cat : n \\ gend : f \\ case : nom \end{bmatrix} \rightarrow \text{Marie}$$

$$[cat{:}s] \rightarrow \begin{bmatrix} cat : n \\ num : pl \\ case : nom \end{bmatrix} \quad \begin{bmatrix} cat : vp \\ num : pl \end{bmatrix}$$

$$\begin{bmatrix} cat = v \\ per = 3 \\ tns = prs \\ vce = act \end{bmatrix} \rightarrow \text{cvičí}$$

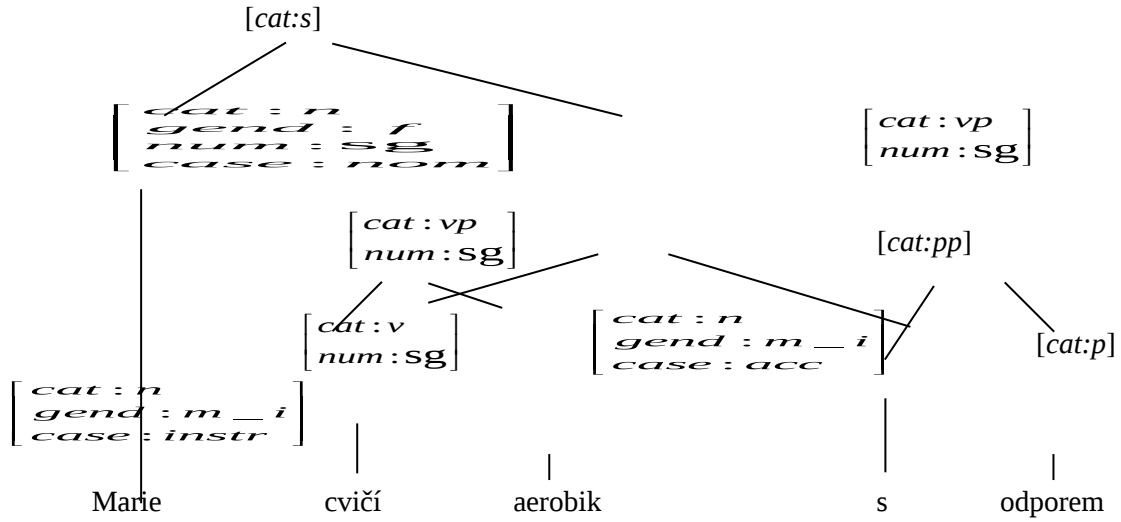
[2] Níže si ukážeme, jak je pomocí pokročilejších technik možné dosáhnout toho, aby byla shoda vyjádřena korektně a přitom počet pravidel nevzrostl. Zatím však takové prostředky nemáme zavedeny a nezbývá tedy, než se pro tuto chvíli smířit s narůstem počtu pravidel.

$\begin{bmatrix} cat:vp \\ num:sg \end{bmatrix} \rightarrow \begin{bmatrix} cat:v \\ num:sg \end{bmatrix} \; |cat:n| \qquad \begin{bmatrix} cat:n \\ gend:m\_i \\ case:acc \end{bmatrix} \rightarrow$ aerobik

$\begin{bmatrix} cat:vp \\ num:pl \end{bmatrix} \rightarrow \begin{bmatrix} cat:v \\ num:pl \end{bmatrix} \; |cat:n| \qquad [cat:p] \rightarrow$ s

$[cat:pp] \rightarrow [cat:p] \; [cat:n] \qquad \begin{bmatrix} cat:n \\ gend:m\_i \\ case:instr \end{bmatrix} \rightarrow$ odporem

Pokud chceme tuto gramatiku aplikovat na příkladovou větu, je okamžitě jasné, že při aplikaci pravidel nevystačíme s identitou symbolů a že je nutné užít unifikaci. Ta nám ovšem do výsledné struktury přinese neterminály, které se v takové formě v původní gramatice nevyskytovaly – totiž neterminály, které jsou rozšířením (pomocí unifikace) neterminálů z gramatiky.
Důležité je také, že příkladovou větu *"Marie cvičí aerobik s odporem"* je možné podle této gramatiky přiřadit dvě různé struktury. Jedna z nich, ta, ve které jsou slova *"Marie"* a *"cvičí"* chápána jako slova stojící v singuláru, vypadá následujícím způsobem (singulárová čtení jsou graficky zvýrazněna) :



Druhá možná struktura se od ní liší v tom, že slovům *"Marie"* a *"cvičí"* jsou v ní přiřazeny struktury obsahující rys *num:pl* (jde tedy o větu, kde se mluví o více Mariích, které všechny cvičí aerobik s odporem). Je samozřejmě důležité si všimnout toho, že unifikace (spolu s tvarem neterminálů v pravidlech gramatiky) zajišťuje, že nemůže dojít k "překřížení" – k tomu, aby jedno z těchto slov bylo v singuláru a druhé v plurálu; konkrétně je to zajištěno tím, že díky unifikaci užité pří aplikaci pravidel jsou slova *"Marie"* a *"cvičí"* (přesněji řečeno: příslušné preterminály) ve výsledné struktuře spojeny řetězcem neterminálů s identickou hodnotou rysu *num*. (Tato technika zajištění shody, obecně vztahů mezi "vzdálenými" neterminály se anglicky nazývá *"feature threading"*, český termín zatím neexistuje)

4.4.4. Zavedení struktury rysů a proměnné jako možných hodnot rysu

Jako možný typ hodnoty rysu v popisovaném formalismu jsme zatím připustili pouze identifikátory a čísla (viz str. XXX). Takovým hodnotám říkame atomické; k jejich výhodám patří, že operace sse strukturami rysů, jejichž hodnoty jsou takto jednoduché, nimi jsou přehledné a že je možné je také velmi jednoduše implementovat. Z takové jednoduchosti však vyplývá i jeden velmi zásadní nedostatek: pomocí atomických hodnot nelze přehledně popsat takové objekty, jejichž vlastnosti mají vlastní vnitřní strukturu.
Uveďme si nejprve jako odlehčující motivaci triviální příklad zcela mimo oblast lingvistiky. Představme si, že máme strukturou rysů popsat jídelní lístek slavnostního oběda[3]. Takový oběd se typicky skládá z více chodů - předpokládejme, že v našem případě to budou předkrm, polévka, hlavní chod a zákusek. Předkrm bude chřestový salát s bylinkovou zálivkou a obzvláště křupavými rohlíčky z pekárny "U císařovy pochoutky", polévka bude sýrová, jako hlavní jídlo zvolíme toufu à la bažant s rýží a jako desert krém z mascarpone šlehaného s borůvkami. Pokud budeme chtít takový jídelní lístek popsat "přirozeným" způsobem pomocí struktury rysů, zřejmě nevystačíme s atomickými hodnotami. Na základní úrovni totiž bude zřejmě vhodné uvést jen rysy *předkrm*, *polévka*, *hlavní_jídlo* a *zákusek*, nikoliv už ale jakékoliv další "vlastnosti" jídelního lístku. Přitom ovšem např. rys *předkrm* má (v našem příkladě) hodnotu, která by měla být strukturována - předkrm se skládá ze salátu a přílohy, a dokonce i tento salát je "dále vnitřně strukturován" - skládá se z chřestu a zálivky. Prostředky na vnitřní strukturaci hodnot nám přitom dosavadní formalismu, připouštějící pouze hodnoty atomické, nedává. Proto je vhodné tento formalismus rozšířit, konkrétně (v tuto chvíli) tak, že jako hodnotu rysu (tj. výraz stojící napravo od dvojtečky v zápisu *<název_vlastnosti> : <hodnota_vlastnosti>*) povolíme kromě identifikátorů (posloupností písmen, číslic a podtržítek, začínajících malým písmenem) a číslic také sestavy rysů. Zápis jídelního lístku z příkladu tedy bude nyní moci vypadat jako následující (správně tvořená) sestava rysů:

Je samozřejmé, že takové rozšíření sestav rysů nebylo zavedeno kvůli popisu jídelních lístků, zároveň je však právě na takovém - jinak profánním - příkladě dobře patrné, jaký je skutečný účel toho, aby byly jako hodnoty připuštěny i sestavy rysů: tímto účelem je těsně spolu sdružit ty rysy, které k sobě z nějakého důvodu patří, a moci takovouto související skupinu rysů i v popisu pojímat jako samostanou jednotku, oddělenou od rysů ostatních.
Důvodem, proč sdružovat rysy v syntaktickém popisu, je typicky ten fakt, že několik rysů spolu popisuje určitý jazykový jev. Typickým lingvistickým jevem, který téměř volá po takové sdružením rysů, je tak např. shoda mezi podmětem a přísudkem - takovou shodu na jedné straně pokládáme za jediný jazykový jev, který ale na druhé zasahuje více rysů: podmět a přísudek se (v češtině) shodují v osobě, čísle a (jmenném) rodě[4]. Např. ve větě "*Já, děvenka moje starostlivá, jsem se vypravila do města pro trochu toho šafránu"* se zájmeno *"já"* stojící v (holém) podmětu a slovesný tvar *"jsem se vypravila"* tvořící (holý) přísudek shodují v osobě (ta je u obou *první*), čísle (*singulár*) a jmenném rodě (*femininum*).
Nazveme-li sdružený rys pro shodu podmětu s přísudkem *s_v_agr* (z anglického *subject-verb agreement*), mohla by gramatika pro věty *"Josef cvičil aerobik s nadšením"* a *"Josefové cvičili aerobik s nadšením"* (které jsou, jak patrno, variacemi na příkladovou větu minulou, zvolenými tak, aby "důležitá" slova v nich obsažená nesla morfologicky pokud možno jednoznačné rysy) vypadat následujícím způsobem[5]:

[3] Například pohoštění pro autory skript, z nichž jsme studovali na úspěšnou rigorosní zkoušku.
[4] Že se jedná o rod jmenný, je potřeba zdůraznit proto, aby byl odlišen od rodu slovesného (aktivum vs. pasívum).
[5] Jak vidno, není jen tato věta variací na větu předchozí, ale je - a to především - tato gramatika variací na gramatiku předchozí.

$[cat{:}s] \rightarrow \begin{bmatrix} cat:n \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \\ case:nom \end{bmatrix} \begin{bmatrix} cat:vp \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix}$

$[cat{:}s] \rightarrow \begin{bmatrix} cat:n \\ s\_v\_agr:\begin{bmatrix} num:pl \\ gend:m\_a \\ per:3 \end{bmatrix} \\ case:nom \end{bmatrix} \begin{bmatrix} cat:vp \\ s\_v\_agr:\begin{bmatrix} num:pl \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix}$

$\begin{bmatrix} cat:vp \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix} cat:v \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix} \quad \lfloor cat:n \rfloor$

$\begin{bmatrix} cat:vp \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix} \rightarrow \begin{bmatrix} cat:v \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \end{bmatrix} \quad \lfloor cat:n \rfloor$

$[cat{:}pp] \rightarrow [cat{:}\ p] \quad [cat{:}n]$

$\begin{bmatrix} cat:n \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \\ case:nom \end{bmatrix} \rightarrow$ Josef $\qquad \begin{bmatrix} cat:n \\ s\_v\_agr:\begin{bmatrix} num:pl \\ gend:m\_a \\ per:3 \end{bmatrix} \\ case:nom \end{bmatrix} \rightarrow$ Josefové

$\begin{bmatrix} cat:v \\ s\_v\_agr:\begin{bmatrix} num:sg \\ gend:m\_a \\ per:3 \end{bmatrix} \\ tns:prs \\ vce:act \end{bmatrix} \rightarrow$ cvičil $\qquad \begin{bmatrix} cat:v \\ s\_v\_agr:\begin{bmatrix} num:pl \\ gend:m\_a \\ per:3 \end{bmatrix} \\ tns:prs \\ vce:act \end{bmatrix} \rightarrow$ cvičili

$\begin{bmatrix} cat:n \\ gend:m\_i \\ case:acc \end{bmatrix} \rightarrow$ aerobik $\quad [cat{:}p] \rightarrow$ s $\quad \begin{bmatrix} cat:n \\ gend:m\_i \\ case:instr \end{bmatrix} \rightarrow$ odporem

Tato gramatika je zapsána sice tak, že rysy a atomickými hodnotami relevantními pro shodu podmětu s přísudkem jsou v ní sdruženy do rysu s komplexní hodnotou, avšak přínos takového zápisu je - zdá se alespoň - nulový či přímo záporný. Neušetřili jsme si totiž žádné psaní, naopak, museli jsme psát ještě více. Především se ale nepovedlo vyjádřit lingvistický fakt, že podmět a přísudek se shodují v patřičných kategoriích bez toho, abychom tyto hodnoty museli vždy explicitně vypisovat, což má za následek, že oproti původní gramatice s jednoduchými symboly (tj. gramatice bez použití sestav rysů) vzrostl počet pravidel.

Pokud si předchozí odstavec promyslíme, zjistíme, že problém je zřejmě v tom, že formalismus nás nutí používat pouze konstanty - proto musíme hodnoty vždy vypisovat (bez ohledu na to, zda jde o hodnoty atomické nebo komplexní). Např pokud chceme, aby byly hodnoty čísla na dvou místech pravidla stejné (tj. aby se např. podmět a přísudek spolu v čísle shodovaly), musíme pravidla dvě, jedno, kde na obou místech uvedeme singulár, a druhé, kde na obou místech uvedeme plurál. Proto, abychom si tuto práci mohli ušetřit, a především proto, abychom byli schopni vyjádřit relevantní lingvistické zobecnění, že podmět a přísudek se (prostě) shodují v osobě, čísle a jemnném rodě, ať už jsou hodnoty těchto rysů jakékoliv, zavedeme další možný typ hodnoty rysu: proměnnou. Značit budeme takovou hodnotu jako (malé) přirozené číslo uzavřené mezi svislé čáry zleva a zprava, např. tedy |1|, |2|, |15| jsou proměnné[6].

Dva výskyty téže proměnné na místě hodnoty dvou rysů v jedné sestavě rysů nebo v jednom pravidle pak budeme interpretovat tak, že (možná) neznáme, jaké konkrétní hodnoty tyto dva rysy mají, ale víme, že jejich hodnoty jsou shodné (používáme proměnnou tedy zcela standardně jako kde koliv jinde v matematice). Důležité je si uvědomit, že právě uvedené tvrzení je zřejmým příkladem "informace o informaci" (možná nevíme, jaká konkrétní informace na těch dvou místech stojí, víme ale, že na obou místech je to informace identická). V tomto smyslu je dvojí (či vícenásobný) výskyt proměnné ve struktuře či pravidle "naznačenou" či "skrytou" unifikací[7] informací stojících jako hodnoty na příslušných místech.

[6] V literatuře se běžně používá notace, kdy je číslo uzavřeno v "krabičce". My jsme tuto notaci nezavadli jen proto, že je typograficky náročnější.

[7] Dvojí výskyt proměnné je tedy - v tomto ohledu - velmi podobné psaní zlomků: dva výskyty jedné proměnné je "naznačená unifikace", zlomek je (jak známo ze základní školy) "naznačené dělení".

Důležité přitom je, že zavedení proměnných jako hodnot rysů umožňuje svyjadřovat identitu rysů bez toho, že bychom museli vypisovat konkrétní hodnoty rysů (konstanty). To potom dovoluje následující zápis gramatiky, kterou lze vytvořit všechny naše předchozí věty o *"Josefovi/Josefech"* a *"Marii/Mariích"*. Tato gramatika má stejný počet pravidel (pokud v gramatikách nepočítáme pravidla přepisující preterminály na neterminály) jako původní gramatika s jednoduchými neterminálními symboly, na rozdíl od této původní gramatiky však vyjadřuje i shodu podmětu s přísudkem (uvádíme jen pravidla neobsahující terminální symboly - slova).

$$[cat{:}s] \rightarrow \begin{bmatrix} cat : n \\ s\_v\_agr : \boxed{1} \\ case : nom \end{bmatrix} \quad \begin{bmatrix} cat : vp \\ s\_v\_agr : \boxed{1} \end{bmatrix}$$

$$\begin{bmatrix} cat : vp \\ s\_v\_agr : \boxed{1} \end{bmatrix} \rightarrow \begin{bmatrix} cat : v \\ s\_v\_agr : \boxed{1} \end{bmatrix} \quad \boxed{cat : n}$$

[*cat:pp*] → [*cat: p*] [*cat:n*]

Výšeuvedené rozšíření typů hodnot rysů si automaticky vynutí i některé změny (rozšíření) další. NOTACE pokud je známa hodnota a navíc je tam proměnná.

Dále je potřeba zmínit, fakt, že i nadále sice samozřejmě zůstává platit omezení, že <*název_vlastnosti*> musí být identifikátor, ale je nutné nyní přesněji formulovat omezení na výskyt identických názvů vlastnosti. Obecně se totiž nyní může nastat, že jeden název vlastnosti se může vyskytnout v platné komplexní sestavě rysů víckrát, vždy jako nejvíce vnější název vlastnosti v určité (pod)sestavě rysů. Jako ještě jeden nelingvistický příklad si uveďme ve formalismu sestav rysů zapsanou personálně-organizační strukturu nějakého koncernu.

$$\left[\begin{array}{l} \text{ředitel : jan\_novák} \\ \text{první\_závod : } \left[\begin{array}{l} \text{ředitel : honza\_novák} \\ \text{první\_továrna : } [\text{ředitel : honzík\_novák}] \\ \text{druhá\_továrna : } [\text{ředitel : honcek\_nováků}] \end{array}\right] \\ \text{druhý\_závod : } \left[\begin{array}{l} \text{ředitel : jenda\_novák} \\ \text{první\_továrna : } [\text{ředitel : jeník\_novák}] \\ \text{druhá\_továrna : } [\text{ředitel : janek\_novákojc}] \end{array}\right] \\ \text{samostatný\_zahr\_dodavatel : } [\text{ředitel : john\_newman}] \end{array}\right]$$

Jak je z příkladu zřejmé, vyskytují se názvy vlastností *ředitel*, *první_továrna* a *druhá_továrna* v této komplexní struktuře víckrát, ke konfliktu však zřejmě nedochází, a to proto, že každý takový název vlastnosti je jediný (unikátní) ve své doméně platnosti, kterou je v případě struktur rysů nejbližší názvu vlastnosti bezprostředně nadřazená struktura rysů ("nejbližší nadřazené hranaté závorky"). Omezení výskytu názvů vlastností je tedy potřeba přeformulovat tak, že žádný <*název_vlastnosti*> se v sestavě rysů nesmí vyskytnout více než jednou ve stejné doméně platnosti. (Toto znění je přitom zřejmě konzervativním rozšířením formulace minulé, která byla speciálním případem pro situaci, že existuje pouze jediná doména platnosti).

Alternativní notace: DAGs. Ne každá SR je vyjádřitelná jako DAG. Často je proto SR definována méně obecně.

Dále je zřejmě potřeba rozšířit (zevšeobecnit) definici unifikace.

