

Úvod do počítačové lingvistiky (pro informatiky)

Úvod do Úvodu

Co je počítačová lingvistika?

Obor, zabývající se formálním popisem vlastností přirozených jazyků a jejich automatickým zpracováním (vytváření automatických systémů, modelujících užívání přirozeného jazyka).

Je to mezní obor, využívající výsledků (a také přispívající k dalšímu rozvoji):

- teoretické lingvistiky
- teoretické informatiky (viz Chomského hierarchie jazyků)
- umělé inteligence
- psychologie
- logiky
- matematiky (statistiky)

Jiné názvy: matematická lingvistika, algebraická lingvistika, statistická lingv.

Podobory (počítačové) lingvistiky

- Rozpoznávání a generování mluvené řeči
- Fonetika (zkoumá zvuky, fóny, třídí je a klasifikuje – nauka o **tvorbě** hlásek)
- Fonologie (zabývá se pouze těmi zvukovými rozdíly, které nesou význam, základní jednotkou *foném*, je to nauka o **funkci** hlásek)
- Morfologie (tvarosloví)
- Syntaxe (skladba)
- Sémantika (význam)
- Strojový (automatický) překlad
- Formalismy (syntaktické)
- Korpusová lingvistika
- Statistická lingvistika (dříve kvantitativní, nyní modeluje užívání jazyka)
- ...

Funkce přirozených jazyků

- popis reálných věcí v okolním světě, zavedení pojmů
- objektivní popis abstraktních vztahů a pojmů, zobecňování
- rekurzivní modelování komunikačního partnera
- přijímání nebo zamítání kooperativních řešení
- definice sociálních vztahů mezi partnery (vykání apod.)
- komunikační prostředek o jazyce
- ...

Zásady komunikace v přirozeném jazyce

1. Všeobecnost

Přirozený jazyk je nejšířeji používaný standard komunikace

2. Využitelnost

Formálních jazyků pro výpočetní procesy je příliš mnoho, přirozený jazyk je v sobě obsahuje všechny, nezastarává.

3. Obsah

Přirozený jazyk je samostatný. Vše, o čem se dá komunikovat, je vyjádřitelné přirozeným jazykem.

4. Vágnost

Jistá míra vágnosti je užitečná, je to základ inovativního myšlení.

5. Vícevrstevnost

Řeč může být i předmětem rozhovoru, v komunikaci člověk – stroj umožňuje např. vysvětlující dialogy

6. Zkratkovitost

Dotaz v přirozeném jazyce bývá kratší, srozumitelnější a jednodušší než dotaz pomocí formálního dotazovacího jazyka

Problémy s významovou ekvivalencí

- Karel prodává auta.
Od Karla se kupují auta.
- Nakrájel salám na pět kusů.
Nakrájel ze salámu pět kusů.
- Na Moravě se mluví česky.
Česky se mluví na Moravě.

Víceznačnost a vágnost jazyka

- Bramborové knedlíky a švestkové knedlíky
- Kritika brazilského delegáta byla ostrá.
- V místnosti stojí zelený stůl a židle.
- Na recepci se dostavil i ředitel banky roku.
- Vysoká škola lesnická v Trutnově otevřela novou fakultu.
- Často loví tlouště na višni.
- Když slepice málo snáší, tak se vejce špatně shání.
- Páry vycházejí z lesa.
- Včera jsem viděl Frantu v tramvaji.
- Dědeček se rozložil na gauči.
- Závodnice se před závodem se soupeřkami oddávala sexu.
- Dálnice z Žiliny do Bratislavy postavená Ruskem stála 10 miliard.

Příklady z tisku

- Stát se skvělou vědkyní jde i se dvěma dětmi
- Otec Emmons bude trénovat Australany.
- Loprais upustil kola a ujížděl.
- Padl návrh vyhodit Čunka.
- Důvod nechutného útoku? Nová láska, vůle vrátit se zpět do ringu a [propuštění nevlastního otce ze služeb svého manažera](#).
- Na trase C spadl člověk do kolejiště metra, nahradí ho autobusy
- Padlým námořníkům se znovu rozsvítilo
- V hotelu Corrado se za jeho nejslavnější éry scházely prostitutky. Často tam bydlely špičky ČSSD jako Miloš Zeman, Jiří Paroubek nebo Petr Benda.
- Na českých silnicích umírá více lidí než ve zbytku EU, hůř jsou na tom jen v Polsku

Příklady z tisku (2)

- Házel rohlík na strop v naději, že na něm ulpí
- Počítače nahradí výpravčí s plácačkami.
- Co vám uniklo: policie zničila Rathovy miliony i nejvyšší věž Prahy.
- Zástupci židovské obce se zúčastnili slavnosti odhalení.
- Štěpánka čeká v Šanghaji souboj s Berdychem.
- Soud shledal Hanu P. vinnou, že od června 2010 do srpna 2011 měla v lesíku v Sobědruhách na Teplicku pohlavní styk s tehdy třináctiletým hochem.
- Rosbergovi narazil v dvousetkilometrové rychlosti do helmy pták, srážku přežil bez úhony
- Sarah Palinová řekla Ne, Obamovi se nepostaví.

Historie morfologie

Morfologie je asi nejstarším odvětvím lingvistiky. První počátky se datují 4.stoletím př.n.l., kdy indický lingvista [Pānini](#) formuloval 3 959 pravidel morfologie sanskrtu v díle [Ashtadhyayi](#). Jeho popis se blížil tzv. složkové gramatice.

Ashtadhyayi je nejstarší známou gramatikou sanskrtu a zároveň také nejstarší známou prací deskriptivní lingvistiky, generativní lingvistiky, a společně s prací jeho bezprostředních předchůdců ([Nirukta](#), [Nighantu](#), [Pratishakyas](#)) prakticky znamená začátek lingvistiky jako vědního oboru.

Řecko-římská gramatická tradice se také zabývala morfologickou analýzou, protože stará řečtina a latina patřily mezi flektivní jazyky.

Panini měl také značný vliv na moderní lingvistiku, otec strukturalismu, [Ferdinand de Saussure](#), byl profesorem sanskrtu.

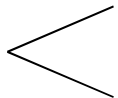
Morfologie

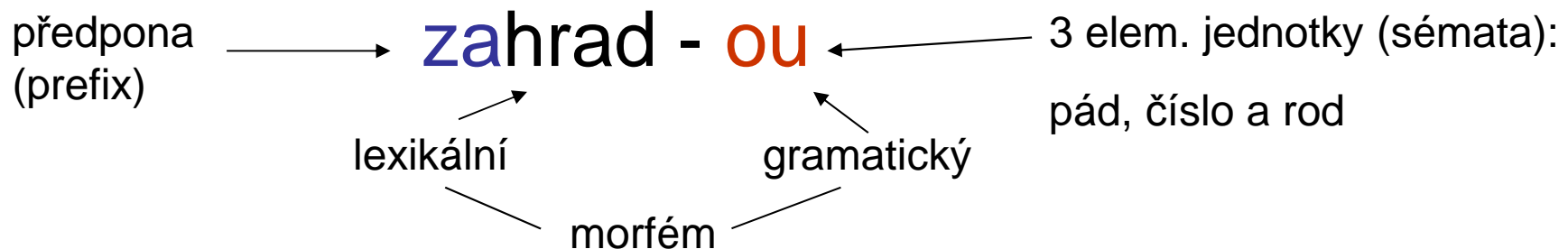
Předmětem morfologie je studium vnitřní struktury slov.

Lexikologie - slova jsou studována jako jednotky slovní zásoby

Lexikografie – sestavování slovníků

Základní jednotkou studia morfologie je **morfém** – nejmenší znaková jednotka jazyka nesoucí význam

Morfém  **lexikální** – nese význam slova jako takového
gramatický – určuje gramatickou roli slovního tvaru



Další pojmy

Morfologie studuje způsoby **skloňování** (deklinace) a **časování** (konjugace).

Tvaroslovné **dublety** – stejné slovní tvary odvozené od dvou nebo více slovních základů (žena, tři, hnát, stát, už apod.)

Alternace – změna hlásek uvnitř kmene

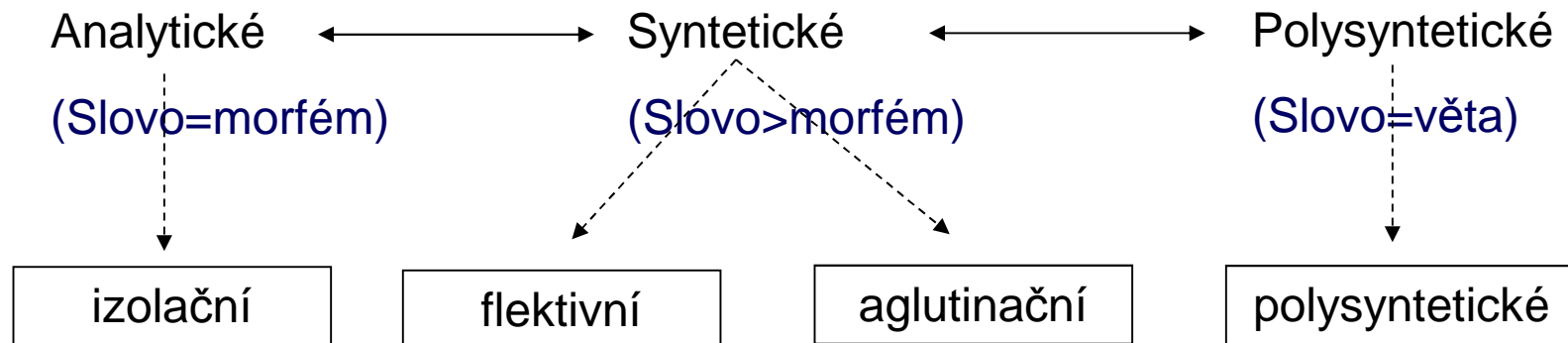
vůz – vozu; švec – ševce; prkno - prken

Alomorfy – varianty kmene odvozené od stejného slovního základu

Autosémantická (plnovýznamová) slova

Synsémantická (pomocná) slova

Morfologická typologie jazyků



Izolační: vietnamština, čínština

flektivní: latina, stará řečtina, slovanské jazyky

aglutinační: maďarština, japonština

polysyntetické: eskymácké a indiánské jazyky

Přístupy ke zpracování morfologie

Morfologie založená na **morfémech**

- vidí slovo jako řetízek morfémů, jako korálky na niti

Morfologie založená na **lexémech**

- vidí slovo jako výsledek aplikace pravidel, která slovo mění a tím vytváří nový slovní tvar

Morfologie založená na **slovech**

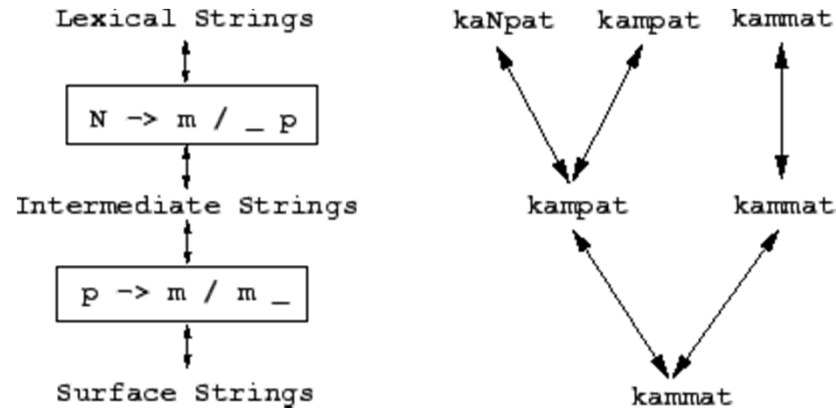
- centrální roli mají **vzory**. Pokud máme základní tvar a známe, ke kterému vzoru slovo patří, dokážeme vygenerovat všechny jeho tvary. Tento přístup je vhodný i tam, kde předchozí dva selhávají, např. tam, kde jeden morfém reprezentuje více gramatických kategorií (např. 3.os.č.j.r.ž).

Two-Level Morphology

Systém zpracování morfologie vyvinutý Lauri Karttunenem a Kimmo Koskeniemi na začátku 80.let.

Jednalo se o první obecný model zpracování morfologie přirozeného jazyka. Byl založen na konečňestavových automatech, pro každý jazyk bylo nutné vytvořit slovník a pravidla, společný byl mechanismus morfologie.

Tradiční počítačové zpracování morfologie se orientovalo na generování výsledných tvarů slov z nějakého tvaru základního a nebralo příliš v úvahu, že opačný směr (analýza) může být víceznačný.

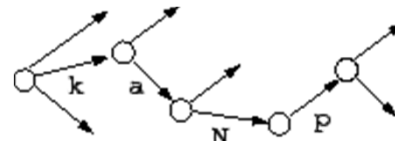
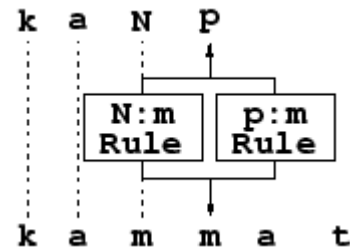


Two-Level Morphology

2 úrovně – lexikální a povrchová

Základní myšlenky:

- pravidla se aplikují paralelně, nikoli sekvenčně
- podmínky se mohou vztahovat k jedné z úrovní nebo k oběma zároveň
- lexikální vyhledávání (trie, letter tree) a morfologická analýza probíhají současně



Česká morfologie

Vyvíjena od r.1989 zejména prof.Hajičem. Využívá pozičních značek, každá pozice má svůj jednoznačně určený význam. Značky jsou 15 místné, ovšem rozeznává se pouze 13 kategorií:

Category	# of values	Example(s)
POS	10	N (noun), Z (punctuation)
SUBPOS	75	P (personal pron.), U (possessive adj.)
GENDER	8	I (masc. inanimate), X (any), - (N.A) -----
NUMBER	4	P (plural), D (dual)
CASE	9	1 (nominative), 6 (locative)
POSSGENDER	4	M (masc. animate), F (feminine) ↓
POSSNUMBER	3	S (singular), P (plural)
PERSON	5	1 (first), ...
TENSE	4	P (present), M (past)
GRADE	5	3 (superlative)
NEGATION	3	A (affirmative), N (negative)
VOICE	3	A (active), P (passive)
VAR	11	1 (1 st variant), 6 (colloq. style), 8 (abbrev.)

Česká morfologie

- Značka: 13 kategorií

– Příklad: AAFP3-----3N-----

Adjective

Regular

Feminine

Plural

Dative

no poss. Gender

no poss. Number

no person

no tense

superlative

negated

no voice

reserve1

reserve2

base var.

Ex.: nejnezajímavější
“(to) the most uninteresting”

- Lemma: jednoznačný identifikátor
 - Stát/slov. -> stát-1, šel -> jít

Česká morfologická analýza - příklad

Vstup:

Prezident rezignoval na svou funkci.

Výstup:

```
<csts>
<f cap>Prezident<MMI>prezident<MMt>NNMS1-----A----
<f>rezignoval<MMI>rezignovat_:T<MMt>VpYS---XR-AA---
<f>na<MMI>na<MMt>RR--4-----<MMt>RR--6-----
<f>svou<MMI>svůj-1_^(přivlast.)<MMt>P8FS4-----1<MMt>P8FS7-----1
<f>funkci<MMI>funkce<MMt>NNFS3-----A-----<MMt>NNFS4-----A-----
<MMt>NNFS6-----A-----
<D>
<d>.<MMI>.<MMt>Z:-----
</csts>
```

Činnosti využívající morfologii

Morfologická analýza – výsledkem je seznam lemmat a značek popisujících jednotlivé kombinace gramatických kategorií spjatých s daným vstupním slovním tvarem.

Morfologické značkování (tagging) – proces výběru jediné správné značky v daném kontextu (statistické metody).

Částečná morfologická desambiguace založená na pravidlech (Oliva, Petkevič) – pomocí spolehlivých pravidel redukuje počet značek, odstraňuje nevhodné, ponechává všechny, které nelze spolehlivě odstranit.

Lemmatizace – proces výběru (správného) základního tvaru, ze kterého byl odvozen daný vstupní tvar. Je to klíčová operace pro vyhledávání v textech.

Stemming – odříznutí koncovky, na rozdíl od lemmatizace je základním tvarem kmen slova. Populární je tzv. Porterův stemmer.

Generování – proces výběru správného slovního tvaru, pokud známe lemma a příslušnou kombinaci gramatických kategorií.