Ukázka toho, že unifikace dvou DAGů nemusí být DAG - implementační problém z toho plynoucí.

TO BE WRITTEN

4.4.5. Několik algebraických vlastností

V minulých kapitolách jsme zavedli sestavy rysů a operaci na nich (unifikaci) a ukázali jsme si, jak je možné s pomocí takového aparátu vytvářet gramatiky. Takové gramatiky je však nejen potřeba vytvářet, ale také (pro praxi) implementovat a (pro teorii) studovat co do formálních vlastností. Jak k

jednomu tak k druhému je ovšem potřeba především postavit vše - a zejména unifikaci - na formálnější základ. To je obsahem této kapitoly.
Představme si nyní, že nyní pomocí sestav rysů popisujeme množinu Ω objektů, z nichž každý má konečný počet vlastností. Představme si, že v libovolné sestavě rysů nepoužijeme vlastnost, která by nebyla přítomna alespoň v jednom objektu z množiny Ω (tj. omezíme se jen na ty vlastnosti, které jsou skutečně přítomny na nějakých – alespoň jednom – objektech z Ω; takové omezení je zřejmě rozumné, např. i proto, že jedině s tímto omezením jsou sestavy rysů schopny efektivně rozlišovat jednotlivé objekty). V takovém případě zřejmě platí, že i počet všech možných (a navzájem různých) sestav rysů, které s různou mírou informace správně (tj. pravdivě, odpovídajícím způsobem) popisují nějaký objekt z množiny Ω, je konečný počet. Přidejme nyní k množině všech takových sestav ještě dvě sestavy speciální, totiž sestavu ⊥ zavedenou výše a sestavu, která nenese žádnou informaci ("prázdnou" sestavu – dosud jsme o ní nemluvili, je však zřejmé, že její zavedení dává smysl, taková sestava je totiž formálním odrazem situace, kdy o popisovaném objektu nevíme nic). Takovou sestavu bychom mohli značit např. [] (tj. jen "sestavové závorky", bez jakékoliv vlastnosti v nich uzavřené), my však pro ni zavedeme značku T (důvod pro takové značení – jakož i pro značení ⊥ - vyplyne níže). Je zřejmé, že sestavy ⊥ a T jsou přidány "rozumně" či "organicky", že rozšiřují základní množinu sestav přirozeným způsobem, jako specální případy sestav popisujících informaci o vlastnostech objektů z množiny: T je sestavou, která o objektu dává nulovou informaci, ⊥ je sestavou která poskytuje přespříliš mnoho informace. Označme tuto množinu sestav rysů (tj. po rozšíření o ⊥ a T) nad množinou objektů Ω jako M.
V následujícím přehledu vlastností algebraické struktury (M, ∪), tj. struktury tvořené množinou M a operací ∪ na M definovanou, budeme všechna tvrzení uvádět bez formálních důkazů, půjde však vždy o tvrzení z názoru naprosto zřejmá. Formálněji orientovaného čtenáře odkazujeme na práci (Carpenter, 19XX).

Jako první bod bude důležité si uvědomit, jaký je vztah mezi strukturou rysů jakožto prvkem množiny M a množinami popisovaných objektů (tj. množinami prvků z Ω). Platí totiž, že každá sestava rysů σ ∈ M je popisem nějaké (určité) podmnožiny množiny Ω – takovou podmnožinu nazýváme *denotací sestavy rysů σ* a budeme ji v následujících odstavcích značit Den(σ). Den(σ) je tedy taková množina objektů, které (z logického hlediska) splňují podmínky kladené σ na vlastnosti těchto objektů. Platí přitom následující vztahy:

$$Den(T) = \Omega$$

Denotace T, tj. sestavy rysů, která nenese žádnou informaci, je celá množina Ω. Plyne to z toho, že sestava rysů T nijak nedefinuje (tj. ani neomezuje) vlastnosti objektu/objektů, jejichž „popisem" by měla být – T je tedy popisem libovolného objektu z Ω.

$$Den(\perp) = Ø$$

Denotace ⊥, tj. sestavy rysů, která nese "příliš mnoho" informace, je prázdná množina. Plyne to z toho, že sestava rysů ⊥ je "vnitřně sporná", tj. popisuje vlastnosti, které jsou samy se sebou ve sporu. Je samozřejmé, že takové vlastnosti nemůže mít žádný objekt z Ω, tj. že množina objektů majících takové vlastnosti je prázdná.

$$\forall \sigma,\tau \in M : Den(\sigma \cup \tau) = Den(\sigma) \cap Den(\tau)$$

Toto tvrzení říká, že množina všech objektů, které vyhovují popisu vzniklému unifikací dvou popisů σ a τ, je rovna průniku množin objektů, které odpovídají každému z popisů σ,τ vzatých zvlášť. V podstatě jde tedy jen o formální vyjádření zcela zřejmého faktu, že objekty popsané unifikací dvou popisů jsou právě ty objekty, které odpovídají oběma popisům. Toto tvrzení je také zřejmě v dobré shodě s intuitivní představou, že "čím podrobnější je popis (tj. čím více informací), tím méně je objektů, které tomuto popisu (těmto informacím vzatým najednou) odpovídají".

Uvědomme si dále na tomto místě, že výše uvedenou formulaci, že každá sestava rysů σ je popisem nějaké (určité) podmnožiny Ω nelze obrátit – tj. že neplatí, že každá podmnožina množiny Ω je posána nějakou (určitou) sestavou rysů. Ukažme si to na (proti)příkladě.

Příklad: Vezměme za Ω množinu, jejímiž prvky jsou (právě) tří objekty O1, O2 a O3, o nichž máme jen tu informaci, že O1 má bílou barvu, O2 ma zelenou barvu a O3 má černou barvu. Každému objektu O1, O2 a O3 pak (po řadě) odpovídá jedna sestava rysů

$$\lfloor barva : bílá \rfloor \qquad \lfloor barva : zelená \rfloor \qquad \lfloor barva : černá \rfloor$$

Neexistuje však sestava rysů, která by popisovala libovolnou množinu obsahující právě dva objekty (např. bílý a zelený). Všimněmě si, že se jedná o situaci velmi podobnou té, kdy se nám nepovedlo popsat slovo *"kniha"* kvůli tomu, že jsme neuměli přesně určit jeho pád, současně jsme jej však nechtěli popsat tak, jako by šlo o slovo, jehož pád vůbec neznáme. Tento problém tedy – jak slíbeno - vyřešíme níže.

Podívejme se nyní na další vlastnosti algebraické struktury dané množinou M a operací $\cup$ na M definovanou (unifikací). Platí totiž následující tvrzení.

Operace $\cup$ je uzavřená na M, tj. platí

$$\forall \sigma,\tau \in M : \sigma \cup \tau \in M.$$

Náš názor se zde bude velmi silně opírat o pojem denotace zavedený výše, konkrétně o to, že každou sestavu rysů σ je možno chápat jako popis množiny Den(σ). Uvedené tvrzení pak platí proto, že M je množina popisů všech možných podmnožin určité množiny objektů (včetně popisu vnitřně nekonsistentního, značeného $\perp$, který je popisem prázdné množiny), a jsou-li tedy σ,τ popisy množin objektů Den(σ) a Den(τ) (po řadě), pak i jejich kombinace $\sigma \cup \tau$ je zřejmě popisem množiny objektů Den($\sigma \cup \tau$) (případně popisem nekonsistentním neboli popisem prázdné množiny objektů), a je tedy $\sigma \cup \tau$ prvkem množiny M popisů všech možných podmnožin uvedené množiny objektů.

Operace $\cup$ je na M asociativní, tj. platí

$$\forall \sigma,\tau,\upsilon \in M : (\sigma \cup \tau) \cup \upsilon = \sigma \cup (\tau \cup \upsilon)$$

Z názoru je toto tvrzení zřejmé proto, že existují-li tři popisy objektů (nikoliv nutně navzájem kompatibilní), pak je zřejmě výsledek jejich kombinace stejný, ať již jsou zkombinovány tak, že nejprve jsou zkombinovány první dva popisy a pak teprve k nim přidán popis třetí, nebo pokud je první popis zkombinován s výsledkem kombinace popisu druhého a třetího – a to vše i v případě, že výsledkem (či libovolným mezivýsledkem) je popis vnitřně sporný (v zavedené notaci značený $\perp$).

Operace $\cup$ je na M komutativní, tj. platí

$$\forall \sigma,\tau \in M : \sigma \cup \tau = \tau \cup \sigma$$

Z názoru je toto tvrzení zřejmé proto, že existují-li dva popisy téhož objektu (nikoliv nutně navzájem kompatibilní), pak je zřejmě výsledek jejich kombinace stejný, ať již jsou rysy v nich obsažené zkombinovány v libovolném pořadí, opět i v případě, že výsledkem je popis vnitřně sporný (v zavedené notaci značený $\perp$).

Operace $\cup$ je na M idempotentní, tj. platí

$$\forall \sigma \in M : \sigma \cup \sigma = \sigma$$

Z názoru je toto tvrzení zřejmé proto, že pokud k informaci σ přidáme tutéž informaci (tj. opět σ) , nese výsledná kombinace přesně stejné množství informace jako je obsaženo v σ (přidali jsme informaci identickou, již známou, tj. stávající informaci jsme o nic nerozšířili).

Další zajímavou a významnou vlastností struktury (M, $\cup$) je existence „jednotkového“ prvku (kterým je T), což plyne z toho, že platí

$$\forall \sigma \in M : \sigma \cup T = T \cup \sigma = \sigma$$

(prvek T se nazývá „jednotkový“ kvůli analogii s operací násobení čísel, kde také platí, že libovolné číslo po znásobení číslem 1 zůstane samo sebou). Z názoru je zřejmé, že uvedené tvrzení platí proto, že unifikace s T vlastně znamená přidání prázdné informace (tj. nepřidání žádné informace), čímž se tedy evidentně množství informace nemůže změnit.

Analogická existenci „jednotkového“ prvku je existence „nulového“ prvku (kterým je $\perp$), což plyne z toho, že platí

$$\forall \sigma \in M : \sigma \cup \perp = \perp \cup \sigma = \perp$$

(prvek $\perp$ se nazývá „nulový“ opět kvůli analogii s operací násobení čísel – výsledek násobení libovolného čísla nulou je nula). Z názoru je zřejmé, že uvedené tvrzení platí proto, že unifikací $\perp$ s libovolnou strukturou rysů nelze dosáhnout toho, aby ve výsledné struktuře nebylo „příliš mnoho“ informace – tak „mnoho“, že tato struktura je vnitřně nekonsistentní. (Takového výsledku by šlo dosáhnout „ubráním“ informace, ne však – jakýmkoliv – jejím přidáním.)

Dále je vhodné (a intuitivní) si uvědomit, že v určitých případech je možné srovnávat dvě sestavy rysů z množiny M podle množství informace, kterou každá z nich obsahuje. Tak např. je asi rozumné (tj. odpovídající naší intuici), že ze sestav σ a τ definovaných následujícím způsobem

$$\sigma = \begin{bmatrix} \textit{číslo} : \textit{jednotné} \\ \textit{pád} : 4 \end{bmatrix} \qquad \tau = \begin{bmatrix} \textit{číslo} : \textit{jednotné} \\ \textit{rod} : \textit{ženský} \\ \textit{slovní\_druh} : \textit{adjektivum} \\ \textit{stupeň} : 3 \\ \textit{pád} : 4 \end{bmatrix}$$

obsahuje sestava σ striktně méně informace a sestava τ obsahuje informace více. Na druhé straně je ale zřejmé, že provádět taková srovnání množství informace obsažené ve dvou strukturách není vždy možné, např. není (rozumně) možné srovnávat množství informace obsažené v následujících dvou strukturách.

$$\begin{bmatrix} \textit{číslo} : \textit{jednotné} \\ \textit{pád} : 4 \end{bmatrix} \qquad \begin{bmatrix} \textit{rod} : \textit{ženský} \\ \textit{slovní\_druh} : \textit{adjektivum} \\ \textit{stupeň} : 3 \\ \textit{pád} : 4 \end{bmatrix}$$

Důležité přitom je, že množství informace v sestavách rysů nelze (lépe: není rozumné) srovnávat proto, že každá z nich obsahuje alespoň částečně jinou informaci – sestava na levé straně obsahuje informaci o čísle, která není obsažena v sestavě na straně pravé, a ta zase obsahuje infomaci o rodu, slovním druhu a stupni, která se nevskytuje v sestavě na straně levé (z čehož je mj. vidět, že

porovnávání informačního obsahu sestav se řídí takříkajíc kvalitou, nikoliv kvantitou – alespoň ne kvantitou měřenou počtem rysů).
Další situace, v níž sestavy nelze porovnávat podle množství informace v nich obsažené, nastává tehdy, když sestavy obsahují informaci navzájem konfliktní. Např tedy sestavy

$$\begin{bmatrix} \check{c}\acute{i}slo : mno\check{z}n\acute{e} \\ p\acute{a}d : 1 \end{bmatrix} \qquad\qquad \begin{bmatrix} \check{c}\acute{i}slo : mno\check{z}n\acute{e} \\ p\acute{a}d : 5 \end{bmatrix}$$

zřejmě nelze (co do množství obsažené informace) porovnávát.

Jinými slovy, dvě sestavy lze porovnávat jenom tehdy, když se informace obsažené v každé z nich „plně překrývají“ – množina rysů (jak co do jmen vlastností, tak co do jejich hodnot) obsažená v jedné ze sestav je plně obsažena také v sestavě druhé.

DOPLNIT: DAG !

Důležitým příkladem dvou sestav rysů, které lze porovnat co do obsahu informace, je dvojice sestav rysů

$$\begin{bmatrix} podmet : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \\ přísudek : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} f : \left| \begin{matrix} osoba : 2 \\ rod : fem \end{matrix} \right| \\ g : \left| \right| \end{bmatrix}$$

Je potřeba si uvědomit, že sestava nalevo obsahuje striktně méně informace než sestava napravo: levá sestava říká, že hodnotou rysu *f* je sestava $\begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix}$ a že hodnotou rysu *g* je sestava $\begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix}$, sestava napravo obsahuje tutéž informaci a navíc informaci o tom, že sestavy, které jsou hodnotami rysů *f* a *g*, jsou identické (že je to tedy vlastně sestava jediná). Tento rozdíl můžeme samozřejmě vyjádřit i formálně, totiž unifikační rovností

$$\begin{bmatrix} podmet : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \\ přísudek : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \end{bmatrix} \cup \begin{bmatrix} podmet : |1| \\ přísudek : |1| \end{bmatrix} = \begin{bmatrix} f : \left| \begin{matrix} osoba : 2 \\ rod : fem \end{matrix} \right| \\ g : \left| \right| \end{bmatrix}$$

ze které fakt, že v levé z výše uvedených sestav je striktně méně informace než v sestavě pravé, vyplývá zvláště zřetelně.

Rozdílnost mezi těmito dvěma si dále lze zvláště dobře uvědomit, pokud si promyslíme, jak se každá z těchto sestav chová při unifikaci např. se sestavou $\left[podmet : \left[číslo : pl \right] \right]$. Výsledky této unifikace jsou totiž rozdílné, v prvním případě dojde jen ke změně hodnoty rysu *podmět*, ve druhém se samozřejmě změní hodnota obou rysů *podmět i přísudek* (prostě proto, že oba tyto rysy mají identickou hodnotu). Konkrétně vypadají tyto unifikace následujícím způsobem.

$$\begin{bmatrix} podmet : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \\ přísudek : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \end{bmatrix} \cup \left[podmet : \left[číslo : pl \right] \right] = \begin{bmatrix} podmet : \begin{bmatrix} osoba : 2 \\ rod : fem \\ číslo : pl \end{bmatrix} \\ přísudek : \begin{bmatrix} osoba : 2 \\ rod : fem \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} podmet : \left| \begin{matrix} osoba : 2 \\ rod : fem \end{matrix} \right| \\ přísudek : \left| \right| \end{bmatrix} \cup \left[podmet : \left[číslo : pl \right] \right] = \begin{bmatrix} f : \left| \begin{matrix} osoba : 2 \\ rod : fem \\ číslo : pl \end{matrix} \right| \\ g : \left| \right| \end{bmatrix}$$

Důležité zde je, že v prvním z případů se nyní hodnoty rysů *podmět* a *přísudek* od sebe liší, zatímco v druhém jsou (samozřejmě) i nadále identické.

Vzhledem k tomu všemu je zřejmě vhodné na množině M zavést reflexivní částečné uspořádání odrážející množství informace v jednotlivých sestavách rysů , tj. relaci, kterou budeme značit

symbolem $\subseteq$ (který si – z dobrých důvodů – opět vypůjčíme ze symboliky množinové) a která má následující vlastnosti:

$\subseteq$ je reflexivní, tj. platí $\forall \sigma \in M : \quad \sigma \subseteq \sigma$
$\subseteq$ je antisymetrická, tj. platí $\forall \sigma,\tau \in M : \quad \sigma \subseteq \tau \,\&\, \tau \subseteq \sigma \Rightarrow \sigma = \tau$
$\subseteq$ je tranzitivní, tj. platí $\forall \sigma,\tau,\upsilon \in M : \quad \sigma \subseteq \tau \,\&\, \tau \subseteq \upsilon \Rightarrow \sigma \subseteq \upsilon$

Pokud pro dvě sestavy rysů σ,τ platí $\sigma \subseteq \tau$, budeme říkat, že σ je méně informativní (menší) než τ.

Je dobré si na tomto místě uvědomit, že operace $\cup$ (unifikace) a relace $\subseteq$ (uspořádání podle informativity) spolu velmi úzce souvisejí. I když je jejich intuitivní obsah zřejmě rozdílný, jsou natolik svázány, že lze jednu definovat pomocí druhé. Platí totiž následující vztah logické ekvivalence vyjadřující zaměnitelnost definicí unifikace a uspořádání podle informativity:

$\forall \sigma,\tau \in M : \quad \tau \cup \sigma = \tau \iff \sigma \subseteq \tau$

Tento vztah opět uvádíme bez (formálního) důkazu (ten lze nalézt v literatuře), i on však patří ke tvrzením zřejmým z názoru. Pokud platí levá strana ekvivalence, totiž to, že přidáním informace σ k informaci τ se informace τ nijak nerozšíří, znamená to, že informace v σ je již zcela obsažena v informaci τ, neboli že σ je méně informativní (obsahuje méně informace) než τ, což formálně zapisujeme právě jako $\sigma \subseteq \tau$, tj. jako pravou stranu ekvivalence. Pokud naopak platí pravá strana ekvivalence, tj. σ je méně informativní než τ, pak platí levá strana téměř triviálně – pokud přidáme k informaci τ informaci σ, která je v τ již obsažena, pak se tím množství informace v τ zřejmě nezmění, což je přesně obsah fomálního zápisu na levé straně. Jako speciální případy tohoto tvrzení pak platí zřejmá tvrzení o minimálním a maximálním prvku vzhledem k uspořádání $\subseteq$

$\bot$ je maximální prvek vzhledem k $\subseteq$, formálně $\forall \sigma \in M : \sigma \subseteq \bot$

Toto tvrzení, které říká, že $\bot$ obsahuje více informace než kterákoliv jiná sestava rysů, je triviální – sestava $\bot$ je definována právě touto vlastností (viz též výše).

T je minimální prvek vzhledem k $\subseteq$, formálně $\forall \sigma \in M : \quad \mathsf{T} \subseteq \sigma$

Toto tvrzení je opět triviální, neboť sestava T je definována právě tou vlastností, že je v ní obsaženo méně informace než v jakékoliv jiné sestavě rysů (zcela přesně: v T není obsažena informace vůbec žádná).

Na samý závěr této kapitoly ještě podotkněme dvě věci:
1. že je rozumné zavést i speciální značení pro negaci relace $\subseteq$. Pokud pro dvě sestavy rysů σ,τ neplatí relace $\sigma \subseteq \tau$, a to ať už proto, že platí $\tau \subseteq \sigma$ (přitom $\tau \neq \sigma$), nebo proto, že σ a τ jsou co do obsaženého množství informace v našem smyslu neporovnatelné, budeme tento fakt zapisovat jako $\sigma \not\subset \tau$. Například tedy platí následující vztahy:

$$\begin{bmatrix} slovní\_druh : zájmeno \\ osoba : 2 \\ číslo : množné \end{bmatrix} \not\subset \begin{bmatrix} osoba : 2 \\ číslo : množné \end{bmatrix}$$

$$\begin{bmatrix} pád : 1 \end{bmatrix} \not\subset \begin{bmatrix} slovní\_druh : zájmeno \\ osoba : 2 \\ číslo : množné \end{bmatrix}$$

$$\begin{bmatrix} osoba : 2 \\ číslo : množné \end{bmatrix} \not\subset \begin{bmatrix} slovní\_druh : zájmeno \\ osoba : 3 \\ číslo : množné \end{bmatrix}$$

2. Že jsme kvůli jednoduchosti v celé kapitole používali jen takové příklady sestav rysů, v nichž hodnoty rysů byly vždy atomické. Formální tvrzení, která jsme uvedli, však platí samozřejmě i obecně, což ostatně by se dalo ukázat i na složitějších příkladech - takových, kde hodnotami rysů jsou také sestavy rysů nebo proměnné.

4.4.6. Rozšíření formalismu o negaci a disjunkci

motivace
notace
unifikace takových struktur, spolu s uspořádáním vzniká svaz
unifikace jako konjunkce, dopracování implikace, ekvivalence aj.
vzájemná distributivita konjunkce (unifikace) a disjunkce, de Morganovy vztahy atd.
implementační problémy

TO BE WRITTEN

4.4.7. Rozšíření formalismu o typování

TO BE WRITTEN

4.4.8. Závěr
další rozšíření – formální: funkční/relační závislosti

- lingvistická: principy
- speciální rozšíření pro slovník (jehož popis už ale přísně vzato překračuje rámec unifikačních gramatik)

TO BE WRITTEN

4.5. Zavedení struktury rysů a proměnné jako možných hodnot rysu}
motivace zavedení struktury rysů jako hodnoty rysu
rozšíření unifikace
motivace zavedení proměnných jako hodnot rysů
sdílení proměnné jako naznačená unifikace
TO BE WRITTEN

4.6. Několik algebraických vlastností

4.7. FUNKČNÍ GENERATIVNÍ POPIS ČEŠTINY

Funkční generativní popis češtiny (FGD, Functional Generative Description), navazující na tradice pražské lingvistické školy, se rozvíjí od 60. let 20. století (Sgall (1967), Sgall et al (1986), Sgall (1992) a Hajičová et al (2000)). Zde uvádíme jen velmi stručný přehled, více viz Hajičová et al (2002).

FGD je závislostní typ formalismu, který byl navržen pro účely teoretického popisu struktury českých vět pomocí generativní procedury. Velmi dobře se osvědčuje i jako podklad pro analytické procedury.[8]

Základní charakteristikou FGD je **stratifikační přístup** k popisu jazyka – popis jazyka je rozčleněn do několika rovin. Každá z rovin je množinou zápisů vět, každá má svou syntax (různé elementární jednotky se skládají v jednotky komplexní – jsou ve **vztahu kompozičním** (vztah C)).

[8] FGD slouží jako teoretický základ při budování Pražského závislostního korpusu (PDT, Prague Dependency Treebank, http://..), tedy pro analýzu vět na dvou rovinách větné stavby.

Nejvyšší rovina odpovídá významovému plánu jazyka, nejnižší plánu výrazovému. Jednotky jednotlivých sousedních rovin jsou ve vzájemném **vztahu formy a funkce – vztahu reprezentace** (vztah R; jednotka vyšší roviny je funkcí jednotky nižší roviny, jednotka nižší roviny je její formou).

V "klasické" verzi FGD (Sgall et al (1986)) se rozlišují dvě roviny syntaxe – **rovinu hloubkové struktury,** která odpovídá jazykovému významu (také tektogramatická či podkladová rovina (TR); v anglické terminologii *level of underlying representation* nebo *tectogrammatical level*) a **rovinu** jeho **povrchová realizace** (také tzv. větně členská rovina (VČR); anglicky *level of surface structure*). Ačkoliv teoretická opodstatněnost roviny povrchové syntaxe je v novějších verzích FGD zpochybněna (Sgall (1992)), je výhodné tuto rovinu (nebo nějakou její obdobu)[9] využívat alespoň jako pomocnou rovinu při aplikovaných úkolech NLP. Navíc představuje dobrý základ pro další výklad, neboť povrchová syntax (tj. větná stavba) se vyučuje na českých základních i středních školách.

FGD dále pracuje s nižšími rovinami jazykového popisu – s rovinou **morfematickou** (MR; *morphemic level*) a rovinou **fonologickou** (FR; *phonological level*).[10]

n = 4	———————————	rovina podkladové reprezentace
n = 3	———————————	rovina povrchové syntaxe
n = 2	———————————	morfematická rovina
n = 1	———————————	fonologická rovina

Roviny popisu jazyka (pro automatické systémy)

FGD využívá **závislostního formalismu** – na tektogramatické a na povrchové rovině je věta zachycena jako závislostní strom, což je spojitý acyklický orientovaný graf s kořenem, kde do každého uzlu kromě kořene vede právě jedna hrana. Hrany stromu reprezentují vztah (syntaktické) závislosti – za závislý člen se považuje ten, který lze vynechat, aniž by věta ztratila (syntaktickou) správnost.[11] Uzly stromu jsou lineárně uspořádány, jejich pořadí odpovídá (hloubkovému nebo povrchovému) slovosledu. K popisu věty na nižších rovinách už není potřeba stromová struktura, stačí řetězec.

4.7.1. Základní charakteristika rovin

Fonologická rovina odpovídá výrazovému plánu jazyka. Rozlišuje pojem hlásky (dané způsobem tvoření) a pojem fonému. Fonémy jsou nejmenší významotvorné jednotky daného jazyka (dvě různé hlásky mohou označovat jediný foném – pokud v češtině neexistuje slovo, které by tyto hlásky rozlišovaly, jako např. hlásky *n* a *ŋ* (*ven – venku*)). Řetězy fonémů se nazývají morfy.

Funkcí morfů jsou morfémy, které popisuje **morfematická rovina**. Rozlišují se morfémy lexikální (kmeny slov – např. lexikální morfém pro slovo *matka* se skládá z morfů (variant) *matk-*, *matc-*, *matč-* a *matek-*; patří sem i odvozovací morfémy – např. předpony (morfy) *od-* a *ode-* tvoří jediný lexikální morfém) a morfémy gramatické (typickým příkladem jsou pádové (deklinační) koncovky substantiv).

Rovinu povrchové syntaxe lze brát jako souhrn čistě syntaktických jevů. Zachycuje strukturu větných členů (tj. přísudek, podmět (subjekt), předmět (objekt), příslovečná určení atd.), včetně lexikální a morfologické specifikace. Protože lineární uspořádání uzlů stromu odpovídá

[9] Např. pro PDT je pro tyto účely definována tzv. analytická rovina.

[10] Fonologická rovina nahrazuje roviny morfonologickou (*morphophonemic level*) a fonetickou (pro mluvenou řeč), resp. grafematickou, se kterými se pracovalo ve starších verzích FGD.

[11] Ve dvojicích, kde nelze vynechat žádný ze členů, rozhoduje analogie: např. předmět i podmět jsou pokládány za závislé na slovese, neboť existují slovesa, která nemají předmět, příp. podmět. Zcela jiný je problém koordinace a apozice (přístavku) – k jejich zachycení je potřeba nové dimenze **???**.

povrchovému slovosledu, může být závislostní strom reprezentující povrchovou strukturu věty neprojektivní.

Rovina hloubkové struktury odpovídá významovému plánu jazyka a zachycuje významové jednotky, jejich lexikální a morfologické vlastnosti a vzájemné vztahy. Zápis věty na TR je zbaven homonymie (víceznačnosti). Oproti VČR závislostní stromy na TR splňují podmínku projektivity – pořadí uzlů (jejich lineární uspořádání) zachovává hloubkový slovosled. Ten odpovídá tzv. výpovědní dynamičnosti, a spolu s určením základu (tématu, východiska) a jádra (rematu, ohniska) udává aktuální členění věty.

Funkční generativní popis byl navržen a je dále rozvíjen jako popis z hlediska lingvistiky adekvátně zachycující češtinu, a zároveň vhodný jako teoretický podklad pro automatické a poloautomatické zpracování přirozeného jazyka.

4.7.2. TEORIE VALENCE

Teorie valence je důležitá jako teoretický koncept pro popis přirozeného jazyka, v našem případě češtiny. Valenční požadavky sloves (ale i substantiv a adjektiv) hrají ovšem klíčovou roli i při syntaktické analýze vět přirozeného jazyka, tedy pro aplikovaný úkol.[12] Co je míněno pojmem valence?

Podle autorů valenčního slovníku Slovesa pro praxi (Svozilová et al (1997)): "*Valencí rozumíme v lingvistice schopnost lexikální jednotky, především slovesa, vázat na sebe jiné výrazy a mj. tak zakládat větné struktury*".

The Concise Oxford Dictionary of Linguistics uvádí podobnou definici: "*The range of syntactic elements either required or specifically permitted by a verb or other lexical unit...*"

Teorie valence se ve světě široce studuje již od 60. let. Na evropskou syntax, která vychází z centrálního postavení slovesa ve větě, zejména na výsledky francouzského lingvisty L. Tesnièra (1959), navázal Charles Fillmore (1968, 1977) svou "pádovou gramatikou" studující sémantické role jednotlivých slovesných doplnění. Česká tradice navazovala rovněž na strukturně pojatý popis valence slovenského lingvisty E. Paulinyho (1943). Později se vztahy mezi syntaxí a sémantikou jako vztahy mezi gramatickými a sémantickými větnými vzorci zabýval Fr. Daneš a Z. Hlavsa (1981).

V rámci Funkčního generativního popisu (FGD) byla teorie valence rozpracována nejprve pro slovesa jako pro centrální jednotky věty (viz zejména Panevová (1974-1975), Hajičová (1979), Panevová (1980) a (1994)), posléze byla studována i valence dalších slovních druhů – substantiv (Novotný (1980), Panevová (2000)) a adjektiv (Piťha (1982), Panevová (1998)). Teorii valence je věnována také kapitola v Hajičové et al (2002), odkud přejímáme i některé příklady.

FGD pracuje s podkladovou strukturou věty a její povrchovou realizací. Pro ilustraci – jednotky jako subjekt (podmět) a objekt (předmět) slovesa, které popisují povrchovou stavbu věty, mají různé funkce v jazykovém systému: subjekt obvykle (u aktivního slovesa) vyjadřuje konatele, ale jindy (u pasiva) odpovídá patientu, tj. zasaženému objektu:

(1a) *Nakladatelství Torst*(subjekt) *vydalo mnoho dobrých knih*(objekt).

(1b) *Mnoho dobrých knih*(subjekt) *bylo vydáno nakladatelstvím Torst.*

Věta (1a) je aktivní, rozvitý podmět *nakladatelství Torst* vyjadřuje konatele, předmět *mnoho dobrých knih* je zasaženým objektem. Cítíme, že v pasivní větě (1b) zůstává význam jednotlivých větných členů stejný (*nakladatelství Torst* stále vyjadřuje konatele, *mnoho dobrých knih* zůstává zasaženým objektem), jejich povrchové vyjádření se však zaměnilo. Tyto **významové jednotky** (tj. konatel, zasažený objekt a další) zachycují hloubkovou stavbu věty čili její zápis na tektogramatické rovině.

Pojmy valence, valenčního rámce a valenčních doplnění se týkají především roviny hloubkové struktury. Pro potřeby automatické syntaktické analýzy (a tedy i pro NLP vůbec) je ovšem

[12] Máme zde na mysli především metody NLP založené na pravidlech. Přitom je potřeba zdůraznit, že i parsery navržené pro zpracování českých vět, které se deklarují jako čistě syntaktické, pracují se zjednodušenými valenčními rámci sloves.

nutné zabývat se také jejich realizací v povrchové struktuře věty, to znamená vhodně je interpretovat na rovině povrchové syntaxe.

4.7.2.1. Valenční rámce sloves

Sloveso je ve FGD, podobně jako v jiných popisech, chápáno jako centrum věty – proto je potřeba studovat v prvé řadě syntaktické vlastnosti sloves. Přitom je kladen důraz na formulaci ověřitelných kritérií pro určování jednotlivých doplnění, které tvoří valenční rámec, i pro určování jejich vlastností. Tento požadavek je zvlášť podstatný, chceme-li tento popis využít pro automatické zpracování.

Jako **slovesná doplnění** jsou označovány výrazy (jednoduché i komplexní), které mohou rozvíjet nějaké sloveso (v závislostní reprezentaci věty tvoří dceřinné uzly slovesa; seznam typů doplnění viz Hajičová et al (2000)).

Slovesný valenční rámec (v širokém smyslu) je tvořen všemi doplněními, které mohou dané sloveso v daném významu rozvíjet. Dále budeme termínu valenční rámec užívat v jeho užším smyslu – **slovesný valenční rámec** je tvořen **aktanty a obligatorními volnými doplněními** slovesa, tzv. valenčními doplněními slovesa či jeho valenčními členy (jednotlivé typy doplnění jsou popsány v následujících odstavcích). Přitom každé sloveso má alespoň jeden valenční rámec, často ale má rámců více.[13]

Charakteristika valenčních rámců spolu s možnými povrchovými formami aktantů (jejich možným morfematickým vyjádřením) musí být obsažena v slovníkovém hesle daného slovesa.

Slovesná doplnění se dělí podle následujících kritérií (i) a (ii) na aktanty a na volná doplnění.

Aktanty jsou taková doplnění slovesa (tzv. vnitřní doplnění, argumenty, participanty), která jsou charakterizována dvěma podmínkami:

(i) nemohou se vyskytovat více než jedenkrát (bez koordinace nebo apozice) jako rozvití konkrétního slovesa;

(ii) jejich kombinace je charakteristická pro jednotlivá slovesa (musí být specifikována ve valenčním slovníku).

FGD na základě empirických zjištění rozlišuje na tektogramatické rovině pět aktantů. Jsou to aktor (konatel, Agens, Actor, dále Act), patient (zasažený objekt, Pat), adresát (Addressee), původ (Orig) a výsledek (Eff).

Všech pět aktantů rozvíjí např. sloveso *předělat* ve větě (2). Naopak, existují i slovesa, která ve svém valenčním rámci nemají žádný aktant, např. sloveso *pršet* (3).

(2) *Matka*.Act *předělala dětem*.Addr *loutku*.Pat *z Kašpárka*.Orig *na čerta*.Eff.

(3) *Venku prší.*

Naproti tomu **volná doplnění** (adverbiální doplnění, adjunkty – odpovídají příslovečným určením) – např. místa, času, způsobu, prostředku, podmínky atd., seznam viz Hajičová et al (2000) – se mohou objevovat u téhož slovesa vícekrát a mohou rozvíjet jakékoliv sloveso (případná omezení jsou obsahové, nikoli gramatické povahy):

(4) *V Praze se sejdeme na Hlavním nádraží u pokladen.*

(5) *Včera přišel večer domů pozdě.*

(6) *Kvůli dešti musel čekat pod střechou, protože neměl deštník.*

---

[13] Má-li sloveso n valenčních rámců, má minimálně n **významů A CO ALTER???**. Pokud mají dvě slovesa -???jde o různá slovesa- **stejný valenční rámec** (stejný soubor valenčních členů, včetně obligatornosti) a navíc totožné morfematické vyjádření valenčních členů, říkáme, že mají **totožnou valenci**.

Pro některé úkoly stačí rozlišovat pouze tzv. syntaktické rámce – všechna slovesa s totožnými valencemi reprezentuje jediný **syntaktický rámec**.

Aktanty i volná doplnění mohou být u jednotlivých řídících sloves **obligatorní** (tj. pro dané sloveso povinně přítomny v hloubkové struktuře věty) nebo **fakultativní**.[14] Podobně i volná doplnění mohou být pro konkrétní sloveso obligatorní, příp. obligatorní vypustitelná.

Kritériem pro obligatornost aktantů a volných doplnění je tzv. **dialogový test** (viz Panevová (1974-1975)). Tento test slouží jako měřítko pro obligatornost doplnění, je-li zkoumaný člen vypuštěn na povrchové rovině:

(7a) A: *Přátelé už přišli.*
B: *Kam?*
A: **Nevím.*

Už otázka mluvčího B signalizuje neúplnost výpovědi mluvčího A, jeho odpověď pak činí dialog zcela deviantní (A musí vědět, o jakém místě mluví) – sloveso *přijít* má obligatorní doplnění směru "kam". Jiná situace nastává u doplnění směru "odkud":

(7b) A: *Přátelé už přišli.*
B: *Odkud?*
A: *Nevím.*

Zde je odpověď mluvčího A zcela v pořádku – sloveso *přijít* nemá obligatorní doplnění směru "odkud". Dialogový test pomáhá určit všechna doplnění konkrétního slovesa, která musí být přítomna ve větě, aby (vytržená z kontextu) byla úplná, a tím stanovit jeho valenční rámec.

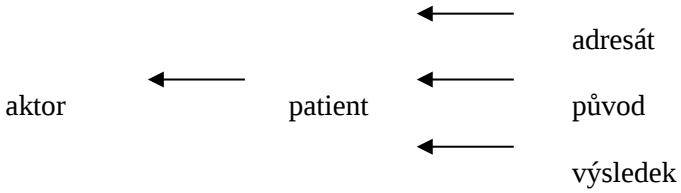
Fakultativní aktanty je třeba odlišovat od **všeobecného aktantu** (General Participant) – všeobecný aktant je jen zvláštním případem lexikálního obsazení obligatorního aktantu (Panevová (1992), Panevová a Řezníčková (2001)).