Text

Morfologická analýza

Značkování

Fantastickým	fantastický AAFP3----1A---- AAIP3----1A---- AAIS6----1A---7 AAIS7----1A---- AAMP3----1A---- AAMS6----1A---7 AAMS7----1A---- AANP3----1A---- AANS6----1A---7 AANS7----1A----	fantastický AAIS7----1A----
finišem	finiš NNIS7----A----	finiš NNIS7----A----
si	být VB-S---2P-AA--7 se_ ^ (zvr._zájmeno/částice) P7-X3-----	se_ ^ (zvr._zájmeno/částice) P7-X3-----
však	však J^-----	však J^-----
Neumannová	Neumannová_;S NNFS1----A---- NNFS5----A----	Neumannová_;S NNFS1----A---- -
doběhla	doběhnout_:W VpQW---XR-AA—1	doběhnout_:W VpQW---XR-AA-- 1
pro	pro-1 RR--4-----	pro-1 RR--4-----
vytoužené	vytoužený_ ^ (*3it) AAFP1----1A---- AAFP4----1A---- AAFP5----1A---- AAFS2----1A---- AAFS3----1A---- AAFS6----1A---- AAIP1----1A---- AAIP4----1A---- AAIP5----1A---- AAMP4----1A---- AANS1----1A---- AANS4----1A---- AANS5----1A----	vytoužený_ ^ (*3it) AANS1----1A-- -- (AANS4----1A----)
olympijské	olympijský AAFP1----1A---- AAFP4----1A---- AAFP5----1A---- AAFS2----1A---- AAFS3----1A---- AAFS6----1A---- AAIP1----1A---- AAIP4----1A---- AAIP5----1A---- AAMP4----1A---- AANS1----1A---- AANS4----1A---- AANS5----1A----	olympijský AANS1----1A---- (AANS4----1A----)
zlato	zlato NNNS1----A---- NNNS4----A---- NNNS5----A----	zlato NNNS1----A---- (NNNS4-- ---A----)
.	. Z:-----	. Z:-----

Aplikace morfologie

Kontrola překlepů

Požadavky:

1. Musejí být nalezeny všechny překlepy a nalezené musejí být opraveny
2. Musejí být přezkoušeny kontextové podmínky korigované verze
3. Slova, která systém nezná, by neměla být hlášena jako chyby
4. Systém by neměl dávat falešná chybová hlášení
5. Korektura musí být co nejvíce automatická
6. Čas zpracování musí být velmi krátký

Tyto požadavky je v praxi velmi obtížné splnit.

Kontrola překlepů

2 základní metody

1. Porovnávání řetězců se slovy ve slovníku

- buď seznam všech možných slovních tvarů daného jazyka (wordlist)
- nebo slovník lemmat + morfologická analýza

Výhoda: Spolehlivé a jednoduché

Nevýhoda: pomalé, náročné na kvalitu slovníku, na místo, nerozezná chybná slova od neznámých, každé zlepšení musí zařídit autor nebo uživatel

2. Srovnávání skupin znaků (dvojice, trojice,...), hledání nedovolených kombinací znaků

Výhoda: nezávislé na slovníku, rychlé

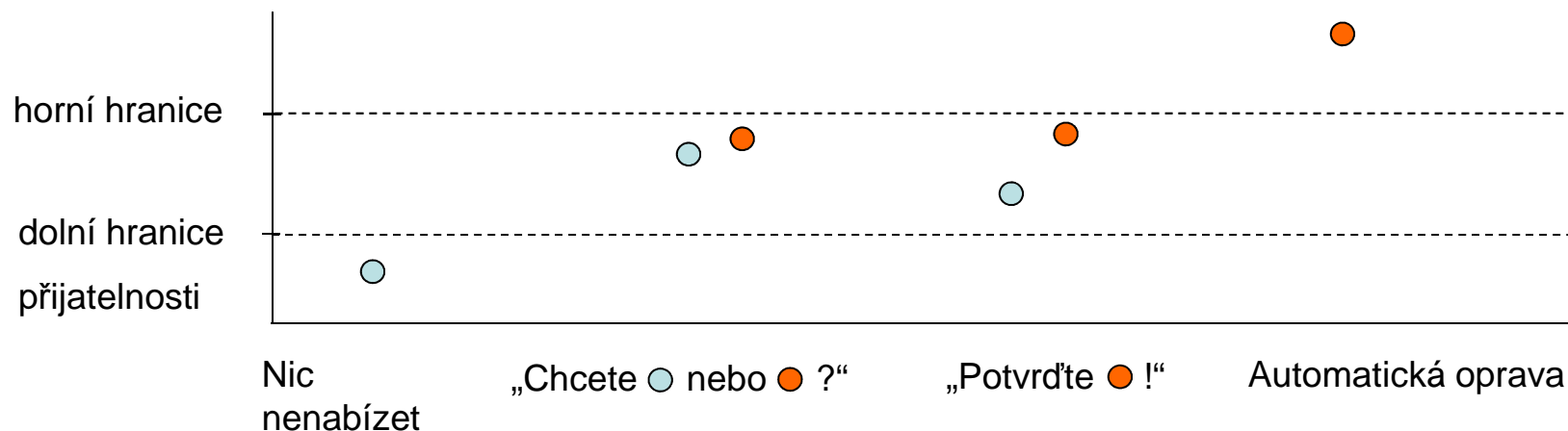
Nevýhoda: velmi neúplné, i řada chybných slov se skládá z vhodných kombinací znaků

Kontrola překlepů

Možná vylepšení

- vzít v úvahu okolnosti vzniku chyb (blízké klávesy apod.)
- zohlednit statistiku chyb
- zohlednit možné pravopisné chyby (mně x mě, jsem x jsme)
- heuristika na oddělení chyb a neznámých slov
- zapojení syntaxe a sémantiky
- pracovat s kontextem (porovnávat s korpusy apod.)

Komunikace s uživatelem



System ASIMUT

Automatická Selekcce Informací Metodou Úplného Textu (Králíková, Panevová '90)

2 základní moduly – jazykový a vyhledávací

Vyhledávací modul

výrazy (složené z podstatných a přídavných jmen) v základním tvaru, doplněné a sadu operátorů:

! vyskloňovat slovo

-1- obě slova musejí být bezprostředně vedle sebe

-2- mezera mezi oběma slovy nesmí obsahovat více než dvě slova

-3- obě slova se musejí nacházet ve stejné větě

-4- ... ve stejném odstavci

Příklad:

Dotaz : *vzdálenost!*, *odstup!* -3- *rodinný!* -1- *domek!*

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

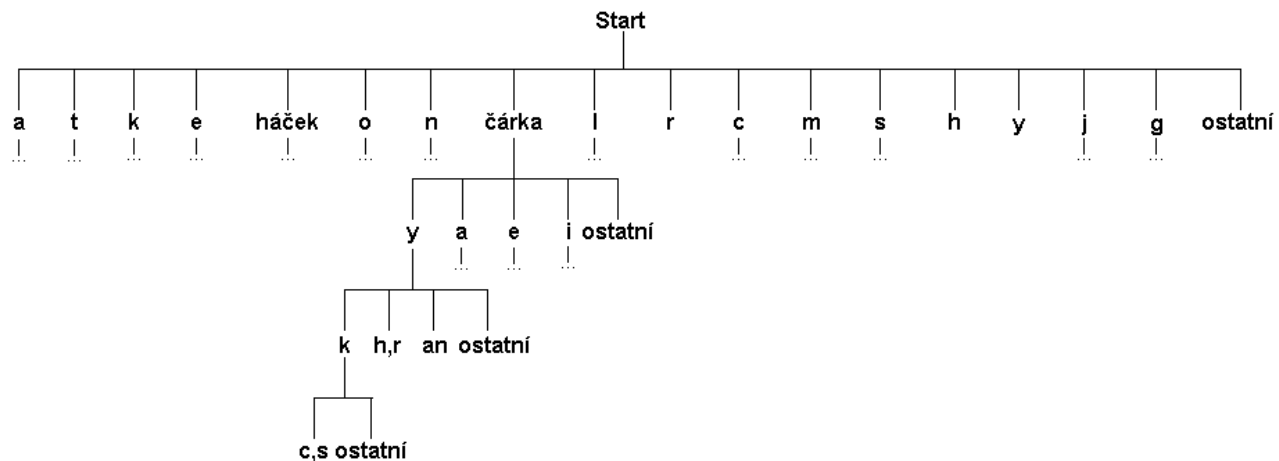
System ASIMUT

Jazykový modul

Neobsahuje žádný rozsáhlý slovník. Je založen na retrográdním slovníku dr.Slavíčkové (1975) – mnoho slov, která mají v základním tvaru stejný koncový segment, se stejně skloňuje. Výjimky je možné uložit do zvláštního slovníku výjimek, jsou jich řádově pouze stovky (při důkladnějším zpracování max. tisíce).

Algoritmus:

Porovnávají se jednotlivé znaky zákl.tvaru slova odzadu (háček a čárka jsou zvláštní znaky) dokud není možné (až na výjimky) jednoznačně určit, jak slovo skloňovat. Poté slovnímu základu (event. základům v případě změn v kmeni) přidáme všechny vhodné pádové koncovky.