Příkladem všeobecného aktantu je např. patient slovesa *prodávat* ve větě (8):

(8) A: *Soňa prodává u Bati.*
B: *Co?*
A: *Nevím, boty. To, co se u Bati prodává.*

Mluvčí A sice neví, co konkrétně je patientem, je ale schopný ho specifikovat (v tomto případě jako určitou skupinu výrobků).

Pro určení aktoru a patientu FGD užívá **syntaktická kritéria** (v duchu přístupu L. Tesnièra (1959)), pro určení dalších aktantů užívá kriteria **sémantická** (tento přístup využívá zejména Fillmore (1968), (1977), u nás např. Daneš a Hlavsa (1981)). FGD vychází z toho, že se realizuje tzv. **posouvání aktantů** – prvním aktantem je vždy aktor, druhým patient. U sloves s více aktanty přistupují ohledy sémantické (viz následující schéma).



Má-li tedy sloveso jediný aktant, jde o aktor, slovesa se dvěma aktanty mají aktor a patient; teprve u sloves se třemi a více aktanty se uplatňují sémantická kritéria:

Příklady

Ve větách (9a) a (9b) se patient se posouvá do pozice aktoru (tj. aktoru odpovídá doplnění, které by primárně bylo vyjádřeno patientem):

(9a) *Kniha*.Act *vyšla.*
(9b) *Jan*.Act *spadl se skály.*

V dalších větách je aktor zaplněn, do pozice patientu se posouvá adresát (10a), (10b), efekt (11a), (11b) nebo původ (12):

[14] Je ovšem potřeba rozlišovat mezi obligatorností a nutností povrchového vyjádření konkrétního doplnění. Některá obligatorní doplnění nemusí být vždy realizovány v povrchové reprezentaci věty – říkáme jim (obligatorní) valenční členy **vypustitelné** (např. aktor je v české větě vypustitelný vždy). O povrchové vypustitelnosti viz dále.

(10a) *Jan*.Act *oslovil Marii*.Pat *(na ulici).*
(10b) *Učitel*.Act *vyvolal žáka*.Pat.
(11a) *Petr*.Act *vykopal jámu*.Pat.
(11b) *Jan*.Act *(za tu dobu) vyspěl (z jinocha*.Orig*) v muže*.Pat.
(12) *Z banálního nachlazení*.Pat *se vyvinulo závažné onemocnění*.Act.

Následují rámce několika častých českých sloves (nejsou zde uvedeny rámce odpovídající některým okrajovým významům[15]; v příkladech uvádíme i fakultativní volná doplnění):

Příklady doplněny příklady tázacích a rozkazovacích vět, vyhozena poznámka

chodit$_1$ Act *Petr*.Act *chodí rychle*.Mann.
Choď pomalu.Mann*!* (ty=Act)
chodit$_2$ Act Intt *Marie*.Act *chodí na borůvky/na nákup/nakupovat*.Intt.
chodit$_3$ Act Pat *Petr*.Act *chodí s Marií*.Pat. (idiom)

čekat$_1$ Act Pat (Orig) *Petr*.Act *čekal od Jirky*.Orig *omluvu/že přijde*.Pat.
čekat$_2$ Act Pat *Rodiče*.Act *čekají na dítě*.Pat *před školou*.Loc.
Kdo.Act *čeká na dítě*.Pat *před školou*.Loc*?*
Čekej na mě.Pat *před školou*.Loc. (ty=Act)[16]
čekat$_3$ Act Pat *Věřitel*.Act *jim*.Ben *čeká s dluhem*.Pat. (idiom)
čekat$_4$ Act Pat *Marie*.Act *čeká s Petrem*.Acmp *dítě*.Pat. (idiom)

hovořit$_1$ Act (Addr) Pat *Petr*.Act *o tom*.Pat *hovořil s přítelem*.Addr.
hovořit$_2$ Act *Petr*.Act *hovoří plynule/anglicky*.Mann.
hovořit$_3$ Act (Addr) (Pat) *Petr*.Act *o tom/na toto téma*.Pat *hovořil k publiku/ na přítele*.Addr (a ten ho nevnímal).

informovat$_1$ Act Addr Pat *Petr*.Act *informoval rodiče*.Addr *o svém návratu*.Pat.
informovat$_2$ Act Addr (Pat) Eff
Petr.Act *informoval rodiče*.Addr (*o noční bouřce*.Pat), *že byla zlá*.Eff.

informovat se$_1$ Act Pat *Petr*.Act *se na ně/o nich*.Pat *informoval*.
informovat se$_2$ Act (Pat) Eff *Petr*.Act *se* (*o nich*.Pat) *informoval, že mluví pravdu*.Eff

jednat$_1$ Act *Petr*.Act *začal jednat* (=konat).
jednat$_2$ Act Addr Pat *Petr*.Act *s nimi*.Addr *jedná* (=vyjednává) *o investicích*.Pat.
jednat$_3$ Act Pat Mann *Učitel*.Act *jedná* (=zachází) *se žáky*.Pat *špatně*.Mann.

jednat se Act Pat *Jedná se mi*.Act *o princip*.Pat.

nabízet Act Addr Pat (Eff) *Petr*.Act *jim*.Addr *nabízí služby*.Pat *za úplatu/jako úplatu/ úplatou*.Eff. (ale: *nabízí mu auto za pakatel*.Ext)

odpovídat$_1$ Act Addr (Pat) Eff *Petr*.Act *jim*.Addr *na dotaz*.Pat *odpovídal, že ...* Eff
(ale: *odpovídal mu na jeho dotaz činy*.Means)
odpovídat$_2$ Act Addr Pat *Petr*.Act *mu*.Addr *odpovídal pravdu/že ...*Eff
odpovídat$_3$ Act Pat *Řešení*.Act *odpovídá* (=je ve shodě) *požadavkům*.Pat.
odpovídat$_4$ Act (Addr) Pat *Rodiče*.Act *odpovídají* (=mají odpovědnost) *za své děti*.Pat. (ale: *odpovídá za ztrátu svým majetkem*.Means)

říkat$_1$ Act (Addr) (Pat) Eff *Petr*.Act *mu*.Addr *říkal o Marii*.Pat, *že je chytrá*.Eff.
říkat$_2$ Act Addr Pat *Petr*.Act *mu*.Addr *říkal o katastrofě*.Pat

[15] Indexy u sloves rozlišují jednotlivé rámce sloves, a tedy i jejich významy.

[16] Pro užití slovesa v **tázací větě** platí stejná pravidla jako pro věty oznamovací; ve **větách rozkazovacích** není vyjádřen aktor, ostatní doplnění zůstávají zachována.

říkat$_3$ Act (Addr) Pat *Petr*.Act *(si) mu*.Addr *marně říkal* (=požádal) *o vodu*.Pat

vyhrát$_1$ Act Pat (Orig) *Petr*.Act *na něm*.Orig *vyhrál v kartách*.Reg *pět korun*.Pat.
vyhrát$_2$ Act (Addr) (Pat) *Petr*.Act *s ním/proti němu/nad ním*.Addr *vyhrál zápas*.Pat. (ale: *Buffalo s Haškem*.Acmp *v bráně vyhráli s Penguins 2:0*)

zahájit Act Pat *Petr*.Act *zahájil schůzi*.Pat *projevem*.Means.

žít$_1$ Act *Petr*.Act *žil v Praze*.Loc.
žít$_2$ Act Pat *Petr*.Act *žal palouk*.Pat *kosou*.Means.

Valenční rámce se uvádějí pro činný (aktivní) tvar slovesa. V **pasivní větě** je valenční rámec zachován, dochází však k systémovým změnám v povrchové realizaci valenčních členů (viz dále).

4.7.2.2. Interpretace valenčních rámců na povrchové rovině

Řekli jsme, že pro potřeby automatické syntaktické analýzy je nutné zabývat se realizací jednotlivých valenčních doplnění v povrchové struktuře věty. **Slovesná doplnění** mohou být vyjádřena buď jednotlivými slovy – zejména podstatnými jmény a zájmeny v určitém pádu, ale i např. přídavnými jmény, příslovci, slovesy v infinitivu. Dále skupinami slov – rozvitými větnými členy, jmennými či předložkovými skupinami, a koordinovanými větnými členy. Další možností jsou slovesa v infinitivu a vedlejší věty (uvozené podřadícími spojkami, tázacími zájmeny a příslovci).[17]

Aktanty mají pro každé sloveso pevně danou **morfematickou formu** (zpravidla jedinou, někdy jich může být více) uchovávanou ve slovníku, volná doplnění (obligatorní i fakultativní) jsou určena sémanticky (typicky mají možnost různého morfematického vyjádření). Jednotlivé aktanty mohou být vyjádřeny u daného slovesa **prototypicky** (v aktivní větě aktor v nominativu, akuzativ pro patient, dativ pro adresáta), u jiného neprototypicky.

Vraťme se k **nutnosti povrchového vyjádření** konkrétního doplnění. Zdá se totiž, že neexistuje člen valenčního rámce, který by byl (absolutně) nevypustitelný (alespoň ve specifických kontextech, jako je např. odpověď na otázku). Na druhé straně zřejmě pro vypouštění platí nějaká omezení – pro konkrétní sloveso realizace některého aktantu na povrchové rovině vyžaduje vyjádření jiného aktantu (alespoň formou zájmena), především patientu.[18]

4.7.2.3. Obohacené rámce

"Klasické" valenční rámce nezachycují fakultativní volná doplnění. Pro NLP je však zřejmě výhodné využívat co nejbohatší slovníkové informace – je možné navíc pracovat s **kvazivalenčními doplněními**, což jsou "obvyklá doplnění" charakteristická pro konkrétní slovesa, která mohou specifikovat jeho význam, příp. jeho posunuté či přenesené užití (např. volné doplnění směru u slovesa *jet* (kam?) nebo *přijít* (odkud), doplnění prostředku (Means) u slovesa *hrát – hrát na kytaru* či doplnění účelu (Aim) u sloves *potřebovat* a *poskytnout*). Kromě kvazivalenčních doplnění lze zachycovat i **typická volná doplnění** (např. určení prospěchu (Ben) u slovesa *čekat*$_3$ *– čekat někomu (s dluhem)*, určení způsobu (Mann) u slovesa *hovořit*$_2$ *– hovoří plynule/anglicky* nebo určení zřetele (Reg) u slovesa *vyhrát*$_1$ *– vyhrál v kartách*) (více viz Straňáková (2001)).

---

[17] Navíc je pro některá slovesa možné u recipročních (vzájemných) vztahů některá doplnění zaplnit reflexivním zájmenem (viz Panevová (1999)):

Přátelé se sešli (=přítel.Act *s přítelem*.Pat*).*

Každý den na sebe čekají před školou (=jeden.Act *na druhého*.Pat*).*

[18] Např. valenční rámec slovesa *předělat* obsahuje obligatorní aktor, patient, původ a výsledek a fakultativní adresát. Varianty bez vyjádřeného patientu nejsou zřejmě správnými větami.

Matka.Act *předělala dětem*.Addr *loutku*.Pat *z kašpárka*.Orig *na čerta*.Eff.

**Matka*.Act *předělala dětem*.Addr *na čerta*.Eff.

**Matka*.Act *předělala dětem*.Addr.

Matka.Act *předělala loutku*.Pat.

4.7.2.4. Valenční rámce substantiv

Z hlediska valence je vhodné zabývat se zvlášť podstatnými jmény odvozenými od sloves a primárními substantivy typicky rozvíjenými doplněním v genitivu.

Substantiva odvozená od sloves (slovesné deriváty) vpodstatě dědí valenční rámec zdrojových sloves (viz Panevová (2000)). Je tedy vhodné uvažovat stejný repertoár aktantů a volných doplnění jako u sloves. Přitom **slovesná** substantiva (tzv. syntaktické deriváty jako *dělání*, *bodnutí*) stejně jako tzv. **široce dějová jména** (např. *odpověď*, *nominace*) sdílejí rámec s originálním slovesem, dochází u nich k systémovým změnám v povrchovém vyjádření (viz níž). "Vzdálenější" odvozeniny (tj. **lexikální deriváty** jako konatelská jména – *učitel*, *svářeč*, či názvy artefaktů – *dopis*, *povídka*) si ještě můžou zachovávat slovesná doplnění (tj. doplnění typicky rozvíjející slovesa), přistupují k nim ještě nominální doplnění.

Speciální nominální doplnění, doplnění **primárních substantiv** jsou partitiv (Part, někdy též material, Mat, v širokém smyslu kvantitativní určení), přináležitost (App, appurtenance), identita, restriktivní a deskriptivní přívlastek. Partitiv a identitu lze považovat za aktanty (partitiv u některých jmen obligatorní – *skupina*, *hektar*, u jiných fakultativní – *sklenice*, *koš*), ostatní jsou volná doplnění (přináležitost je však např. u relačních jmen obligatorní – *bratr*, *přítel*). Při derivaci často dochází k posunu významu (např. *náklady*) nebo ke konkretizaci (např. *půllitr*).

Valenční rámce substantiv, resp. pravidla, podle kterých se odvozují (včetně nesystémových změn u jednotlivých lexikálních položek), spolu s možnými povrchovými formami jednotlivých doplnění, je nutné uchovávat ve slovníku.

4.7.2.5. Povrchová rovina

Všechny členy valenčních rámců substantiv jsou v povrchové realizaci věty **vypustitelné** (přestože se u slovesných derivátů může zachovávat obligatornost na podkladové rovině).

U některých typů derivace zůstává rámec formálně nezměněn (některá dějová jména), při derivaci dochází k **regulárním změnám** v povrchové realizaci valenčních členů, příp. doplnění mohou nabývat několika forem. Přitom se nepřevádějí jednotlivé aktanty, ale celé rámce (Panevová (2000)),[19] jako v následujícím příkladu:

Příklad

Podstatné jméno *návrh* je odvozeno od slovesa *navrhovat*, přitom valenční rámec zůstává zachován:

navrhnout Act Pat (Addr) → *návrh* Act Pat (Addr)

(Např. *vláda navrhuje (parlamentu) rozpočet/zachovat rozpočet/že zachová rozpočet*.)

Mění se jen povrchové vyjádření jednotlivých aktantů (rámec lze převést čtyřmi způsoby):

někdo$_{\text{Nom}}$ *navrhuje (někomu*$_{\text{Dat}}$*) něco*$_{\text{Acc}}$ / inf / *že* …

→ *něčí*$_{\text{pos-subj}}$ *návrh něčeho*$_{\text{Gen-obj}}$ */ na něco*$_{Acc}$ / inf / *že* …

(Např. *ministrův*.Act *návrh rozpočtu/na zachování rozpočtu/zachovat rozpočet/že bude zachován rozpočet*.Pat.)

→ *něčí*$_{\text{pos-subj}}$ *návrh někomu*$_{\text{Dat}}$

(Např. *ministrův*.Act *návrh parlamentu*.Addr.)

→ *něčí*$_{\text{pos-subj}}$/*někoho*$_{\text{Gen-subj}}$ *návrh na něco*$_{Acc}$ / inf / *že* …

(Např. *ministrův*.Act *návrh na zachování rozpočtu/zachovat rozpočet*.Pat; *návrh vlády*.Act *na zachování rozpočtu/zachovat rozpočet*.Pat.)

→ *něčí*$_{\text{pos-subj}}$/*někoho*$_{\text{Gen-subj}}$ *návrh někomu*$_{\text{Dat}}$

(Např. *ministrův*.Act *návrh parlamentu*.Addr; *návrh vlády*.Act *parlamentu*.Addr.)

4.7.2.6. Kvazivalenční doplnění substantiv

Stejně jako pro slovesa, i pro substantiva lze využít bohatší informace – kvazivalenční doplnění substantiv zachycuje především **terminologická a frazeologická** spojení vlastní odbornému

[19] Obvykle se mění doplnění v nominativu (subjekt) na genitiv (subjektový) nebo přivlastňovací adjektivum (Adj$_{\text{pos-subj}}$), případně je vyjádřen instrumentálem nebo Pg *od*+Gen. Slovesná doplnění vyjádřená akuzativem často mění formu na genitiv objektový. Naproti tomu doplnění v genitivu, v dativu a v instrumentálu typicky zůstávají zachována, stejně jako předložkové skupiny.

názvosloví (*tečna v bodě x*, *souměrnost podle osy x*, Panevová (1966)). Při zpracování např. žurnalistických textů se zdá vhodné přistupovat stejným způsobem také k **‚obvykle užívaným spojením‘**. Spadají sem například ustálená spojení typická pro ekonomické či právní texty (které přitom nelze označit za odbornou terminologii (*řízení na (vedoucí) funkci*, Regard), je možné pracovat zde se slangem (ve sportovních komentářích) – podle typu textů, které mají být zpracovány. Dalším typem kvazivalenčních doplnění jsou některá **‚zděděná‘ doplnění** – obligatorní volné doplnění slovesa, případně jeho kvazivalenční doplnění se může transformovat v ‚obvyklé‘ doplnění substantiva odvozeného od tohoto slovesa (*pobyt*, kde?, *tlak na nápravu*, Pat).

Kvazivalenčních doplnění, včetně možného povrchového vyjádření, musí být vyznačeny ve slovníku. Předpokládáme přitom, že kvazivalenční doplnění substantiva může mít několik víceméně synonymních forem (*daň na něco* / *daň za něco* / *daň z něčeho*, Pat).

4.7.2.7. Valenční rámce adjektiv

Adjektiva odvozená od sloves také určitým způsobem zachovávají rámec původních sloves (viz zejména Panevová (1998b)). Proto se předpokládá stejný repertoár aktantů a volných doplnění, jako mají slovesa. Dochází ke dvěma regulárním změnám:
(i) Adjektivum přichází o jedno doplnění z ‚úplného‘ valenčního rámce slovesa – to je zaplněno, ‚obsazeno‘ slovem rozvíjeným zkoumaným adjektivem (obvykle jde o aktor, *zpívající dívka*, nebo patient, případně adresát, *oslovené obecenstvo*).
(ii) Na **povrchové rovině** není žádné z obligatorních doplnění povinné, všechna doplnění jsou vypustitelná.

Adjektiva odvozená od sloves mohou s původním slovesem sdílet i jeho kvazivalenční doplnění a doplnění fakultativní volná doplnění.

Příklad
Adjektivum *omezený* je odvozeno od slovesa *omezit*, dědí přitom valenční doplnění, ‚obsazen‘ je patient:

omezit Act Pat (Eff) → *omezený* Act Eff
(Např. *Petr*.Act *omezuje kouření*.Pat *na minimum*.Eff.)
někdo$_{Nom}$ *omezuje něco*$_{Acc}$ *na něco*$_{Acc}$ → *něco*$_{Nom}$ *omezené (někým*$_{Ins}$*) na něco*$_{Acc}$
(Např. *kouření omezené (Petrem*.Act*) na minimum*.Eff.)
Sloveso *omezit* je často rozvíjeno volným doplněním prostředku (Means, v instrumentálu) a zřetele (Regard, s formou *v*+Loc). Tato doplnění často rozvíjejí i adjektivum:
někdo$_{Nom}$ *omezuje někoho*$_{Acc}$ *(v něčem*$_{Loc}$*) (něčím*$_{Ins}$*) (na něco*$_{Acc}$*)*
→ *někdo*$_{Nom}$ *omezený (v něčem*$_{Loc}$*) někým*$_{Ins}$*/něčím*$_{Ins}$ *(na něco*$_{Acc}$*)*
(Např. *Pavel omezený v kouření*.Reg *Petrem*.Act */ žvýkačkami*.Means *na minimum*.Eff.)

Primární adjektiva

Také **adjektiva, která nejsou odvozena od sloves**, mají valenční rámce. Valencí primárních adjektiv se zabývá zejména Prouzová (1983), Piťha (1982) a Panevová (1998b). Lze shrnout:
(i) Adjektiva, která nejsou odvozena od sloves, mají stejný repertoár aktantů a volných doplnění jako slovesa.
(ii) Slovo, které je rozvíjeno adjektivem, váže/zaplňuje podle typu adjektiva (obvykle) aktor (participium aktivní) nebo patient (participium pasivní).
(iii) Navíc existují speciální doplnění, která rozvíjejí druhý a třetí stupeň adjektiv (pokud je lze vytvořit).

4.7.2.8. Kde lze valenční informace hledat

Pro češtinu sice již existuje několik valenčních slovníků sloves, nicméně buď je jejich rozsah omezený a forma neumožňuje automatické zpracování (**Svozilová at al** (1997)), nebo nejsou dostatečně spolehlivé díky automatickému zpracování (**Skoumalová** (2001)), případně vůbec

neobsahují podkladovou strukturu (**Pala a Ševeček** (1997)). Proto v současné době vzniká bohatě anotovaný valenční slovník popisující podkladovou strukturu, který je "čitelný" pro člověka a zároveň poskytuje data pro NLP. Tento slovník čerpá z existujících zdrojů, důraz je kladen zejména na konzistenci a úplnost zachycovaných údajů (**Straňáková-Lopatková a Žabokrtský** (2002)).

4.8. FORMÁLNÍ SPECIFIKACE PODKLADOVÉ STRUKTURY (TEKTOGRAMATICKÉ REPREZENTACE FGD)

Každý úplný formální popis podkladové reprezentace věty přirozeného jazyka (tak, jak je chápána v rámci Funkčního generativního popisu češtiny, tedy reprezentace na tektogramatické rovině, TR, viz Sgall – nebo 1. díl, krátké shrnutí zde v ...), musí zachycovat následující jevy:

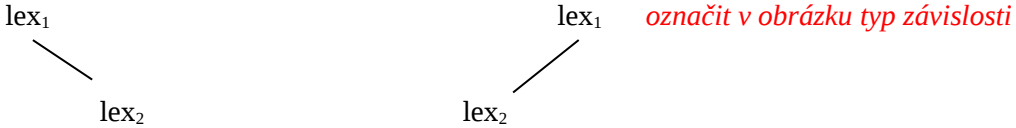
- ✓ závislostní vztahy založené na valenčních rámcích (vztahy podřízenosti)
- ✓ koordinace a apozice
- ✓ aktuální členění (topic-focus articulation)
- ✓ hloubkový slovosled
- ✓ gramatická koreference

Koncept *komplexní závislostní struktury* (CDS), která umožňuje formální zachycení všech těchto jazykových jevů, nabízí {petkevic95}. CDS lze přirozeně graficky znázornit jako strom se dvěma typy hran. Jde o datový typ, který umožňuje reprezentaci vět přirozeného jazyka na tektogramatické rovině FGD.
Zde přebíráme hlavní myšlenky článku {petkevic95}. Petkevič předpokládá podrobnou znalost FGD – pro čtenáře, kterým pojmy teoretického popisu přirozeného jazyka nejsou blízké, nabízíme opačnou cestu: přes formálně definovaný datový typ se přiblížit lingvistice. Po úvodní formální definici komplexní závislostní struktury ukážeme, jakým způsobem jsou v navrženém formalismu jednotlivé jevy zachyceny.[20]

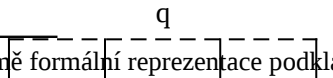
4.8.1. Formální definice (CDS)

Komplexní závislostní struktura (complex dependency structure, CDS) nad slovníkem LEX, množinou DEP symbolů a množinou Q symbolů, je řetězec symbolů nad abecedou
LEX ∪ {;,<,>,[} ∪ {>d; d∈DEP} ∪ {d<; d∈DEP} ∪ {]q; q∈Q}
pro který platí tvrzení (i)-(iv):
(i) pro všechna lex, lex_1, lex_2 ∈ LEX, d ∈ DEP
jsou lex, lex_1 d<lex_2> , <lex_2>d lex_1 CDS;
(ii) pro všechna lex_i ∈ LEX, 1<=i<=k, k>=2, q ∈ Q
jsou [lex_1; ... lex_k]q CDS;
(iii) jsou-li C_1 a C_2 CDS, potom je CDS i řetězec získaný substitucí jakéhokoliv lex ∈ LEX z C_1 za C_2;
(iv) neexistují jiné CDS než ty, které splňují podmínky (i)-(iii).

Interpretace zápisu lex_1 d<lex_2> , resp. <lex_2>d lex_1: lex_2 závisí (je závislý) na lex_1 zprava, resp. zleva, a to typem závislostní relace d; graficky znázorňujeme podstromy:



Interpretace zápisu [lex_1; ... lex_k]q: lex_1, ... lex_k jsou koordinované, resp. aponované; graficky znázorňujeme:



[20] Petkevič (1995) kromě formální reprezentace podkladové struktury a definice tektogramatického slovníku navrhuje i generátor se zásobníkem (pushdown store generator) , který generuje tektogramatickou reprezentaci věty. Zde se omezujeme na (statický) formální popis podkladové struktury.

lex_1 lex_2 lex_3 lex_4

Graficky CDS odpovídá tzv. závislostní strom (tj. spojitý acyklický orientovaný graf s kořenem, kde do každého uzlu kromě kořene vede právě jedna hrana, uzly jsou lineárně uspořádány) – jednotlivé prvky LEX jsou reprezentována uzly stromu, jejich pořadí dává lineární uspořádání uzlů a lomené a hranaté závorky představují dva typy hran.

strom s jednoduchými symboly
<[(A,t); (B,t)]q_1 d_1<<(C,t)>d_2 (D,t) d_3<(E,t)>>>d_4 (F,f) d_5<<(G,f)>d_6 (H,f)>

4.8.2. Lingvistická interpretace TADY

CDS má přirozenou lingvistickou interpretaci – slovník LEX obsahuje lexikální jednotky plnovýznamových slov, lomené závorky <> udávají typ jazykové závislosti (podřízenosti)[21] a hranaté závorky [] nový typ uzlu odpovídající koordinaci nebo apozici.
V grafické reprezentaci jde o závislostní strom, jehož uzly jsou ohodnoceny plnovýznamovými slovy a hrany typem závislosti (podřízenosti), resp. typem koordinace / apozice.

strom s větou

(2) Jan a Marie, kteří žijí v Bostonu, jsou moji přátelé. strom
< [(Jan,t); (Marie,t)]q_1 d_1< <(který,t)>d_2 (žít,t) d_3<(Boston,t)> > >d_4 (být,f) d_5< <(můj,f)>d_6 (přítel,f)>

Slovník

FGD předpokládá, že pouze plnovýznamová slova (jako jsou slovesa, podstatná a přídavná jména) jsou na TR reprezentována uzly závislostního stromu, slova pomocná (jako pomocná slovesa či předložky) jsou zachycena jako atributy (gramatémy) slov významových.
Proto slovník **LEX** obsahuje právě plnovýznamová slova, **komplexní terminální sémantémy** (cts) a jejich vlastnosti.

cts = (wb, cl, gr, tf, relpath)

Každá jednotka (slovo) w∈LEX je dál vnitřně strukturovaná - význam slova (lex) a jeho syntakticko-sémantické vlastnosti (properties) tvoří tzv. bázi slova (wb, 'word basis'),

wb = (lex, properties).

Nejpodstatnější vlastností slova je slovní druh (cl, 'word class').
Další vlastnosti se vztahují ke konkrétnímu užití slova ve větě - gr je množina gramatémů (jako je např. osoba, číslo, čas, vid u slovesa, rod, číslo, pád u podstatného jména, stupeň u přídavného jména), tf zachycuje aktuální členění a relpath gramatickou koreferenci (viz níž). Při grafickém vyjádření jsou slova w∈LEX reprezentována uzly stromu.

Zachycení závislostí (podřízenosti)

Antisymetrická relace mezi řídícím slovem a jeho doplněním, tzv. **relace závislosti** (podřízenosti), je zachycena lomenými závorkami <>d, resp. d<>, d∈DEP, v nichž je uzavřen rozvíjející člen.
DEP je množina přirozených čísel, které kódují typ doplnění (viz {valence}), tedy jednotlivé aktanty a volná doplnění. Přitom pro každý slovní druh je dána množina $DEP_{cl} \subseteq DEP$ všech možných doplnění.

cl → $DEP_{cl} \subseteq DEP$

[21] Termínem závislost se běžně označují různé věci – jednak formální vztah dvou jednotek, jednak lingvisdtický jev,kdy jedna jazyková jednotka (plnovýznamové slovo) rozvíjí jinou jednotku. Abychom tyto jevy terminologicky rozlišili, užíváme termín závislost pouze pro formální vztah. Lingvistický jev nazýváme podřízenost.

Navíc je pro každou lexikální jednotku w∈LEX dána podmnožina $DEP_w \subseteq DEP_{cl}$ doplnění, které tvoří její **valenční rámec** (ať již pracujeme s 'klasickými' či obohacenými valenčními rámci, viz {valence}).

$$w \rightarrow DEP_w \subseteq DEP_{cl}$$

Množina DEP_w se dále dělí podle podle typu doplnění (na aktanty a volná doplnění) a podle obligatornosti (obligatorní a fakultativní, případně též kvazivalenční a typická doplnění).
Graficky jsou relace závislosti (podřízenosti) znázorněny jako hrany vedoucí vždy od řídícího slova (‘rodičovský’uzel) ke slovu závislému (podřízenému) (‚dceřinný‘ uzel).

Koordinace a apozice

Koordinaci v jazyce lze ilustrovat příklady *Jan, Petr a Marie* či *Jan, Petr nebo Marie*. Za apozici jsou považována spojení typu *Hamlet, princ dánský* nebo *kralevic Ivan*.
Relace koordinace a apozice je relace symetrická. Pro její adekvátní??? popsání je vhodné využít bohatší datový typ než je stromová struktura s jedním typem hran – obohacujeme stromovou strukturu o nový typ komplexního uzlu.
Relace koordinace a apozice (c/a konstrukce) je zachycena hranatými závorkami []q, q∈Q, v nichž jsou uzavřeny členy této konstrukce. Množina **Q** je udává typ koordinace, resp. apozice.

Zavedení komplexních uzlů reprezentujících c/a konstrukce s sebou nese nutnost specifikace jejich vlastností a vztahů k jednotlivým členům:

- **zobecnění CDS**

Podobně jako u jednotlivých slov definujeme i bázi c/a konstrukce, která se skládá z podmnožin jednotlivých kategorií (každá podmnožina je sjednocením hodnot příslušných kategorií jednotlivých členů c/a konstrukce). Obdobně pro množinu gramatémů, která se pro c/a konstrukci skládá z podmnožin hodnot jednotlivých členů.

- **závislostní relace** (podřízenosti) **c/a konstrukce a jejích jednotlivých členů**

Je-li celá c/a konstrukce rozvíjena doplněním d∈DEP, potom toto doplnění rozvíjí i všechny její členy. Z toho vyplývá:
- Je-li taková konstrukce rozvíjena nějakým aktantem[22], už žádný její člen nemůže být tímto aktantem rozvíjen.
- Je-li alespoň jeden člen c/a konstrukce rozvíjen nějakým aktantem, celá konstrukce už tímto aktantem nemůže být rozvíjena

Z definice CDS plyne, že celá c/a konstrukce i její jednotlivé členy mohou být rozvíjeny a mohou rozvíjet stejně jako lexikální jednotky. Navíc c/a konstrukce mohou být zanořeny - celá c/a konstrukce i její člen mohou být rozvíjeny jinou c/a konstrukcí.
Tyto vlastnosti mají i koordinační a apoziční vztahy ve větách přirozeného jazyka, jak ukazují následující příklady:

(3) *Marie a její přítel Jan, který žije v Bostonu, jsou moji přátelé.*

dopsat linearizované stromy ???

(4) *Jan, který žije v Bostonu a učí angličtinu, je mým přítelem.*
(5) *Jan a Marie, kteří žijí v Bostonu a učí angličtinu, jsou moji přátelé.*
(6) *Marie a její přítel Jan, který žije v Bostonu a učí angličtinu, jsou moji přátelé.*

Ve (3) je jeden člen koordinace, *Jan* je rozvíjen přívlastky shodnými a přívlastkovou větou.
Ve (4) je podstatné jméno *Jan* rozvíjeno koordinovanými přívlastkovými větami.
V (5) jsou koordinovaná podstatná jména *Jan a Marie* rozvíjena koordinovanými přívlastkovými větami.
V (6) je jeden člen koordinace *Jan* rozvíjen přívlastky shodnými a koordinovanými přívlastkovými větami.

[22] Aktanty se mohou vyskytnout pouze jednou jako doplnění konkrétního slovesa, viz {valence}.

4.8.3. Hloubkový slovosled a aktuální členění (topic-focus articulation)

Každé plnovýznamové slovo (cts) v konkrétní podkladové reprezentaci je buď kontextově zapojené, nebo kontextově nezapojené (více viz FGD). Kontextová zapojenost / nezapojenost slova v konkrétní větě je vyznačena hodnotou tf ∈ {t, f}.
Kontextově zapojená slova (CB, contextually bound), která nevyjadřují novou informaci, patří k tématu věty (t, topic). **Kontextově nezapojená** slova (NB, contextually non-bound) tvoří réma (f, focus), tj. nesou novou informaci.
Hloubkový slovosled věty, který odpovídá výpovědní dynamičnosti, je dán lineárním uspořádáním jednotlivých plnovýznamových slov v CDS.

4.8.4.Gramatická koreference

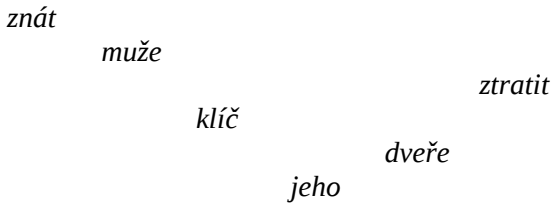
Gramatická koreference zachycuje vztah mezi dvěma výrazy, tzv. antecedentem (plnovýznamovým slovem) a výrazem, který k němu odkazuje. Lze ji ilustrovat např. vztahem mezi podstatným jménem a vztažným zájmenem.

Protože gramatické kategorie (osoba, číslo) vztažného zájmena musí souhlasit s gramatickými kategoriemi vztažného zájmena, je nutné koreferenci zaznamenat i ve formálním modelu.

(8) *Jan, který žije v Bostonu, je můj přítel.*
(9) **Marie, který žije v Bostonu, je moje přítelkyně.*

Formálně lze tento vztah zachytit pomocí cesty (`relative path‘) v CDS vedoucí od antecedentu přes hlavní sloveso vztažné vět ke vztažnému zájmenu – každé slovo v CDS má atribut relpath, která nabývá hodnot 0 (slovo neleží na žádné cestě), RELSTART (začátek cesty, tj. sloveso ve vztažné větě), nebo RELCONT (toto slovo leží na cestě, resp. jde o vztažné zájmeno).
Ve větě (8) vztažné zájmeno *který* závisí bezprostředně (je podřízeno) na slovese *žít*, které je hlavním slovesem vztažné věty. Vztažné zájmeno však může být zanořeno hlouběji, viz nap. větu (10):

(10) *Znám muže, klíč od jehož dveří jsem ztratil.*



ZÁVĚR ???

5. **Matematizace sémantiky**
Po té co se Chomskému podařilo iniciovat plodnou interakci mezi lingvistikou a matematikou, se přirozeně vynořila otázka, zda by bylo možné aplikovat matematické metody, které se ukázaly tak užitečné pro analýzu syntaxe, také na sémantiku. Sémantika je ale zřejmě oblastí neskonale problematičtější než syntax: zatímco předmětem syntaxe jsou samy výrazy a jejich skladba, předmětem sémantiky jsou významy, u nichž je zásadně problematické se shodnout byť i jen na tom, čím vlastně jsou, natož pak na tom, jak je uchopit.

První pochomskyovské pokusy o matematické uchopení sémantiky přirozeného jazyka velice přímočaře přejímaly metody, které se osvědčily v rámci generativní či transformační gramatiky. Asi

nejvýznačnějším z pokusů tohoto druhu byla tzv. *generativní sémantika* Lakoffa a McCawleyho (viz Lakoff, 1971). Jinou teorií sémantiky tohoto druhu byla například teorie J.J. Katze a P. Postala (1964). Také v rámci teorií samotného Chomského a jeho následovníků začaly postupně hrát čím dál větší roli roviny, které se zdály být spíše záležitostí sémantiky než syntaxe, zejména tzv. *logická forma* (která ovšem nijak přímo nesouvisela s logikou) – viz např. Chomsky (1986) nebo Jackendoff (1990). V Čechách se v rámci Sgallovy varianty generativní gramatiky, *funkčního generativního popisu*, rozvinula koncepce *tektogramatiky* (viz Sgall et al., 1987).

Aspirace teorií tohoto druhu na to, aby byly prohlášeny za skutečné teorie sémantiky, však narazily na odpor některých filosofů a logiků, kteří měli za to, že sémantika nějak zásadně souvisí s pravdivostí a pravdivostními podmínkami – a že teorie, které tuto souvislost neexplikují, nejsou skutečnými teoriemi sémantiky. Nejilustrativnější je v tomto ohledu polemika Davida Lewise (1972) s koncepcí J.J. Katze. Lewis argumentuje, že ‚lingvistické' teorie sémantiky nejsou skutečnou explikací významů, ale jenom překlady z jednoho jazyka (toho přirozeného) do jazyka jiného, jakési ‚stromovštiny'.