System ASIMUT

Problémy

- Není vždy možné jednoznačně určit vzor
- Příliš hrubá klasifikace, pádové koncovky mají varianty => přegenerování
- Malý rozsah Retrográdního slovníku => je nutné přidávat výjimky
- Nefunguje tak spolehlivě pro slovesa (příliš velká víceznačnost koncových segmentů základních slovesných tvarů)

Další pojmy

Negativní slovník

Obsahuje slova, která nejsou důležitá při dotazování (spojky, citoslovce apod.) – tato slova jsou odstraněna při předzpracování textu

Konkordance

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.

System MOZAIKA

MOSAIC - Morphemic Oriented System of Automatic Indexing and Condensation

Standardní přístup k indexaci

- slovníky klíčových slov, dokumenty indexovány těmito slovy, v úvahu se bere četnost výskytu

MOZAIKA

- využívá toho, že řada přípon a koncovek nese význam, např.:
v angličtině *-er* nebo *-or* je **konatel** děje, *-tion* **činnost**, *-ity* nebo *-ness* **vlastnosti**;
v češtině *-ič*, *-ač*, *-čka*, *-ér*, *-or*, *-dlo*, *-metr*, *-graf*, *-fon*, *-skop* jsou **nástroje** nebo **přístroje**, *-ace*, *-kce*, *-áž*, *-ní*, *-za* **procesy** nebo **činnosti**,
-ost, *-ita*, *-nce* **vlastnosti** a přídavná jména končící na *-aný*, *-ený* jsou **výsledky procesů** zatímco *-ací*, *-ecí* značí **účel**.

Pro pokrytí tématické oblasti elektrických obvodů stačilo 800 přípon, technickou terminologii by pokrylo cca 2000 přípon.

System MOZAIKA

Algoritmus

- Vstupem je nijak nepředzpracovaný text, u kterého je zachováno typografické členění.
- Lematizace a morfologická analýza poskytnou lemata a morfologické značky.
- Nalezená lemata jsou profiltrována a jsou odstraněna např. ta, jejichž kmen nemá vztah k dané tematické oblasti (negativní slovník – malý, řádově desítky slov) či ta, která jsou příliš krátká nebo obsahují nevhodné kombinace hlásek.
- Syntaktická analýza jmenných skupin pomocí jednoduché gramatiky v jazyce Systémů Q pomůže odhalit několikaslovné termíny (*zesilovač* obsah textu charakterizuje mnohem méně než termín *operační zesilovač TESLA KC 415*), u nich se započítávaly i termíny v nich obsažené.
- Přiřazení vah na základě výskytu termínů v různě důležitých místech textu (nadpisy, první a poslední odstavce, první a poslední věty v odstavcích apod.).
- Normalizace vah vzhledem k délce dokumentu umožňuje porovnávat relevanci různě dlouhých dokumentů.
- Výstupem je seznam deseti nejvýznamnějších termínů, seřazený podle četnosti výskytu

System MOZAIKA

Výhody

- není nutné vytvářet slovníky odborných termínů, pouze množiny relevantních přípon a koncovek, doplněné o určitou formu negativního slovníku či pravidel
- lokální syntaktická analýza umožňuje větší flexibilitu při hledání termínů

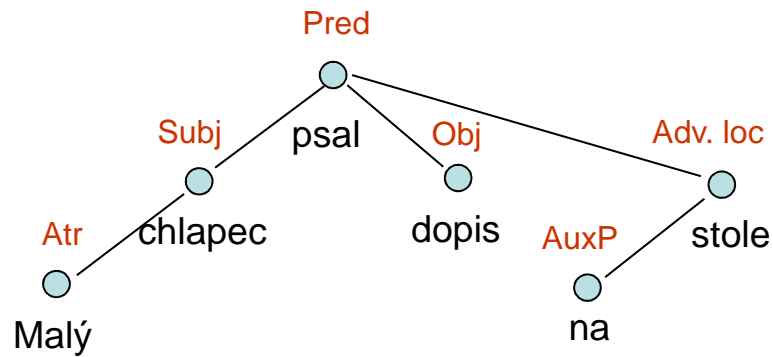
Problémy

- pracné vytváření slovníků a omezujících pravidel v závislosti na tematické oblasti
- neobsahuje řešení odkazů v textech pomocí zájmen, nevyjádřeného podmětu apod., např. „*Jedním z nejvýkonnějších u nás vyráběných přístrojů je operační zesilovač TESLA KC 415. Je vhodným pomocníkem Také je ...*“

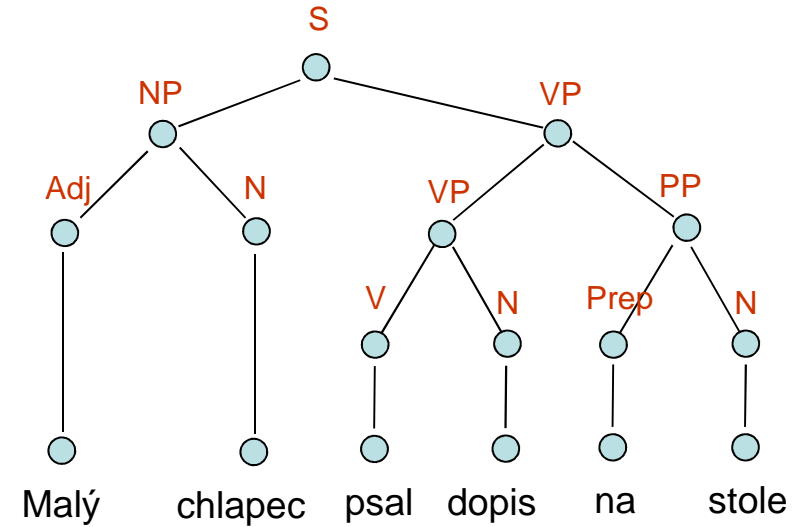
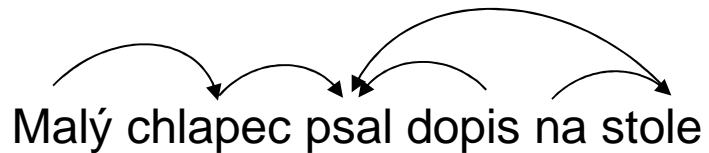
Syntax

Používané datové typy:

Malý chlapec psal dopis na stole.



Závislostní strom



Složkový strom

((Malý chlapec) ((psal dopis) (na stole)))

Syntax

Závislostní strom

- Velmi dobře a přehledně zachycuje vztahy mezi jednotlivými větnými členy.
- Nedává návod, jak strom získat (tj. strom nezachycuje postup výpočtu)
- Zdaleka ne všechny vztahy ve větě jsou přirozeně popsitelné jako závislost, zdaleka ne vždy je jasné, co na čem závisí (např. koordinace, předložky apod.)

Složkový strom

- Odpovídá derivačnímu stromu bezkontextové gramatiky.
- Je méně přehledný, obsahuje mnohdy velké množství nadbytečných uzlů.
- Přirozené jazyky nebývají bezkontextové

Neprojektivní konstrukce

„Nevinné věty“

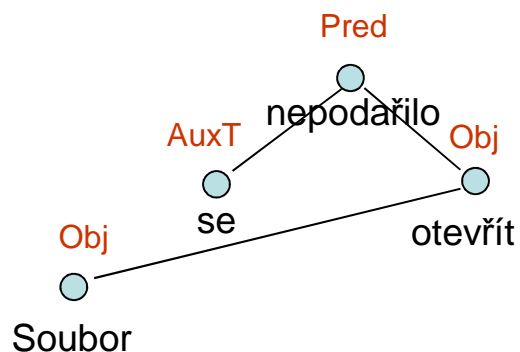
Soubor se nepodařilo otevřít.

Vánoční nadešel čas.

Které děvčata chtěla dostat ovoce?

Tuto knihu jsem se mu rozhodl dát k narozeninám.

Proti odvolání se zítra Petr v práci nakonec důrazně rozhodl protestovat.



?(Soubor ((se nepodařilo) otevřít))?.