V rámci logiky se přitom již několik desetiletí rozvíjely teorie sémantiky formálních jazyků, které z propojení významu s pravdou vycházely. (Sémantika formálních jazyků logiky se totiž od počátku odvíjela od Tarského formální teorie pravdivosti[23].) Způsob, jakým by mohla logika význam matematicky zachytit, nastínil ovšem již na sklonku devatenáctého století jeden ze zakladatelů moderní logiky, německý matematik a filosof Gottlob Frege; Alfred Tarski, Rudolf Carnap a další je potom v podstatě jenom zaintegrovali do rozvíjející se formální logiky. Z hlediska teorií přirozeného jazyka byl ovšem problém ve dvou věcech: (i) jazyky, se kterými logici běžně pracovali a pro které budovali své formální teorie sémantiky, měly ve srovnání s přirozeným jazykem příliš jednoduchou syntaktickou strukturu; a (ii) sémantika, kterou byly opatřovány, nebylo možné považovat za realistickou explikaci významů, jaké mají výrazy v přirozeném jazyce.

Až kolem roku 1970 se situace začala měnit: objevily se formální jazyky, jejichž sémantika již nebyla z hlediska explikace sémantiky jazyka přirozeného tak nepoužitelná jako sémantika 'standardní' logiky. To souviselo s rozvojem *modálních logik* (to jest logik možnosti a nutnosti) a později logik intenzionálních, jejichž sémantika se opírala o pojem *možného světa*. V rámci intenzionálních logik bylo navíc toto sémantické prohloubení zkombinováno s obohacením syntaxe, protože jazyky těchto logik byly založeny na syntakticky bohatém zobecnění ‚standardní' logiky. Největší věhlas v tomto ohledu získal americký logik Richard Montague[24], ale pozoruhodné výsledky paralelně s ním dosáhl i český logik Pavel Tichý[25].

Rozvoj intenzionální sémantiky i dalších formálně-sémantických systémů, které reagovaly na diskuse o jejích omezeních, dal postupně vzniknout interdisciplinárnímu směru, ve kterém se spojili formálně orientovaní lingvisté, filosofové, logikové, i někteří odborníci z oblasti *computer science*. Pro tento směr se někdy používá termín *formální sémantika* a někdy poněkud specifičtější *modelově-teoretická sémantika* (*model-theoretic semantics*; z „teorie modelů", jak se v rámci matematické logiky a algebry nazývá studium sémantických vlastností formálně logických jazyků[26]).

Ve zbytku této kapitoly nejprve objasním základní myšlenku ‚matematizace významu', ze které tento přístup vyšel. Potom v dalších dvou oddílech naznačím, jak lze tuto myšlenku zobecnit na jazyky

[23] Viz o tom Peregrin (1999, §5.2).

[24] Viz Montague (1974); a také Partee a Hendriks (1997).

[25] Původní Tichého myšlenka je načrtnuta v Tichý (); její pozdější, již 'hyperintenzionální' (viz níže) podoba je pak podrobně vyložena v Tichý (1998). Viz také Materna a Štěpán (2000).

[26] Výborný úvod do problematiky sémantiky jazyků logiky představuje Cmorej (2001).

syntakticky tak bohaté, jako jsou ty, které stojí z základu intenzionálních logik (to jest na jazyky v rámci tzv. lambda-kategoriální gramatiky). Potom ukážu, jak se zapojením možných světů dostaneme od extenzionálního k intenzionálnímu modelu sémantiky přirozeného jazyka. V následujících dvou oddílech pak naznačím, jakými směry se vyvíjí post-intenzionální teorie sémantiky – načrtnu základní myšlenky za hyperintenzionálními a dynamickými modely sémantiky. Na závěr naznačím, jakým se může formální sémantika potýkat s lingvisticky netriviálními problémy přirozeného jazyka[27].

5.1.'Fregův manévr': funkcionální koncepce sémantiky

Gottlob Frege přišel s nápadem, že významy některých druhů výrazů, s jejichž explikací si jeho předchůdci příliš nevěděli rady, by bylo možné explikovat jako určité funkce v matematickém slova smyslu[28]. Načrtněme, jak k tomu dospěl.

Jednoduchá věta přirozeného jazyka se v nejjednodušším případě skládá z podmětu (v typickém případě singulární fráze) a předmětu (verbální fráze). Podmět, podle Fregova názoru, obvykle označuje nějakou věc: tak podmět „Jaromír Jágr" označuje Jaromíra Jágra, „Václav Havel" označuje Václava Havla, „prezident České republiky" označuje (v době, kdy je tato kniha psána) také Václava Havla atd. Komplikovanější úvahy pak vedly Frega k závěru, že (oznamovací) věta označuje svou pravdivostní hodnotu: tak věta „Jaromír Jágr je dramatik" označuje *nepravdu* zatímco „Prezident České republiky je dramatik" označuje *pravdu*. Co, položil si potom Frege otázku, za těchto předpokladů označují predikáty, takové jako „být dramatik" či třeba „kulhat"?

Jeho odpověď se odvinula od úvah o tom, jak predikáty v rámci věty fungují. Vezmeme-li predikát P, víme, že spolu s podmětem N_1 vytvoří nějakou větu V_1, s podmětem N_2 vytvoří V_2 atd.:

$P + N_1 = V_1$
$P + N_2 = V_2$
...

Pro konkrétní případ predikátu "být dramatikem" tedy dostáváme

"být dramatikem" + "Jágr" = "Jágr je dramatikem"
"být dramatikem" + "prezident ČR" = "Prezident ČR je dramatikem"
...

Můžeme tedy P vidět jako prostředek přiřazení věty V_1 podmětu N_1, věty V_2 podmětu N_2 atd.:

P: $N_1 \rightarrow V_1$"být dramatikem": "Jágr" → "Jágr je dramatikem"

[27] Podrobněji jsem vše to, co naznačuji zde, vyložil na jiném místě (Peregrin, 1998). Encyklopedií vyčerpávajícím způsobem mapující stav formální sémantiky ve druhé polovině devadesátých let sestavili J. van Benthem a A. ter Meulenová (1997).

[28] Sémantika byla před Fregem často považována za záležitost psychologie. Frege se naproti tomu snažil významy uchopit ne jako něco subjektivně-psychologického, ale jako něco, co existuje objektivně, bez přímé souvislosti s tím, co se děje v hlavách mluvčích jazyka. Měl totiž za to, že jsou-li věty našeho jazyka pravdivé či nepravdivé nezávisle na tom, co si o tom kdo myslí, pak i významy musí být nezávislé na tom, co se komu z mluvčích 'děje v hlavě'. To je dáno i tím, že má-li být význam prostředkem komunikace, musí existovat nějakým způsobem, který umožňuje, aby k němu měli přístup všichni mluvčí jazyka – tedy objektivně.

$N_2 \rightarrow V_2$ “prezident ČR” → “Prezident ČR je dramatikem”

... ...

Předpokládejme nyní, že víme, co je označováno podměty i větami. Použijeme-li pro to, co je označováno výrazem *Y* (říkejme tomu *denotát*), značku $\|Y\|$ (přičemž vynecháváme případné uvozovky), můžeme celou tuto úvahu přenést na úroveň sémantiky. Denotát predikátu *P* tedy můžeme nahlédnout jako prostředek přiřazení denotátu věty V_1 denotátu podmětu N_1, denotátu věty V_2 denotátu podmětu N_2 atd.:

$\|P\|$: $\|N_1\| \rightarrow \|V_1\|$
$\|N_2\| \rightarrow \|V_2\|$
...

$\|$být dramatikem$\|$: $\|$Jágr$\|$ → $\|$Jágr je dramatikem $\|$
$\|$prezident ČR$\|$ → $\|$Prezident ČR je dramatikem $\|$
...

Jsou-li denotáty podmětů jimi pojmenovávané věci a denotáty výroků jejich pravdivostní hodnoty, stává se z $\|$být dramatikem$\|$ funkce, která přiřazuje osobě Gottlobu Fregovi pravdivostní hodnotu *nepravda* (N), osobě Václavu Havlovi pravdivostní hodnotu *pravda* (P) atd. (obecně každému dramatikovi P a každému jinému objektu N). A Frege skutečně navrhl *ztotožnit denotáty predikátů s funkcemi, přiřazujícími věcem pravdivostní hodnoty*. (Vzhledem k tomu že takováto funkce je vždy charakteristickou funkcí[29] nějaké množiny – v našem případě množiny dramatiků – můžeme ji někdy přímo s jí charakterizovanou množinou ztotožňovat a vidět tedy denotát predikátu alternativně jako množinu.)

Fregovská sémantika byla postupně, v první polovině dvacátého století, integrována do rodící se logické teorie formálních jazyků. Zásluhu na tom měl zejména Alfred Tarski, který jako první ukázal, že tímto způsobem lze definovat přijatelnou sémantiku formálních jazyků stejně exaktně jako jejich syntax. Zkoumání takto uchopených sémantických aspektů jazyka elementární logiky (predikátového počtu prvního řádu) pak dalo vznik tzv. *teorii modelů*[30].

‘Fregův manévr’, kterým jsme se dobrali denotátů predikátů, lze ovšem zobecnit. Vezměme třeba spojky, takové jakými jsou “a” či “nebo” (případně jejich formálně-logické koreláty “^” a “˅”). Taková spojka zřejmě ‘vyrábí’ větu z dvojice vět; tedy na úrovni denotátů pravdivostní hodnotu z dvojice pravdivostních hodnot. Její denotát tedy můžeme ztotožnit s funkcí, přiřazující pravdivostní hodnoty dvojicím pravdivostních hodnot. (Jaké konkrétně funkce to budou v případě uvedených spojek je nasnadě: $\|$a$\|$ bude přiřazovat P pouze dvojici <P,P>, zatímco $\|$nebo$\|$ bude přiřazovat P každé dvojici kromě <N,N>.)

Se zobecňováním však můžeme pokračovat i za hranice toho, co je běžné ve standardní logice. Představme si například, že bychom do jazyka chtěli přidat ‘příslovce’, to jest výrazy, které se spojují s (unárními) predikáty v komplexní (unární) predikáty. Fregovskou úvahou pak můžeme denotáty těchto nových typů výrazů stanovit jako funkce, přiřazující denotátům predikátů denotáty predikátů – to jest

[29] Charakteristickou funkcí množiny *M* je funkce, která každému prvku množiny *M* přiřazuje P a každému jinému prvku ze svého definičního oboru přiřazuje N.

[30] Viz Ježek (1976) a Sochor (2001; Kapitola V).

jako funkce, které budou funkcím z individuí do pravdivostních hodnot přiřazovat funkce z individuí do pravdivostních hodnot.

Takováto sémantika je však z hlediska přirozeného jazyka neuspokojivá proto, že fregovské denotáty jistě nejsou přijatelnými explikáty významů v intuitivním slova smyslu: asi nikdo by nechtěl souhlasit s tím, že významem věty je její pravdivostní hodnota, a že tedy například všechny pravdivé věty mají stejný význam!

Frege sám si toho byl dobře vědom, a vedle toho, čemu on říkal "význam" (a co jsme my právě rekonstruovali jako denotát), přisoudil výrazu navíc něco, čemu říkal "smysl" a co *de facto* odpovídalo významu v intuitivním smyslu slova. Rudolf Carnap (1947) pak navrhl nahradit Fregovu dvojici termínů "význam" a "smysl" techničtějšími termíny "extenze" a "intenze" a poukázal na to, že chceme-li skutečně analyzovat význam, musíme se pokusit explikovat intenze nějakým stejně exaktním způsobem, jakým se Fregovi, Tarskému a spol. podařilo explikovat extenze.

Cesta k takové explikaci se ukázala vést přes pojem *možného světa*, zavedeného do sémantiky částečně Carnapem a částečně logiky budujícími sémantiku pro tzv. *modální logiky* (Kripke, 1963). Úvaha byla následující: znát intenzi výrazu znamená být schopen určit jeho extenzi v každém možném světě; a intenzi výrazu tedy můžeme obecně ztotožnit s funkcí, přiřazující každému možnému světu extenzi tohoto výrazu v tomto možném světě. Tak intenzí výroku je funkce, která každému možnému světu přiřadí pravdivostní hodnotu tohoto výroku v tomto možném světě; intenzí singulární fráze je funkce, která každému možnému světu přiřadí objekt označovaný touto frází v tomto možném světě; intenzí predikativní fráze je funkce, která každému možnému světu přiřadí množinu objektů, o kterých je tento predikát pravdivý v tomto možném světě (respektive její charakteristickou funkci) atd.

5.2.Rámec pro formální rekonstrukci sémantiky přirozeného jazyka

Jde-li nám pouze o syntax, můžeme si jazyk představit jako nějaký konečný soubor slov (různých kategorií) plus nějaký soubor pravidel pro skládání složitějších výrazů z výrazů jednodušších. (Terminologií algebry tedy můžeme říci, že jazyk se z tohoto pohledu jeví jako určitá *(mnohasortová) algebra*, jejímiž generátory jsou slova.) Chceme-li do tohoto obrázku dostat i sémantiku, musíme zřejmě každému výrazu přiřadit nějaký denotát.

V předchozím oddíle jsme se při rekonstrukci Fregovy explikace významu predikátů implicitně opřeli o předpoklad, že denotát složeného výrazu je určen denotáty jeho částí. Většinou sémantiků (i když ne všemi bez výjimky) je přijímáno, že tohle je skutečně obecný princip charakteristický pro význam; říká se mu *princip kompozicionality*:

> Význam složeného výrazu je jednoznačně určen významy jeho částí a způsobem jejich kombinace.

Pro sémantický model jazyka to znamená, že máme-li v něm syntaktické pravidlo P, které nám dovoluje zkombinovat výrazy Y_1, ... Y_n ve výraz $P(Y_1,...,Y_n)$, musíme mít paralelní sémantické pravidlo, které nám dovolí zkombinovat denotáty $\|Y_1\|$, ... $\|Y_n\|$ v denotát $\|P(Y_1,...,Y_n)\|$. (Z algebraického hlediska nám princip kompozicionality tedy *de facto* říká, že denotáty musí tvořit algebru, která je podobná algebře výrazů, a že přiřazení denotátů výrazům musí být homomorfismem té druhé do té první).

Sémantický model jazyka je tedy obecně tvořen čtyřmi komponentami:

(i) slovník
(ii) syntaktická pravidla
(iii) přiřazení denotátů slovům
(iv) pravidla pro to, jak „počítat“ denotáty složených výrazů z denotátů jejich složek.

S jistým zjednodušením nyní můžeme říci, že Fregovský model sémantiky je charakterizován tím, že všechna pravidla bodu (iv) mají tvar ‘vezmi denotát jedné složky a aplikuj ho na denotáty těch ostatních‘. Z toho vyplývá, že v tomto modelu mohou existovat jedině pravidla kombinující výrazy, z nichž vždy jeden denotuje funkci aplikovatelnou na denotáty těch ostatních.

Extenzionální sémantické modely jsou navíc charakterizovány tím, že se opírají o dvě základní kategorie výrazů, jejichž prvky nedenotují funkce: o kategorii *termů* (‚jmen‘, ‚podmětů’), jejíž výrazy denotují objekty nějakého ‚univerza diskurzu‘, a kategorii *výroků* (‚oznamovacích vět‘), jejíž prvky denotují pravdivostní hodnoty.

5.3.Funkcionální chápání významu: kategoriální gramatika

Fregův manévr je tedy, jak jsme viděli, založen na předpokladu, že způsob, jak se denotát přísudku kombinuje s denotátem podmětu v denotát jimi tvořené věty, můžeme explikovat jako *funkční aplikaci* toho prvního na ten druhý; a my jsme naznačili, že takový pohled se dá zobecnit na další gramatická pravidla. Zcela důsledným zobecněním této myšlenky je pak tzv. *kategoriální gramatika* (kde slovem gramatika rozumíme jazykový rámec, čili vymezení nějaké třídy jazyků).

Představme si, že máme jazyk, jehož všechna pravidla fungují tímto způsobem. To znamená, že máme-li gramatické pravidlo, které kombinuje výrazy kategorií $K_1 ... K_n$ ve výraz kategorie K, pak denotáty výrazů jedné z kategorií $K_1 ... K_n$, řekněme K_i, musí být funkcemi aplikovatelnými na denotáty výrazů zbylých kategorií $K_1,...,K_{i-1},K_{i+1},...,K_n$. Abychom toto zviditelnili, budeme v takovém případě kategorii K_i označovat indexem $K/K_1,...,K_{i-1},K_{i+1},...,K_n$. Budeme-li tedy kategorie výroků a termů označovat **V** a **T**, pak bude kategorie predikátů tímto způsobem označena jako **V**/**T**. Toto označení přímo ukazuje, že výraz kategorie predikátů doplněný výrazem kategorie **T** dá výraz kategorie **V** (příslušné gramatické pravidlo pak můžeme vyjádřit způsobem opticky připomínajícím krácení zlomků: výraz kategorie **V**/**T** se kombinuje v výrazem kategorie **T** ve výraz kategorie **V**). Podobně můžeme spojky „a“ či „nebo“ nahlédnout výrazy kategorie **V**/**V**,**V**; a ‚příslovce‘ jako výrazy kategorie (**V**/**T**)/(**V**/**T**).

Pro jednoduchost se můžeme omezit ‘unární’ variantu kategoriální gramatiky, to jest variantu, ve které pracujeme jenom s kategoriemi, které mají za lomítkem jedinou kategorii (která ovšem může být složená); obejdeme se tedy bez kategorií typu $K/K_1,...,K_n$ pro $n>1$. Máme-li totiž kategorii $K/K_1,...,K_n$, pak to znamená, že máme pravidlo, které výrazy této kategorie kombinuje s výrazy kategorií $K_1,...,K_n$; a my si toto pravidlo můžeme rozložit na *n* kroků (dílčích pravidel), z nichž v každém se přidává jeden z výrazů kategorií $K_1,...,K_n$. Výrazy kategorie $K/K_1,...,K_n$ tak přejdou ve výrazy kategorie $(K/K_1)/.../K_n$ (či kategorie $(K/K_n)/.../K_1$). Logické spojky, které, jak jsme viděli, vycházejí jako kategorie **V**/**V**,**V**, se tak stanou výrazy kategorie (**V**/**V**)/**V**. (Trik spočívá v tomto případě v tom, že namísto abychom logické spojky viděli jako kombinující se s dvojicí výroků ve výrok, je budeme vidět jako kombinující se s výrokem v něco, co dá výrok, když je to zkombinováno s dalším výrokem.)

Kategoriální gramatika je tedy obecně vymezena následujícím způsobem:

1. *Slovník.* Předpokládáme soubor KAT nějakých primitivních kategorií (KAT se může například skládat z kategorií **V** a **T**). Gramatickou kategorií je pak cokoli, co dostaneme z prvků KAT pomocí lomítka. Přesněji: prvky KAT jsou gramatickými kategoriemi; a kdykoli jsou K_1 a K_2 gramatickými kategoriemi, je gramatickou kategorií i K_1/K_2. (Máme tedy kategorie **V**/**T**, (**V**/**V**)/**V** atd.). Každá kategorie pak obsahuje nejvýše konečný počet slov. (Tak kategorie **T** může obsahovat slova *Jágr*, *Havel*, *prezident ČR*...; kategorie **V**/**T** třeba slova *dramatik*, *hokejista* ...; kategorie (**V**/**V**)/**V** slova *a*, *nebo*, ... atd.)

2. *Syntax.* Každé slovo kategorie K je výrazem kategorie K. Jediným pravidlem pro kombinaci výrazů je pak následující: Je-li *Y* výraz kategorie K_1/K_2 a je-li *Z* výraz kategorie K_2, je *Y*(*Z*) výrazem kategorie K_1. (Takže je-li *dramatik* výraz kategorie **V**/**T** a *Jágr* výraz kategorie **T**, je *dramatik*(*Jágr*) výrazem kategorie **V**. Podobně je-li například *nebo* výraz kategorie (**V**/**V**)/**V** a *V*, *V´* výrazy kategorie **V**, je *nebo*(*V*) výrazem kategorie **V**/**V**; a (*nebo*(*V*))(*V´*) je tedy výrazem kategorie **V**.)

3. *Denotáty slov*. Předpokládáme, že ke každé primitivní kategorii K je dána množina D_K; tzv. *doména* kategorie K. (Tak například doménou D_T přiřazenou kategorii **T** může být nějaká daná množina ‘individuí‘ a doménou D_V přiřazenou kategorii **V** množina {P,N} dvou pravdivostních hodnot). Toto přiřazení rozšíříme na všechny kategorie tak, že za $D_{B/A}$ vezmeme množinu všech funkcí z D_A do D_B, kterou budeme značit $[D_A \Rightarrow D_B]$. (Takže $D_{V/T}$ bude množina $[D_T \Rightarrow D_V]$ funkcí přiřazující individuím pravdivostní hodnoty; zatímco $D_{(V/V)/V}$ bude množina $[D_V \Rightarrow [D_V \Rightarrow D_V]]$ funkcí přiřazující pravdivostním hodnotám funkce přiřazující pravdivostním hodnotám pravdivostní hodnoty.) Předpokládáme, že je-li *Y* výrazem kategorie K, je mu přiřazen denotát $\|Y\| \in D_K$. (Tak například $\|dramatik\|$ může být funkcí přiřazující všem dramatikům P a všem ostatním individuím N; $\|$Jágr$\|$ může být individuum Jaromír Jágr; a $\|nebo\|$ může být funkcí, která přiřadí hodnotě N identickou funkci a hodnotě P funkci přiřazující každé hodnotě P.)

4. *Denotáty složených výrazů*. Denotát $\|Y(Z)\|$ složeného výrazu *Y*(*Z*) je dán jako hodnota $\|Y\|(\|Z\|)$ aplikace funkce $\|Y\|$ na argument $\|Z\|$. (Takže například $\|dramatik(Jágr)\|$ bude hodnotou funkce $\|dramatik\|$ pro argument $\|Jágr\|$, to jest – protože Jágr není dramatik – pravdivostní hodnota N; $\|nebo(dramatik(Jágr))\|$ bude $\|nebo\|(\|dramatik(Jágr)\|) = \|nebo\|$(N), to jest identická funkce; a $\|nebo(dramatik(Jágr))(dramatik(Jágr))\| = \|nebo(dramatik(Jágr))\|(\|dramatik(Jágr)\|) = \|nebo(dramatik(Jágr))\|$(N) = N.)

5.4.Lambda-abstrakce

Kategoriální gramatika bývá obohacována o další pravidlo, které ji dodává větší flexibilitu, takzvané pravidlo lambda-abstrakce. Představme si, že vezmeme nějaký složený výraz a ‘uděláme do něj díru‘, tj. odstraníme z něj nějakou složku a nahradíme ji nějakým dohodnutým symbolem, třeba písmenem „*x*“ (tomu budeme říkat, jak je zvykem v logice, *proměnná*). Takovému ‚děravému‘ výroku pak říkáme *matrice*. Tak z výroku *dramatik*(*Jágr*) můžeme udělat matrici *dramatik*(*x*). (Předpokládáme, že pro zaplňování míst po různých kategoriích výrazů máme různé proměnné; budeme-li tedy pro termy používat *x*, *y*, ..., pak pro predikáty budeme používat třeba *p*, *q*, ... – a budeme tedy mít například matrici *p*(*Jágr*)). Zaplňujeme-li potom díru v matrici různými výrazy stejné kategorie, jaké byl původně odstraněný výraz, dostáváme různé výsledky; a matrice tedy představuje ‚předpis‘ určitého přiřazení či funkce. Například matrice *dramatik*(*x*) může být chápána jako předpis funkce

Jágr ⟶ *dramatik*(*Jágr*)
Havel ⟶ *dramatik*(*Havel*)
...

Taková funkce přiřazuje výrazy výrazům; paralelně však můžeme uvažovat i o odpovídající funkci na úrovni denotátů, tedy v našem případě o funkci

$\|Jágr\|$ ⟶ $\|dramatik(Jágr)\|$
$\|Havel\|$ ⟶ $\|dramatik(Havel)\|$
...

tedy vlastně o funkci

Jaromír Jágr ⟶ N
Václav Havel ⟶ P
...

Myšlenka lambda abstrakce je založena na tom, že zavedeme nový druh výrazu, za jehož význam definitoricky stanovíme právě tuto funkci. Nechť je tedy výše naznačená funkce denotována výrazem $\lambda x.dramatik(x)$. Obecněji, je-li *Y* matrice, nechť $\lambda x.Y$ denotuje funkci *f* takovou, že $f(\|Z\|) = \|Y^{x \leftarrow Z}\|$, kde $Y^{x \leftarrow Z}$ značí variantu výrazu *Y*, ve které byl symbol *x* nahrazen výrazem *Z*. To znamená, že $\|(\lambda x.Y)(Z)\| = \|Y^{x \leftarrow Z}\|$, a výrazy $(\lambda x.Y)(Z)$ a $Y^{x \leftarrow Z}$ jsou tedy z hlediska sémantiky ekvivalentní; ten první, složitější, tudíž můžeme v rámci analýzy kdykoli nahradit tím druhým, jednodušším. Pravidlo nahrazení složitějšího výrazu $(\lambda x.Y)(Z)$ jednodušším $Y^{x \leftarrow Z}$ nazýváme *pravidlem lambda-konverze*. Podle tohoto pravidla můžeme například $(\lambda x.dramatik(x))(Havel)$ převést na *dramatik*(*Havel*) (protože $\|(\lambda x.dramatik(x))(Havel)\| = \|dramatik(Havel)\|$).

Výraz $\lambda p.p(Jágr)$ pak analogicky označuje funkci, která každému prvku $[D_T \Rightarrow D_V]$ přiřadí prvek D_V, to jest pravdivostní hodnotu. Konkrétně jde o funkci, která dané funkci *f* přiřadí P právě tehdy, když $f(\|Jágr\|)$=P, to jest když *f* přiřazuje Fregovi hodnotu P – takže $\lambda p.p(Jágr)$ vlastně označuje něco jako ‘množinu všech množin, do kterých patří Jágr’.

Kategoriální gramatice obohacené o pravidlo lambda-konverze budeme říkat (spolu s Cresswellem, 1973) *lambda-kategoriální gramatika*. (Logický kalkulus založený na tomto typu jazyka je pak znám jako *typovaný lambda-kalkul*[31].)

5.5.Možné světy: intenzionální sémantika

Jak jsme konstatovali, extenzi nelze brát za přijatelnou explikaci významu v intuitivním slova smyslu. Výroky *Praha je město* a *Havel je dramatik* jsou oba pravdivé a mají tedy tutéž extenzi, avšak jistě nemají tentýž význam. V čem spočívá odlišnost jejich významů? Jednou z odpovědí je to, že ačkoli mají stejnou pravdivostní hodnotu, je možné, aby ji stejnou neměly. Mohl by jistě existovat svět, ve kterém by Havel byl dramatik, ale Praha byla pouhou vesnicí, či svět, kde by Praha byla městem, ale Havel byl třeba dělníkem v pivovaru (ten poslední by dokonce odpovídal jednomu z předchozích stádií našeho současného světa). Z tohoto pohledu tedy znát význam výrazu vyžaduje znát nikoli jeho aktuální extenzi,

[31] Viz Zlatuška (1993).

ale být schopen určit jeho extenzi ve kterémkoli možném světě. Takové úvahy stojí v základě návrhů na explikaci pojmu intenze, vycházejících z prací Carnapa, Kripka a dalších: intenze výrazu je funkce, která každému možnému světu přiřadí extenzi tohoto výrazu v tomto možném světě. Tak intenzí výrazu „prezident ČR“ je funkce, která každému možnému světu přiřadí tamního prezidenta ČR (pokud tam existuje), intenzí výrazu „dramatik“ je funkce, přiřazující každému možnému světu množinu tamních dramatiků, a intenzí výroku „Prezident ČR je dramatik“ je funkce, přiřazující P těm možným světům, v nichž je tamní prezident ČR dramatik, a N těm ostatním.

Chceme-li tuto myšlenku realizovat, vidíme, že nyní namísto extenzí, které jsme potřebovali v rámci extenzionální sémantiky, potřebujeme funkce, které mají za definiční obor množinu všech možných světů a za obory hodnot extenze. Takže namísto objektů z D_T potřebujeme objekty $[MS \Rightarrow D_T]$ (kde MS je množina všech možných světů), namísto $[D_T \Rightarrow D_V]$ potřebujeme $[MS \Rightarrow [D_T \Rightarrow D_V]]$ atd. Náš sémantický systém tedy musíme nějak obohatit o množinu MS a o funkce s definičním oborem rovným této množině[32].

Přiřadíme-li ovšem ‘podmětům‘ prvky $[MS \Rightarrow D_T]$, přísudkům prvky $[MS \Rightarrow [D_T \Rightarrow D_V]]$ a výrokům prvky $[MS \Rightarrow D_V]$, nebudeme již zřejmě moci nahlédnout denotát jednoduché věty jako výsledek aplikace denotátu jejího přísudku na denotát jejího podmětu. Její denotát bude výsledkem poněkud komplikovanější operace s denotáty jejích částí. Tímto denotátem totiž zřejmě bude funkce, která dává pro každý možný svět hodnotu, která je výsledkem aplikace hodnoty denotátu jejího přísudku *pro tento možný svět* na hodnotu denotát jejího podmětu *pro tento možný svět*. To znamená, že je-li *w* možný svět, bude

$$\|P(T)\|(w) = (\|P\|(w))(\|T\|(w))$$

Z tohoto pohledu se intenzionální sémantika jeví prostě jenom jako *extenzionální sémantika prováděná pro všechny možné světy*. Situace ovšem není tak triviální. Jak se ukazuje, může být někdy k výpočtu denotátu komplexního výrazu v daném možném světě potřeba nejenom extenze jeho komponent v tomto možném světě. Vezměme si například větu

(1) *Jágr hledá prezidenta ČR.*

Tato věta se, jak se zdá, skládá z binárního predikátu (to jest výrazu kategorie **V**/**T**,**T** – či ve zjednodušené, ‚unární‘ variantě kategoriální gramatiky (**V**/**T**)/**T**) *hledat* aplikovaného na dva výrazy kategorie **T**, *Jágr* a *prezident ČR*. Zdá se tedy, že její pravdivostní hodnota v daném možném světě by měla být dána jako aplikace hodnoty denotátu jejího predikátu v tomto možném světě (což by měl být prvek $[D_T \Rightarrow [D_T \Rightarrow D_V]]$) na hodnoty denotátů termů v tomto možném světě (prvky D_T). Avšak Jágr může jistě hledat prezidenta ČR i v možném světě, ve kterém žádný takový prezident neexistuje (představme si například, že Česko je v tomto světě monarchií, což ovšem Jágr, který je v tomto světě třeba ugandským cestovatelem, neví).

Právě proto, že nyní pracujeme s *intenzionální* sémantikou, nabízí se ovšem řešení. Můžeme *Hledat* nahlédnout nikoli jako vztah mezi extenzemi (individui), ale jako vztah mezi extenzí a intenzí, takže $\|(Hledat(N_1))(N_2)\|(w)$ nebude $((\|Hledat\|(w))(\|N_1\|(w)))(\|N_2\|(w))$, ale $((\|Hledat\|(w))(\|N_1\|(w)))(\|N_1\|)$. Pak ovšem bude *Hledat* nikoli prvkem množiny $[MS \Rightarrow [D_T \Rightarrow [D_T \Rightarrow D_V]]]$, ale množiny

[32] Co to možné světy jsou? O tom existuje mnoho teorií (viz například Materna a Štěpán, 2000). Avšak my se ne vždy, když provozujeme formální sémantiku, musíme pokoušet na tuto otázku odpovídat – podobně jako se často nemusíme zabývat odpovídáním na otázku, co to jsou individua.

[MS⇨ [D_T⇨ [[MS⇨ D_T]⇨ D_V]]]. Pokud výraz, který je součástí nějakého výroku, přispívá k pravdivostní hodnotě tohoto výroku v každém možném světě jen svou extenzí, budeme říkat, že je v tomto výroku v supozici (postavení) *de re*, jinak budeme říkat, že je v supozici *de dicto*[33]. (Tak výraz *prezident ČR* je jako předmět slovesa *Hledat* v supozici *de dicto*; zatímco jako předmět slovesa *najít* by byl v supozici *de re*. To je dáno tím, že hledat můžeme i něco co neexistuje, takže hledání musí být analyzováno jako vztah k intenzi, zatímco najít můžeme jen to, co skutečně existuje, takže nalézání může být bráno jako vztah k extenzi.)

Od extenzionální lambda-kategoriální gramatiky k její intenzionální verzi vedou dvě technicky odlišné cesty (první z nich se vydal již zmíněný Montague, druhou pak již také zmiňovaný Tichý). Ta první spočívá v tom, že intenze chápeme jako něco, co mají výrazy *navíc* kromě svých standardních denotátů (extenzí): takže výraz kategorie K má přiřazen jednak prvek D_K (denotát či extenzi) a jednak prvek [MS⇨ D_K] (smysl či intenzi), a zatímco extenze složeného výrazu se z extenzí jeho částí počítá standardním fregovským způsobem, jeho intenze se počítá tak, že se spočítají extenze ve všech možných světech. Pro výrazy v supozici *de dicto* je pak zaveden mechanismus, který posune jejich intenzi do role jejich extenze. V rámci této varianty intenzionální sémantiky se tak formálně stále pracuje s extenzemi, avšak do role extenze výrazu lze dočasně dosadit jeho intenzi. Bude-li tedy $^{\wedge}V$ výrazem, který denotuje intenzi výrazu *V*, můžeme intenzi extenzi věty (1) (v aktuálním světě) denotovat výrazem

(*Hledat*(*Jágr*))($^{\wedge}$*prezident ČR*)

a jeho intenzi výrazem

$^{\wedge}$((*Hledat*(*Jágr*))($^{\wedge}$*prezident ČR*)).

Alternativní varianta spočívá v tom, že se soubor základních kategorií našeho extenzionálního lambda-kategoriálního jazyka, který je v typickém případě tvořen kategoriemi **V** a **T**, obohatí o ‘kvazikategorii‘ **S**, které bude jako doména odpovídat množina MS možných světů, a výrazy, které jsou při standardní analýze analyzovány jako kategorie K, budou nadále analyzovány jako kategorie K/**S**. Podměty tedy již nebudou analyzovány jako výrazy kategorie **T**, nýbrž **T**/**S**; přísudky jako kategorie (**V**/**T**)/**S**; a výroky jako kategorie **V**/**S**. V tomto případě budou výrazy přímo denotovat intenze, a rozdíl mezi supozicí *de re* a *de dicto* bude reflektován jako rozdíl mezi hodnotou intenze pro možný svět a touto intenzí samotnou. V rámci takového systému pak můžeme díky tomu, že máme k dispozici proměnné pro možné světy, zapsat intenzi výroku (1) jako

λ*w*.(*Hledat*(*w*))(*Jágr*(*w*),*prezident ČR*)

5.6. Konstrukce a situace: hyperintenzionální sémantika

Intenzionální model se tedy jakožto model významu v intuitivním slova smyslu ukázal být mnohem přiměřenější než model extenzionální. I proti jeho přiměřenosti se ovšem později objevily námitky, zejména v souvislosti se sémantickou analýzou vět o tzv. propozičních postojích. Uvažme větu

(2) *Jágr se domnívá, že jedna a jedna jsou dvě*

[33] Viz Tichý (1996b).

Tato věta je tvořena tranzitivním slovesem, jehož podmětem je jméno a předmětem věta; sémanticky se tedy, zdá se, jedná o vztah mezi denotátem jména a denotátem věty. Předpokládáme-li tedy, že věta *Jedna a jedna jsou dvě* je v této větě v supozici *de dicto* (pokud bychom předpokládali supozici *de re*, problémy, o kterých se chystáme hovořit, by to jenom ještě umocnilo), jedná se o vztah mezi Jágrem (tedy extenzí nebo intenzí jména *Jágr*) a intenzí věty *Jedna a jedna jsou dvě*. Protože tato věta je matematickou pravdou a protože matematické pravdy nezávisejí na stavu světa, bude extenzí této věty v každém možném světě pravdivostní hodnota P; a její intenzí tedy bude konstantní funkce přiřazující P každému možnému světu. Tatáž funkce ale zřejmě bude intenzí *jakékoli* matematické pravdy, takže dosadíme-li do (2) za *Jedna a jedna jsou dvě* jakýkoli jiný pravdivý matematický výrok, třeba *Existuje nekonečně mnoho prvočísel*, dostaneme, podle intenzionální analýzy, větu, která říká totéž, co (2). Takže podle intenzionální analýzy by nemohlo nastat, že by byla například věta (2) pravdivá, zatímco věta

(3) *Jágr se domnívá, že existuje nekonečně mnoho prvočísel*

nepravdivá. To se zdá být v rozporu s intuicí: zdá se, že Jágr může (v nějakém jiném možném světě, když ne v tom našem) docela dobře vědět, že jedna a jedna jsou dvě, a současně se mylně domnívat, že prvočísel je jenom konečně mnoho.

Řešení tohoto problému se zdá požadovat další ‘jemnější‘ sémantickou analýzu než je analýza intenzionální; proto se potom někdy hovoří o sémantice *hyperintenzionální*. Myšlenkou, která je společná většině hyperintenzionálních systémů, je myšlenka, že je-li intenze výrazu výsledkem nějaké kombinace intenzí jeho částí, pak bychom význam v intuitivním slova smyslu neměli explikovat jako výslednou intenzi, ale jako nějakou formu zachycení samotného procesu kombinace. Nejjednodušší variantou realizace této myšlenky je ztotožnění denotátu složeného výrazu s uspořádanou *n*-ticí tvořenou denotáty jeho částí. To znamená, že jestliže například intenze výroku (1) vznikne určitou kombinací intenzí $\|Jágr\|$, $\|hledat\|$ *a* $\|prezident\ ČR\|$ (konkrétně, jak jsme viděli, to bude funkce, která každému možnému světu přiřadí hodnotu $(\|Hledat\|(w))(\|Jágr\|(w), \|prezident\ ČR\|)$), pak za denotát tohoto výroku budeme považovat uspořádanou trojici $<\|Jágr\|, \|Hledat\|, \|prezident\ ČR\|>$.