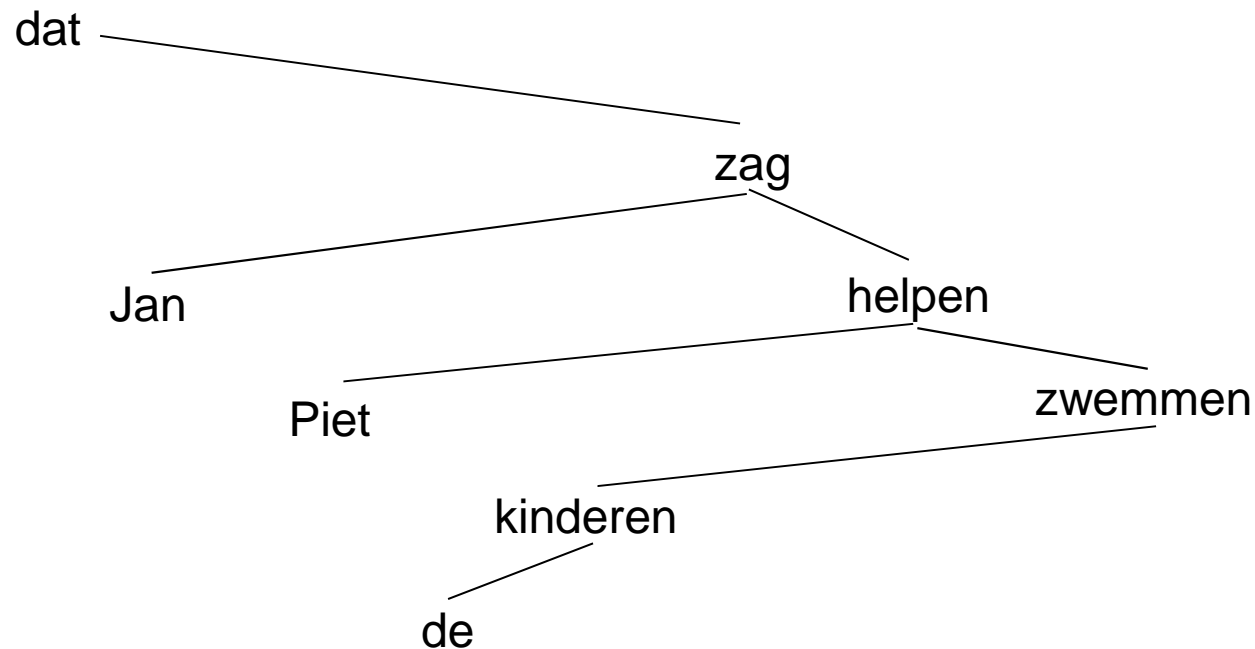
Neprojektivní konstrukce

... existují v mnoha jazycích.

dat Jan Piet de kinderen zag helpen zwemmen

Lit.: that Jan Piet the children saw help swim

that Jan saw Piet help the children swim



Transformační gramatika

Navazuje na předválečnou americkou lingvistiku, na snahu o explicitní popis jazykových pravidel. Předchůdci:

Deskriptivismus (Bloomfield 1933, dále Ch.Hockett a Z Harris) – jazyková fakta klasifikuje a registruje, ale nevysvětluje. Zpracovává zejména povrchovou větnou strukturu.

Analytická syntax (Jespersen 1937)

Logický přístup (Ajdukiewicz 1935) – kategoriální gramatika

Využívá (už tehdy existující) koncept **povrchové** (surface) a **hloubkové** (deep structure) syntaktické struktury. Jedné povrchové reprezentaci může odpovídat více hloubkových (významová víceznačnost) nebo naopak (více možností, jak vyjádřit stejný význam).

Transformační gramatika

Noam Chomsky: Syntactic Structures (1957)
Aspects of the Theory of Syntax (1965)

3 základní komponenty:

Báze: soubor bezkontextových pravidel, generující složkové stromy, tzv. frázové ukazatele (phrase markers)

Transformační komponent: transformační pravidla operující na celých frázových ukazatelích, dělí se na obligatorní a fakultativní

Fonologický komponent: obsahující regulární přepisovací pravidla přidělující řetězům morfémů fonetické interpretace

Transformační gramatika

Množina přijatelných vět daného jazyka je vytvářena **generativní procedurou**, souborem konečného počtu přepisovacích pravidel. Jde v podstatě o bezkontextovou nebo kontextovou gramatiku:

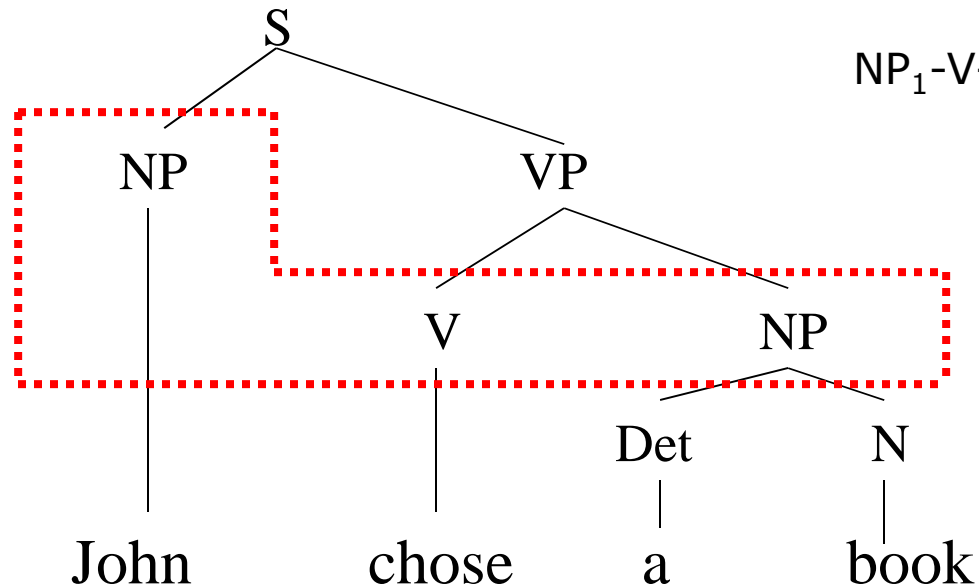
$$\text{VP} \rightarrow \left\{ \begin{array}{cc} V_{\text{intr}}^{\text{sg}} & \text{Adv} \\ V_{\text{tr}}^{\text{sg}} & \text{NP} \end{array} \right\} / \text{NP}^{\text{sg}} _$$

Generativní procedura není schopna zachytit vztahy mezi variantami vět, např. mezi větou tázací a oznamovací.

Transformační gramatika

Transformační složka obsahuje pravidla, která z původních podkladových frázových ukazatelů vytváří povrchovou strukturu věty.

Transformace jsou definovány **strukturním indexem** řetězců a **strukturní změnou**



$NP_1-V-NP_2 \Rightarrow NP_2-was-V+-en-by+NP_1$

Vývoj transformační gramatiky

1965 – **Standard Theory**, popsaná v knize Aspects of the Theory of Syntax.

1968 – **Extended Standard Theory**

zač. 80.let – **Government-binding theory (GB)** – teorie založená na obecných **principech** univerzální gramatiky a **parametrech** platných pro jednotlivé jazyky.

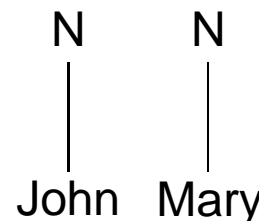
zač. 90.let – **Teorie minimalismu**, obsahuje pouze dvě roviny, rovinu logické formy (LF) a fonetickou rovinu (PF), sloužící jako rozhraní mezi zvukem (PF), reprezentací jazyka a významem (LF)

Tree Adjoining Grammars

pol. 70.let – Joshi, Levy, Takahashi

Substituce stromů

Elementární struktury jsou stromy

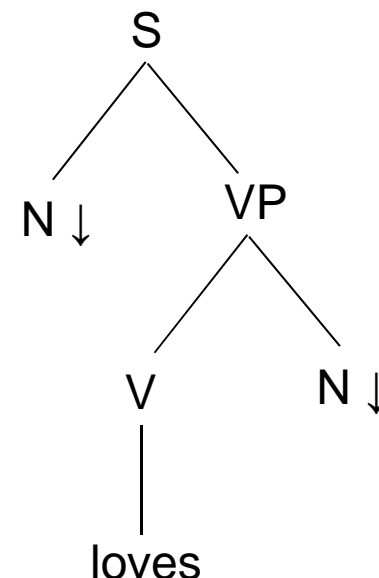


Šipka ↓ označuje, který uzel je možné substituovat

Formalismus se drží myšlenky přepisování

kompletních stromů, nepřepisuje ale řetězce, nýbrž **stromy**

Při substituci se musejí oba neterminály (kořen, list stromu) shodovat



Pořadí substitucí nehraje roli

Proces končí, pokud už žádný neterminál nelze nahradit

Generativní síla odpovídá bezkontextovým gramatikám, po modifikacích mohou být i silnější (kontextové).

Lexical Functional Grammar

Rozlišuje dva základní typy struktur:

- c-structures (constituent structure)
spojování slov do frází
- f-structure (functional structure)
reprezentuje funkční vztahy ve větě (např. vazby sloves)
jiný datový typ – matice atribut-hodnota

[PRED 'David']
	NUM	SG	

[PRED	'spát <SUBJ>']	
	TENSE	PAST			
	SUBJ	[
		PRED 'David'			
		NUM	SG		
]			

- hodnotami atributů mohou být i množiny
- každá c-struktura se spojuje pouze s jednou f-strukturou, ale nikoli naopak

Kategoriální gramatiky

Každému vstupnímu slovnímu tvaru je přiřazena **kategorie**, která fakticky reprezentuje popis syntaktických vlastností dané slovní formy.

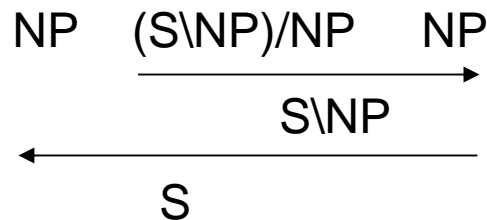
Například sloveso **likes** dostane kategorii (S\NP)/NP

Kategorie mají obecný formát α/β nebo $\alpha\backslash\beta$, kde lomítko určuje pozici argumentu β , tedy zda je vpravo (/) nebo vlevo (\) od α (používají se ale i jiné notace!)

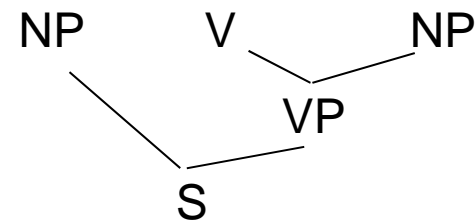
V „čisté“ kategoriální gramatice se používají pouhá dvě pravidla:

$$X/Y \ Y \Rightarrow X \quad \text{a} \quad Y \ X/Y \Rightarrow X$$

Příklad: Dexter likes Waren



Dexter likes Waren



Další typy gramatiky

Unifikační gramatiky

Functional Unification Grammar

Martin Kay

Generalized Phrase Structure Grammar (GPSG)

Gerald Gazdar, Geoffrey Pullum, Ivan Sag, Ewan Klein (1985)

Head Driven Phrase Structure Grammar (HPSG)

Pollard a Sag (1985)

Kategoriální gramatiky

Combinatorial Categorical Grammar

Mark Steedman

Unifikační gramatiky

Popis vlastností objektů

Objekt je reprezentován množinou vlastností (jednoduchých rysů). Popis každé vlastnosti je dvojice `<nazev_vlastnosti>` : `<hodnota_vlastnosti>`

Popis objektu je tvořen neuspořádanou množinou vlastností, tzv. `sestavou rysů` (`feature structure`)

Příklad: Slovo `books`

$$\left[\begin{array}{l} \text{graphematic_form : books} \\ \text{POS : noun} \\ \text{gender : neutral} \\ \text{number : plural} \end{array} \right]$$

Unifikace

Spojování dvou sestav rysů úpopisujících stejný objekt

$$\left[\begin{array}{l} \text{POS : verb} \\ \text{person : third} \\ \text{number : plural} \end{array} \right] \cup \left[\begin{array}{l} \text{gender : masc animate} \\ \text{number : plural} \end{array} \right] = \left[\begin{array}{l} \text{POS : verb} \\ \text{person : third} \\ \text{gender : masc animate} \\ \text{number : plural} \end{array} \right]$$

Unifikace je povolena pouze tehdy, pokud hodnoty všech rysů z určité sestavy neodporují nějaké hodnotě stejného rysu z jiné sestavy.

Pokud dvě sestavy rysů obsahují rozporné informace, potom je výsledkem unifikace vnitřně rozporná sestava rysů obvykle označovaná jako \perp

Sestavy rysů

Základní datová struktura unifikačních gramatik.

Obsahují kombinaci rysů, která popisuje určitý jev (např. shodu apod.)

Hodnotou vlastnosti (rysu) může být také sestava rysů nebo proměnná.

$$\left[\begin{array}{l} \text{subject} : \left[\begin{array}{l} \text{person} : 2 \\ \text{gender} : \text{fem} \end{array} \right] \\ \text{predicate} : \left[\begin{array}{l} \text{person} : 2 \\ \text{gender} : \text{fem} \end{array} \right] \end{array} \right] \quad \left[\begin{array}{l} \text{subject} : |1| \left[\begin{array}{l} \text{person} : 2 \\ \text{gender} : \text{fem} \end{array} \right] \\ \text{predicate} : |1| \end{array} \right]$$

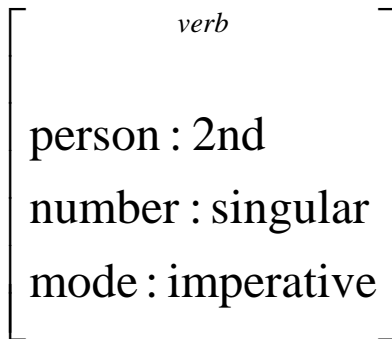
Problém:

Je možné unifikovat vlastnosti, které spolu nijak nesouvisejí.

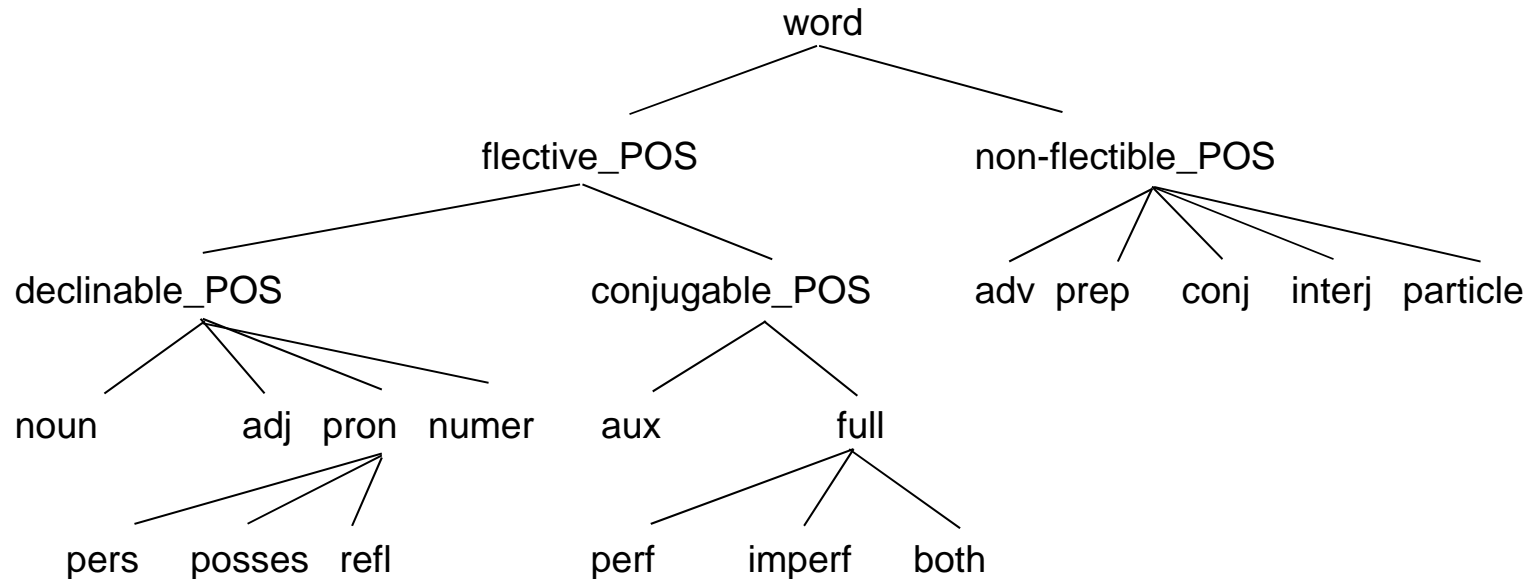
$$[\text{case} : \text{acc}] \cup [\text{mode} : \text{ind}] = \left[\begin{array}{l} \text{case} : \text{acc} \\ \text{mode} : \text{ind} \end{array} \right]$$

Typované sestavy rysů

Typ sestavy určuje její vlastnosti.