Tuto myšlenku je možné dále rozvádět dvojím způsobem. Jednak je možné mít za to, že taková *n*-tice vystihuje něco jako ‚situaci‘, která je danou větou představována. Tak věta *Jágr je dramatik* vyjadřuje situaci tvořenou individuem Jágrem a jemu přináležející vlastností *být dramatikem*; zatímco věta (1) vyjadřuje situaci tvořenou individuem Jágrem, pojmem prezidenta ČR (tj. intenzí výrazu „prezident ČR“) a vztahem *hledat*, který je spojuje. V této variantě ovšem nastane problém s negativními, disjunktivními atd. větami – jakou situaci by totiž vyjadřovala třeba věta *Jágr není dramatik*? Odpovědi na takové a další otázky si klade za cíl poskytnout sofistikovaná teorie situací, která stojí v základě *situační sémantiky* Barwise a Perryho (1983).

Jinou možností je nevidět *n*-tice vyjadřované složenými výrazy jako situace, ale spíše jako něco jako jakési ‚konstrukce‘, které mají primárně co dělat nikoli s tím, jak se věci spojují v rámci světa, ale spíše s tím, jak uživatelé jazyka kombinují významy částí ve významy celků. Podle tohoto pohledu je interpretace věty *Jágr je dramatik*, kterou provádějí mluvčí češtiny, nutně spojena s konstrukcí jí odpovídající intenze z intenzí jejích částí, a význam je třeba explikovat jako právě tuto konstrukci. To ovšem může opět vést k mnohem sofistikovanějším teoriím konstrukcí, než je jejich zachycení pomocí *n*-tic; takovou teorii je možné najít především v pracích Pavla Tichého (1996c; 1988) [34].

Chápání významů jako ‘konstrukcí‘ také představuje potenciální bod kontaktu mezi formální sémantiku a teoriemi různých ‚sémantických struktur‘ vzešlými z rozličných verzí generativní a

[34] Viz též Materna a Štěpán (2000).

transformační gramatiky, jakými je Chomského logická forma či tektogramatická reprezentace. (Lewis, 1972, dokonce na úsvitu hyperintenzionální sémantiky navrhl zcela explicitní propojení: navrhl nahlédnout význam jako stromovou strukturu tohoto typu, na konci jejíchž větví však ‚visejí' intenze.)

5.7. Kontexty a informační stavy: dynamická sémantika

Zhruba v devadesátých letech dvacátého století se pozornost mnoha sémantiků přirozeného jazyka obrátila k jevům, které souvisejí s tím, co by se dalo nazvat jeho *dynamikou* – to jest k sémantickým jevům, které nemůžeme dost dobře vysvětlit, aniž bychom vzali v úvahu to, že jazyk je prostředek diskurzu, který se odvíjí v čase.

Vezměme například zájmena. Jaký je význam zájmena jako třeba „on"? Takové zájmeno pojmenovává nějaké individuum, podobně jako vlastní jméno; ale na rozdíl od vlastního jména nepojmenovává jedno fixní individuum. Avšak na rozdíl od takové singulární fráze, jako je třeba „prezident České republiky", není individuum, které zájmeno pojmenovává, určeno ani možným světem. Je zřejmě určeno spíše *kontextem*, ve kterém je toto zájmeno užito: užiji-li jej například v situaci, kdy ně někoho ukazuji, bude tím pojmenovaným on; a užiji-li jej vzápětí potom, co užiji jméno „Jágr", bude tím pojmenovaným nejspíše Jágr.

Tohle vede k jistému novému pohledu na sémantiku a k novým teoriím formální sémantiky (které jsou někdy předkládány jako nadstavba intenzionální či hyperintenzionální sémantiky a někdy jako jejich alternativa.) Můžeme říci, že tak jako je intenzionální sémantika postavena na pojmu možného světa, je dynamická sémantika postavena na pojmu *kontextu* či *informačního stavu*[35]. Výroky v jejím rámci jsou chápány jako denotující nikoli pravdivostní hodnoty či funkce z možných světů do pravdivostních hodnot, ale funkce z informačních stavů do informačních stavů, tzv. *přechody* (*updates*). To vychází z myšlenky, že z dynamického hlediska je výrok něčím, co je užíváno v nějakém informačním stavu s cílem tento informační stav nějak změnit.

Existují opět dvě hlavní linie konkretizace této myšlenky. Jedna z nich se odvíjí v duchu pojmového rámce situační sémantiky a je reprezentována především Kampovou *teorií reprezentace diskurzu* (*DRT* – viz Kamp a Reyle, 1993). Tato teorie se opírá o struktury podobné situacím (kterým se nyní ovšem říká *struktury reprezentace diskurzu*), avšak soustředí se především na jejich ‚kinematiku', to jest na to, jak se v průběhu diskurzu rozrůstají. Můžeme tedy říci, že v rámci DRT jsou kontexty či informační stavy uchopeny jako ‚reprezentované situace' a věta je chápána jako prostředek přebudování takové reprezentace na nějakou reprezentaci bohatší.

Jiná cesta rozpracování myšlenky vět jako reprezentujících přechody se, podobně jako intenzionální sémantika, opírá explicitněji o logiku a tím se stala výrazným stimulem pro rozvoj tzv. *dynamických logik* (*viz* van Benthem, 1997).

5.8. Přirozený jazyk: sémantika věty

Předpoklady o syntaxi jazyka, které jsme používali při načrtnutí principů formální sémantiky, byly na hranici trivializace. Naznačme nyní alespoň krátce, s jakými potížemi se musí formální sémantika vypořádávat, chce-li poskytnout skutečně realistický sémantický model přirozeného jazyka.

---

[35] Nejpřímočařejší (ale víceméně triviální) možností, jak takovouto dynamickou sémantiku naroubovat na sémantiku intenzionální, je ztotožnit kontexty s množinami možných světů. Užití věty V v kontextu reprezentovaném množinou M možných světů pak vede ke kontextu reprezentovaném množinou $M \cap \| V \|$ – užití věty tedy jakoby z kontextu vylučuje světy nekompatibilní s tím, co tato věta tvrdí.

Předpokládali jsme, že typická jednoduchá věta se skládá z podmětu a přísudku. Avšak přísudek můžeme často dále rozkládat, například na tranzitivní sloveso a předmět (*hledat + prezident ČR*). Obecně lze jednoduchou větu nahlédnout jako sloveso (které může být modifikováno různými příslovečnými určeními) spojené s různými druhy jmenných doplnění (podmětem, předměty). Z tohoto hlediska se může zdát být případné vidět základní kostru jednoduché věty obecně jako *n*-ární predikát aplikovaný na *n* termů. Problém je však v tom, že počet jmenných doplnění slovesa se může větu od věty měnit (přičemž absence některých z nich může znamenat absenci příslušných argumentů na úrovni sémantiky, zatímco absence jiných jenom jejich nevyjádřenost). Tak například z věty

(4) *Karel přednáší posluchačům báseň*

se může zdát, že predikát, který je třeba k analyzování slovesa *přednášet* by měl být ternární; což se nejeví být neudržitelné ani z hlediska věty

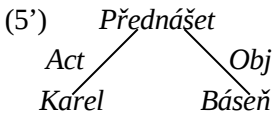
(5) *Karel přednáší posluchačům*

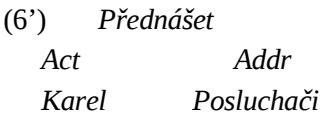
(protože, jak se zdá, i v tomto případě musí existovat něco, co Karel přednáší, i když to ve větě není explicitně zmíněno), avšak nikoli z hlediska věty

(6) *Karel přednáší báseň*

(protože tady se může jednat o skutečnou absenci kohokoli, komu by Karel přednášel).

Slovesa přirozeného jazyka se tedy zdají mít proměnlivý počet argumentů. Navíc na rozdíl od predikátů běžných formálních jazyků jsou tyto argumenty identifikovány komplikovanějším způsobem než jenom pořadím: sloveso může mít ve dvou větách stejný počet argumentů, a přesto se může jednat o argumenty různého typu. (Viz například (5) a (6).) V lingvistických analýzách bývá struktura těchto vět rozlišována označením hran příslušných stromů:





Z hlediska těchto diskrepancí mezi slovesy přirozeného jazyka a predikáty standardních formálních jazyků máme několik možností:

(i) Můžeme trvat na tom, že spojuje-li se totéž sloveso s různým počten argumentů, jde o případ homonymie a je tedy v pořádku, že pro jeho analýzu musíme v každém případě použít jiný predikát (s jinou aritou). To se ale zdá být v příkrém rozporu s intuicí: nezdá se, že by se například v případě (4) a (6) jednalo o dva různé smysly slovesa *přednášet*.

(ii) Můžeme pracovat s formálním jazykem, jehož predikáty mohou mít proměnný počet argumentů (nebo jejichž argumenty nejsou přímo označeními individuí, ale označeními třeba množin individuí). To se nezdá být principiálně neuskutečnitelné, faktem ovšem je, že se jazyky takovéhoto druhu běžně v sémantických teoriích nevyskytují. Navíc se to nezdá řešit problém, že bychom potřebovali zachytit i *typ* vztahu mezi slovesem a jeho jmenným doplněním.

(iv) Mohli bychom ale zvolit zcela jinou strategii a strom znázorňující strukturu věty převést na logickou formuli takovým způsobem, že by se uzly staly termy a označení hran by se stala predikáty. Tak by z (6) vyšlo

$Act(Přednášet,Karel) \wedge Obj(Přednášet,Báseň)$.

Sofistikovanější variantou by bylo předpokládat, že to, o čem věta hovoří, je možné nahlédnout jako nějakou ‘událost‘ (např. událost Karlova čtení básně) a tuto větu tedy analyzovat jako tvrzení existence takové události: $\exists u(Přednášení(u) \wedge Act(u,Karel) \wedge Obj(u,Báseň))$, případně ještě lépe

$\exists u(Přednášení(u) \wedge Act(u,Karel) \wedge \exists x(Báseň(x) \wedge Obj(u,x)))$.

Z hlediska dynamiky diskurzu je potom ve větě třeba rozlišit část, kterou se věta ‘ukotvuje v kontextu‘ (*východisko*, to jest specifikace toho, *o čem* se hovoří) od části, která přináší skutečně novou informaci (*jádro výpovědi*, to jest vyjádření toho, *co* se o tom říká). Tyto dvě části se charakteristicky liší mimo jiné tím, že existence něčeho, o čem hovoří východisko, je často podmínkou pro to, aby byla celá věta vůbec interpretovatelná (zatímco jádro pak má obvykle co dělat již jenom s pravdivostí). Řeknu-li *Prezident ČR je hokejista*, bude to, co říkám, prostě nepravda; zatímco řeknu-li *Český král je hokejista*, bude to spíše než nepravda ne úplně smysluplný výrok, protože není jasné, o čem se vůbec mluví (pokud to ovšem neřeknu třeba v kontextu vyprávění o nějaké době, kdy Česko krále mělo).

V rámci diskurzu se navíc obvykle předpokládá, že co se vypovídá, je z hlediska aktuálního kontextu vyčerpávající či alespoň reprezentativní. Tak například odpovím-li na otázku *Kde se mluví Německy?* třeba *V Hamburku*, bude to odpověď, která sice není nesprávná, ale vzhledem k tomu, že Hamburk jistě není z hlediska území, na kterých se Německy hovoří, reprezentativní, to bude odpověď problematická. Podobně odpovím-li *ano* na otázku *Máš jedno dítě?* v případě, že mám děti dvě, nebude to striktně vzato odpověď nepravdivá, bude ale jistě krajně matoucí.

Zjednodušeně tedy můžeme říci, že pro východisko věty je charakteristický předpoklad existence a pro jádro zase předpoklad reprezentativnosti. Přitom východisko se v nejjednodušším případě kryje s podmětem a jádro s přísudkem; jazyk ovšem disponuje prostředky, které nám dovolují toto členění podle aktuálního kontextu měnit, a to zejména prostředky fonetické (intonace, důraz) a slovosledné. Tak řeknu-li *VÁCLAV HAVEL je prezidentem ČR* (s důrazem na *Václav Havel*) či *Prezidentem ČR je Václav Havel*, je to spíše než výpověď o Václavu Havlovi výpověď o úřadu prezidenta ČR, o kterém se říká, že je zastáván Václavem Havlem. To může znamenat rozdíl, který má co dělat spíše s pragmatikou než se sémantikou, avšak v případě některých druhů vět to může ovlivňovat i sémantiku (viz *V Hamburku se*

mluví v Německu versus *Německy se mluví v Hamburku*; nebo *Každý muž sní o jedné ženě* vs. *O jedné ženě sní každý muž*.)

To naznačuje, že ze sémantického hlediska by mohlo být rozumné považovat za podmět a přísudek ne vždy nutně to, co je podmětem a přísudkem z hlediska syntaxe, a navíc že způsob, jak se denotáty podmětu a přísudku kombinují v denotát věty by měl zohlednit jak požadavek existence, tak požadavek reprezentativnosti. Jedním ze způsobů, jak toto implementovat v rámci intenzionální sémantiky, je zajistit, aby byly extenzí podmětu i přísudku (v každém možném světě) vždy množiny stejného typu; a věta, která vznikne jejich kombinací pak bude pravdivá jestliže (i) ta první množina je neprázdná (předpoklad existence), (ii) ta druhá v ní bude obsažena (‘přísudek o podmětu platí‘) a (iii) bude tvořit ‚reprezentativní část‘ té první (předpoklad reprezentativnosti). V případě věty *Německy se mluví v Hamburku* by východisko *německy se mluví* specifikovalo množinu oblastí, kde se mluví Německy a jádro *v Hamburku* by specifikovalo množinu tvořenou Hamburku (a věta by tedy byla problematická proto, že ta druhá není ‘reprezentativní částí‘ té první). V případě *V Hamburku se mluví německy* by byla první množina tvořena aktuálně relevantními vlastnostmi Hamburku a ta druhá vlastností *mít za jazyk němčinu* (a protože v kontextu řeči o tom, jak se kde mluví, může být to, že se tam mluví německy, *jedinou* relevantní vlastností Hamburku, může být v takovém případě předpoklad reprezentativnosti splněn.)

5.9. Diskurs

6. Lingvistické korpusy

6.1. Úvod a historie

6.2. Typy a realizace korpusů

6.3. Úloha korpusů v lingvistickém výzkumu

6.4. Příklady korpusů

6.5 Jiné zdroje dat pro lingvistický výzkum a aplikace

Literatura

Colmerauer A.: Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Mimeo, Montréal, 1969.

Literatura

A. K. Joshi, L. S. Levy, M. Takahashi: Tree adjunct grammars. Journal Computer Systems Science, 10(1), 1975.
S. Sheiber, Y. Schabes: Synchronous tree adjoining grammars. In Proceedings of the 13th International Conference on Computational Linguistics (COLING ’90), Helsinki, August 1990. Association for Computational Linguistics.

Literatura

Daneš, Hlavsa (1981)
Fillmore (1968, 1977)
Hajičová (1979) - slovesa
Hajičová et al (2000)
Hajičová et al (2002) – jak citovat ???, díl I.

Novotný (1980) - subst
Pala a Ševeček (SSJČ, SSČ, SČS)
Panevová (1966)
Panevová (1974-1975)
Panevová (1980) a (1994) - slovesa
Panevová (1992)
Panevová (1998) – adj
Panevová (2000) - subst
Panevová a Řezníčková (2001)
Pauliny (1943)
Piťha (1982) - adj
Prouzová (1983) – adj
Sgall (1967) - FGD
Sgall (1992) – Rozencvejg
Sgall, P. et al (1986) Úvod do syntaxe a sémantiky
Skoumalová (2001)
Straňáková-Lopatková (2001)
Straňáková-Lopatková a Žabokrtský (2002) – ještě neexistuje
Svozilová, N., Prouzová, H., Jirsová, A. (1997) Slovesa pro praxi, Academia, Praha.
Tesnièra (1959)

V. Petkevič: A New Formal Specification of Underlying Structure. Theoretical Linguistics Vol.21, No.1, 1995.
P. Sgall (1967) Generativní popis jazyka a česká deklinace. Academia, Praha.
1. díl

Literatura

Barwise J. a J. Perry (1983): *Situations and Attitudes*, MIT Press, Cambridge (Mass.).

van Benthem, J. (1997): *Exploring Logical Dynamics*, CSLI, Stanford.

van Benthem, J. a A. ter Meulen, eds. (1997): *Handbook of Logic and Language*, Elsevier/MIT Press, Oxford/Cambridge(Mass.).

Carnap, R. (1947): *Meaning and Necessity*, University of Chicago Press, Chicago.

Cmorej, P. (2001): *Úvod do logickej syntaxe a sémantiky*, IRIS, Bratislava.

Cresswell, M.J. (1973): *Logic and Languages*, Meuthen, London.

Chomsky, N. (1986): *Knowledge of Language*, Praeger, Westport.

Jackendoff, R. (1990): *Semantic Structures*, MIT Press, Cambridge (Mass.).

Ježek, J. (1976): *Univerzální algebra a teorie modelů*, SNTL, Praha.

Kamp, H. a U. Reyle (1993): *From Discourse to Logic*, Kluwer, Dordrecht.

Kripke, S. (1963): 'Semantical Considerations on Modal Logic', *Acta Philosophica Fennica* 16, 83-94.

Lakoff, G. (1971): 'On Generative Semantics', in *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology* (ed. D. D. Steinberg a L. A. Jakobovits), Cambridge.

Materna, P. a J. Štěpán (2000?): *Filosofická logika: nová cesta*, Nakladatelství Olomouc, Olomouc.

Montague, R. (1974): *Formal Philosophy* (*Selected Papers*; ed. R. Thomason), Yale University Press, New Haven.

Partee, H. a R. Hendriks (1997): 'Montague Grammar', in van Benthem a ter Meulen (1997).

Peregrin, J. (1998): *Úvod do teoretické sémantiky*, Karolinum (skripta FF UK), Praha.

Peregrin, J. (1999): *Význam a struktura*, Oikoymenh, Praha.

Sgall, P., E. Hajičová a J. Panevová (1986): *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, Academia, Praha.

Sochor, A. (2001): *Klasická matematická logika*, Karolinum, Praha.

Tichý, P. (1988): *The Foundations of Frege's Logic*, de Gruyter, Berlin.
Tichý, P. (1996a): *O čem mluvíme?* (*Vybrané stati k logice a sémantice*), Filosofia, Praha.
Tichý, P. (1996b): '*De dicto* a *de re*', in Tichý (1996a).
Tichý, P. (1996c): 'Konstrukce', in Tichý (1996a).
Zlatuška, J. (1993): *Lambda-kalkul*, Masarykova univerzita, Brno.