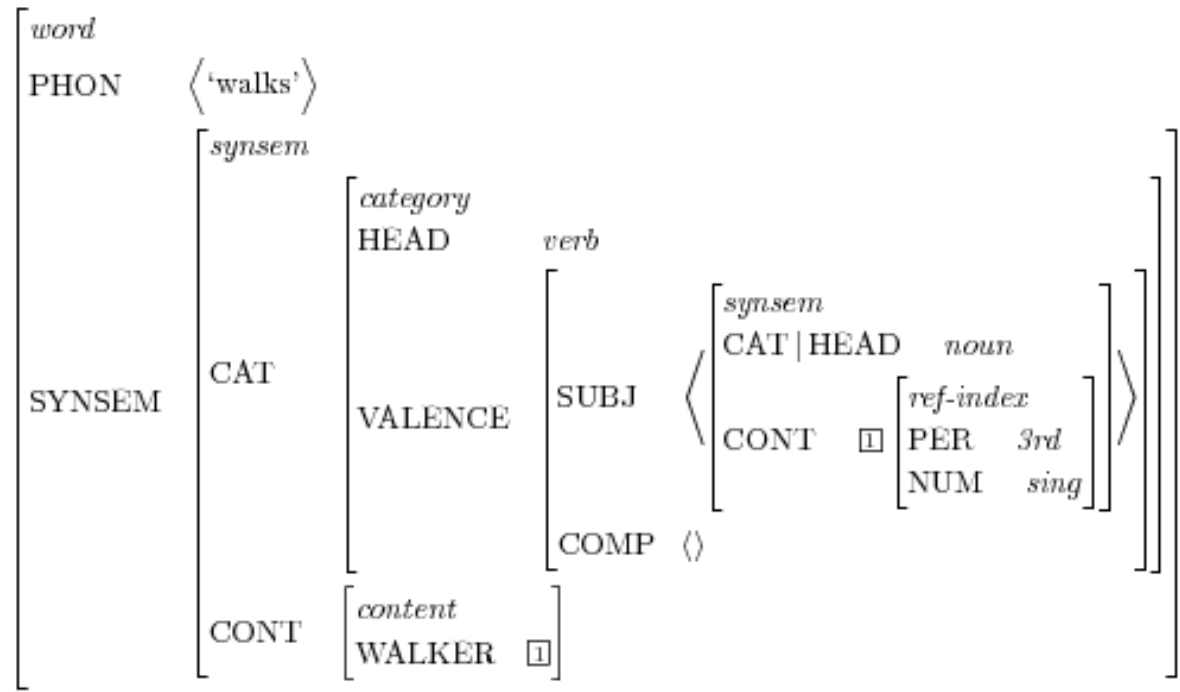
Typy jsou obvykle organizovány hierarchicky.



HPSG

HPSG zahrnuje principy, gramatická pravidla a slovníkové položky. Slovník je bohatě strukturován, položky nesou řadu informací. Základním typem je v HPSG **znak(sign)**. Slova a fráze jsou dva různé podtypy znaku.

Slovo má dva rysy: *[PHON]* (zvuk, fonetickou formu) a *[SYNSEM]* (syntaktické a sémantické informace), oba jsou dále děleny.



Nástroje pro syntaktickou analýzu

Augmented Transition Networks (Woods, 1970)

typy hran: CAT – přechod do dalšího stavu, nalezne-li příslušnou kategorii

JUMP – přechod do dalšího stavu bez hledání kategorií

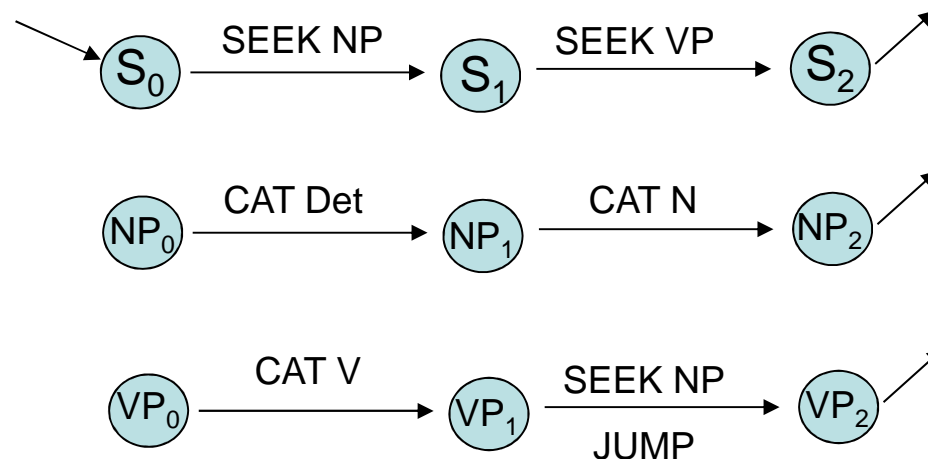
SEEK – přechod k podsíti

S -> NP VP

NP -> Det NP

VP -> V [NP]

The girl saw a boy.



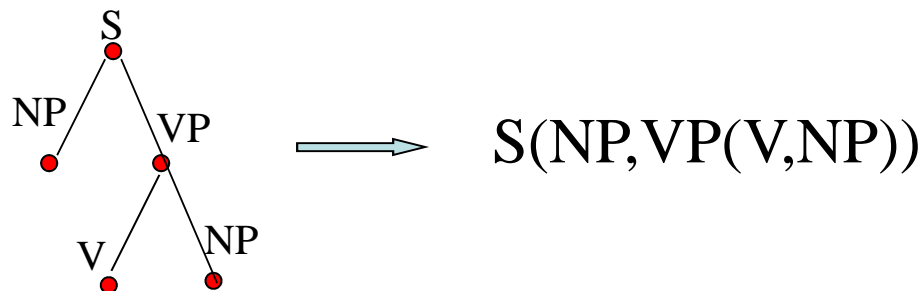
Nástroje pro syntaktickou analýzu

Q-systémy

Alain Colmerauer 1969

Formalismus pro transformaci grafů

Grafy (stromy) jsou linearizovány



Q-Systémy

Základní vlastnosti

Grafový analyzátor (chart parser)

Tři typy objektů: atomy, stromy a seznamy stromů

Implicitní typy proměnných:

- atom (konstanta): písmena z počátku abecedy A-J
- strom: střed abecedy, písmena L-N
- seznam: konec abecedy, písmena U-Z

Operátory -DANS- -HORS- -ET- -NON- -OU- = "

Příklad:

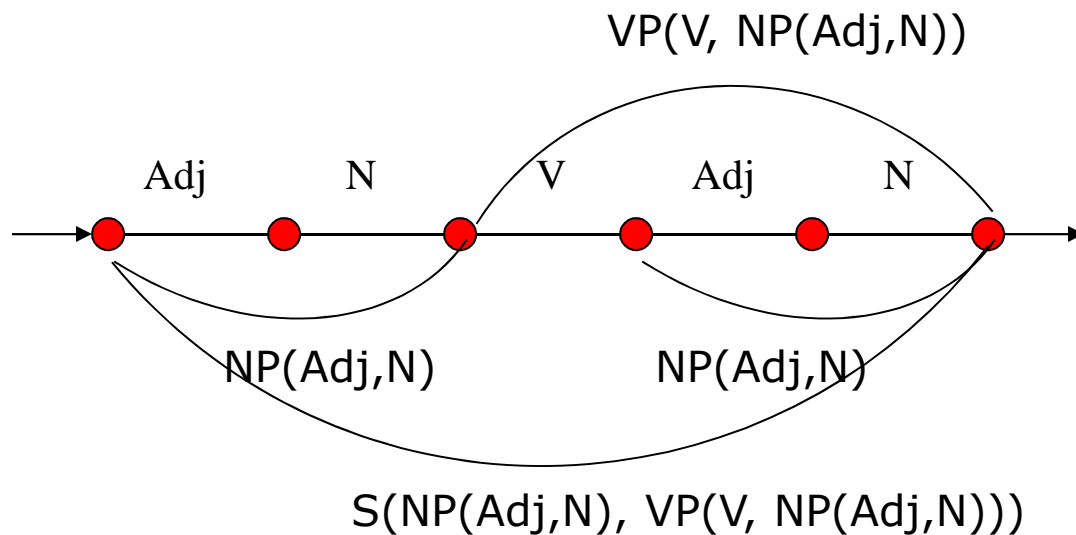
$S(NP, VP(V, NP))$ může být popsáno jako:

$A^*(U^*)$ nebo $S(NP, L^*)$ či M^*

(* signalizuje, že se jedná o proměnnou)

Q-Systemy

Vstupní graf



Pravidla:

$Adj + N \Rightarrow NP(Adj, N)$

$V + NP(U^*) \Rightarrow VP(V, NP(U^*))$

$NP(U^*) + VP(V^*) \Rightarrow S(NP(U^*), VP(V^*))$

Q-Systemy

Reálná česká gramatika (Systém RUSLAN)

** PRIPOJOVANI ADJEKTIV K SUBSTANTIVUM .

** ADJEKTIVUM ZLEVA .

2(U*1,B*(Y*1),U*2,A*,@(W*),#(\$),U*3,1(X*1,\$(\$)))

+ 1(B*,Z*1,/,Y*2,/,Z*2,@(V*),Z*3,\$(Z*4),Z*5)

== 1(B*,/,Y*2,/,@(V*),Z*3,\$(Z*4),

2(A*,U*3,1(X*1,\$(\$))),Z*5)

/ (. Y*2 -DANS- Y*1

-OU- 4(\$2P) -DANS- Z*5 -ET- G(P) -DANS- Y*1 .)

-ET- Z*4 " C1

-ET- RS(+ (SP(V*)), -(V*)) -HORS- U*3

-ET- AD(+ (SP(W*)), -(W*)) -HORS- Z*3

-ET- -NON- (. RS(+) -DANS- U*3 -ET- RS(+ (V*)) -HORS- U*3

-OU- AD(+) -DANS- Z*3 -ET- AD(+ (W*)) -HORS- Z*3.).

Funkční generativní popis

Zakladatel Petr Sgall (od r.1967), později dále
E.Hajičová, J.Panevová

Navazuje na Pražskou lingvistickou školu

Kniha: The Meaning of the Sentence in its Pragmatic Aspects,
1986

Základní vlastnosti

- Stratifikační teorie (5 rovin – fonetická, fonologická, morfématická, povrchová a tektogramatická)
- Formy a funkce – jednotka na vyšší rovině reprezentuje funkci jednotky na rovině nižší (TG rovina je nejvyšší)
- Závislostní reprezentace
- Teorie valence (vazby, vyžadované nebo povolené řídicími slovy, zejména slovesy)

Valence

Teorie valence existuje od 60.let Její základy vytvořili J.Kuryłowicz (1949) a L. Tesnière (1959), rozpracoval ji Charles Fillmore (1968, 1977) ve své „Case Grammar,“ ve které studoval sémantické role jednotlivých slovesných aktantů.

V rámci FGP:

slovesa Panevová (1974-1975), Hajičová (1979), Panevová (1980) a (1994)

substantiva Novotný (1980), Panevová (2000)

adjektiva Piřha (1982), Panevová (1998).

Valenční slovník:

Valex – Lopatková, Žabokrtský 2007

Valence

Dva základní druhy závislých členů na TG rovině:

- Aktanty – Konatel (aktor, agens), Patient, Adresát, Origo, Efekt
každý z nich může být ve větě zastoupen pouze jednou
(i když je samozřejmě lze koordinovat)
- Volná doplnění – mohou se vyskytovat vícekrát

Další dělení

Obligatorní a fakultativní (na TG rovině!)

Obligatorní aktant nesmí ve větě chybět (může ovšem chybět na povrchové rovině, pokud ho známe např. z kontextu)

Dialogový test

Moji přátelé přijeli.

Kam? Odkud?

*Nevím Nevím.

Moji přátelé odjeli.

Odkud? Proč?

*Nevím Nevím.

Valenční rámec – seznam aktantů (i fakultativních) a obligatorních volných doplnění

Vallex

VALLEX 2.5

[alphabet](#)[class](#)[functors](#)[forms](#)[aspect](#)[control](#)[reflex.](#)[recipr.](#)[complexity](#)[miscel.](#)[home](#)[help](#)

darovat

- A (14)
- B (32)
- C (11)
- Č (10)
- D (129)
- E (8)
- F (10)
- G (1)
- H (51)
- CH (22)
- I (17)
- J (13)
- K (73)
- L (37)
- M (53)
- N (133)
- O (220)
- P (528)
- R (104)
- Ř (12)
- S (225)
- Š (13)
- T (61)
- U (173)
- V (376)
- Z (392)
- Ž (11)

- darovat
- dařit se, dařivat se
- dávat, dát
- dávat se, dát se
- dávat si, dát si
- datovat
- datovat se
- dbát
- definovat
- deklarovat
- děkovat
- děkovat se
- dělat, dělávat
- dělat se, dělávat se
- dělat si, dělávat si
- dělit
- dělit se
- demonstrovat
- děsit, děsivat
- děsit se, děsivat se
- diktovat, diktovávat
- dirigovat
- diskutovat
- disponovat
- distancovat
- distancovat se
- distribuovat
- dít_I
- dít_{II} se
- dívat se

darovat^{pf}

1 ≈ dát darem; věnovat

-frame: **ACT**₁^{obl} **ADDR**₃^{obl} **PAT**₄^{obl} **CAUS**_{k+3}^{typ} **AIM**_{na+4}^{typ} **RCMP**_{za+4}^{typ}

-example: daroval dceři k narozeninám knihu; daroval dceři za dobré vysvědčení knihu; zisk darujeme na opravu dětské nemocnice
cor4: chtěla bych se ti darovat

-rfl: cor3: k příštím narozeninám si daruje obrovský dort (ČNK)
pass: ženám se obvykle darují květiny

-rcp: ACT-ADDR: vzájemně si toho mnoho darovali

-class: exchange

2 ≈ odpustit; prominout (idiom)

-frame: **ACT**₁^{obl} **ADDR**₃^{obl} **DPHR**_{to}^{obl}

-example: To ti nedaruju!

-rfl: pass: To se mu příště nedaruje!

-rcp: ACT-ADDR: nic si nedarují

3 ≈ dát milost (idiom)

-frame: **ACT**₁^{obl} **PAT**₃^{obl} **DPHR**_{život}^{obl}

-example: darovat odsouzenému život

cor3: kdybych si tak mohl darovat život

-rfl: pass: podepisovali petici, aby se odsouzenému daroval život

Kontrola gramatické správnosti

Co je vlastně možné kontrolovat?

Př.: Sportovci věnovaly plyšáka. Tatínek šli do práce.

Problémy specifické pro češtinu:

- Gramatická shoda, zejména shoda podmětu s přísudkem
- Interpunkce
- Neprojektivní konstrukce
- Zájmena (mě/mně)

Jak kontrolovat?

Chybové vzorky – vhodnější pro jazyky s pevným slovosledem, kde se chybné konstrukce spíš vyskytují v lokálním kontextu

Gramatika – nelze rozeznat, kdy je konstrukce chybná pouze vzhledem k (neúplné) gramatice a kdy je opravdu špatně

Kontrola gramatické správnosti

RFODG (Robust Free-Order Dependency Grammar)

- jedno pravidlo gramatiky může popisovat správnou i chybnou konstrukci zároveň
 - výpočet probíhá ve fázích, interpret gramatiky rozhoduje, jak se bude stejné gramatické pravidlo používat
 - 3 fáze:
 - pozitivní projektivní
 - negativní projektivní nebo pozitivní neprojektivní
 - negativní neprojektivní
- Novější implementace (Holan 2001) umožňuje ještě plynulejší fázování výpočtu.

RFODG

Ukázka reálného pravidla gramatiky:

```
;-----  
; 6. pravidlo - vytvoreni predlozkove fraze  
;  
PROJEKT TRUE  
  A.syntcl = prep  
  B.syntcl = noun  
  
  A.case ? B.case Case_Dis_Prep_Noun  
  
; IF B.wcl = prn THEN  
;                               B.pprep ? yes NonprepFormOfPronoun  
;                               ELSE ENDIF  
;  
  X:=B  
  X.syntcl := prepfr  
  
  X.prep := A.lexf  
; Kvuli operacim s konkretni predlozkou je treba vedet, o jakou predlozku jde  
OK  
END_P
```

LanGR

autor: P. Květoň 2003

Vlastnosti:

- primárně vyvíjen pro desambiguaci české morfologie
- pracuje s pozitivními a negativními desambiguačními pravidly
- pravidla mohou mít neomezený kontext
- redukční metoda – snaha udržet 100% přesnost
- pravidla jsou psána ručně, avšak na základě dat z korpusu
- pravidla jsou vzájemně nezávislá, neuspořádaná a jsou uplatňována v cyklech
- 4 části: kontext, desambiguační část, report a akce

cont₁ disamb₁ cont₂ disamb₂ ... cont_n disamb_n cont_{n+1} report action

LanGR

Ukážka reálného pravidla gramatiky:

```
/* Neither verb, nor preposition, nor conjunction can immediately follow the (in)definite
article */
rule ArtVerbPrepConj2 {
possart = ITEM Possible Article;
/* this is a disambiguation area: at least one of the interpretations of the word form possart
must be interpretable as an article */
safeverbpreconj = ITEM IsSafe Verb or Preposition or Conjunction;
/* a simple context specifying one corpus element as a verb or preposition or conjunction
only */

REPORT (The article possart cannot immediately precede the form safeverbpreconj!);

/* disambiguation actions: article interpretation (tag) in possart is discarded */
DELETE Article FROM possart;
/* or */
LEAVE ONLY not Article IN possart;

}; // end of rule ArtVerbPrepConj2
```

The rule can be successfully applied e.g. to the following sentence:

(2) *The letter a*(Article | Noun) *from*(Preposition) *the given alphabet is represented in blue.*