ÚVOD

- **Počítačová lingvistika**
  - Obor, zabývající se formálním popisem vlastností přirozených jazyků a jejich automatickým zpracováním (vytváření automatických systému, modelujících užívání přirozeného jazyka).
  - Využívá výsledky a příspívá k dalšímu rozvoji:
    - teoretické lingvistiky
    - teoretické informatiky
    - umělé inteligence
    - psychologie
    - logiky
    - matematiky (statistiky)
  - Podobory:
    - Rozpoznávání a generování mluvené řeci
    - 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 pojmu
  - 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
  - všeobecnost
  - využitelnost
  - obsah
  - vágnost
  - vícevrstevnost
  - zkratkovitost

MORFOLOGIE

- asi nejstarší odvětví lingvistiky
- 1. počátky - lingvista **Panini**, dílo **Ashtadhyayi** - pravidla morfologie sanskrtu
- Morfologie studuje vztahy mezi jednotlivými částmi slov, vnitřní struktury slov. Zabývá se tvořením tvarů slov a jejich významem, dále i tvořením nových slov.
- S morfologií se pojí pojmy:
    - **lexikologie** – slova jsou studována jako jednotky slovní zásoby, z hlediska funkce,
    - **lexikografie** – sestavování slovníků - která slova zahrnout, který význam dát první (příp. překlad)
- Základní jednotkou je **morfém** - nejmenší znaková jednotka (seskupení znaků) nesoucí význam
    - **lexikální morfém** - nese význam slova jako takového, př. kmen slova,
    - **gramatický morfém** – určuje gramatickou roli slovního tvaru.
    - Např. za-hrad-ou má předponu „za“, lexikální morfém „hrad“ a gramatický morfém „ou“, který určuje 3 elementární jednotky (sémata): pád, číslo a rod.
- Studuje způsoby **skloňování** (deklinace) a **časování** (konjugace).
- tvaroslovné **dublety** – stejné slovní tvary odvozené od více slovních základů (žena, tři, hnát, stát, atd.), neboli slova víceznačná, která mají různé slovní druhy,
- **alternace** – změna hlásek uvnitř kmene (vůz → vozu, švec → ševce, prkno → prken),
- **alomorfy** – varianty kmene odvozené od stejného slovního základu (nejvíce změn má matka – matce – matek – matčin).
- Slova se dělí na:
    - **autosémantická** = plnovýznamová,
    - **synsémantická** = pomocná.
- **Morfologická typologie jazyků** dělí jazyky podle toho, kolik morfémů je jedno slovo:
    - **analytické** (slovo = morfém)
        - izolační
        - každé slovo je morfém, bez předpon/přípon
        - vietnamština, čínština
    - **syntetické** (slovo > morfém)
        - **flektivní**
            - mají předpony, přípony, koncovky (ale míra řetězení je nějak omezená)
            - daný tvar morfému nese více významů (koncovka určuje pár, rod, …)
            - latina, stará řečtina, slovanské jazyky
        - **aglutinační**
            - také různé předpony apod., ale jeden morfém nese jeden význam (tedy např. koncovka přidá jeden význam)
            - maďarština, japonština, turečtina, finština
    - **polysyntetické** (slovo = věta)
        - eskymácké a indiánské jazyky.
- **Přístupy ke zpracování morfologie**
    - Morfologie založená na **morfémech**
        - vidí slovo jako řetízek morfémů
        - vhodné pro aglutinační jazyky
    - Morfologie založená na **lexémech**
        - vidí slovo jako výsledek aplikace pravidel, která slovo mění a tím vytváří nový slovní tvar
        - vhodné pro angličtinu
    - Morfologie založená na **slovech**
        - centrální roli mají vzory
        - známe-li základní tvar slova a jeho vzor, dokážeme vygenerovat všechny jeho tvary
        - vhodné i pokud jeden morfém reprezentuje více gramatických kategorií (např. 3.os,
        - sg., r.ž.) - tam předchozí přístupy selhávají
        - čeština - ve skutečnosti asi 250 vzorů (pokud má být jednoznačné)
- **Two-Level Morphology**

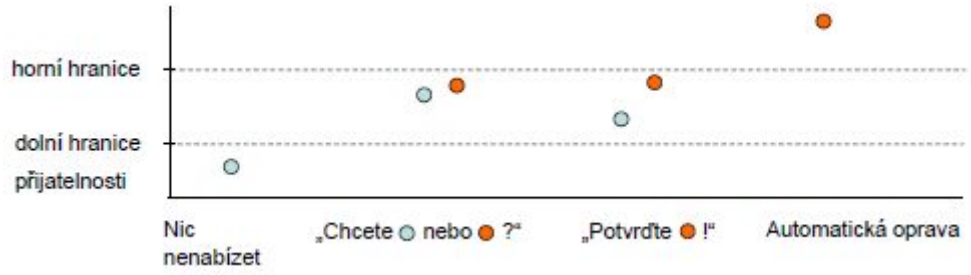
- Systém zpracování morfologie
- Lauri Karttunenem a Kimmo Koskeniemmin, zač. 80. let
- První obecný model zpracování morfologie přirozeného jazyka
- Morfologie jednodušší než gramatika, ale zpracování systému trvalo déle
- Založen na konečných stavových automatech a na nich definovaných oboustranných přechodech.
    - Analýza zpracovaná tímto automatem
    - Mechanismus morfologie byl pro všechny jazyky společný (to byl požadavek, aby to bylo nezávislé na jazyku), ale pro každý jazyk se musel vytvořit slovník a pravidla (přechody mezi stavy).
- 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
- 2 úrovně: první **lexikální**, druhá **povrchová**
- Základní myšlenky:
    - Pravidla se aplikují paralelně, nikoli sekvenčně
    - Podmínky se mohou vztahovat k oběma úrovním zároveň či jen k jedné z nich
    - Lexikální vyhledávání (prohledávání slovníku, trie) a morfologická analýza probíhají současně
- Nehodí se pro jazyky se vzory (čeština) - vzory nejsou pravidla → pro češtinu niky nebylo

- **Česká morfologie**
    - vyvíjena od r. 1989 zejména prof. Hajičem
    - využívá poziční značky, každá pozice má svůj jednoznačně určený význam
        - značky jsou 15 místné, ovšem rozeznává se pouze 13 kategorií (2 jsou rezervní)
        - každá kategorie má své pořadí ve výsledné značce
        - některé kategorie se vzájemně vylučují (např. příslovce nemá osobu) - pak se píše **"-"**
        - kromě značky se každému slovu ještě přiřadí jednoznačné lemma, což je základní tvar slova
            - lemma: jednoznačný identifikátor
        - značka a lemma dohromady jednoznačně určují slovo a jeho tvar se vším všudy
    - POS - part of speech (slovní druh), VOICE - způsob, VAR - např. nespisovné
    - Česká morfologická analýza
        - k některým tvarům daného slova pasuje více značek, pak se tam napíšou všechny
        - u jednoho slova bývá více značek (průměrně 4, nejhorší je to však u měkkých přídavných jmen jako „jarní“, která mají až 27-násobnou víceznačnost).
- **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)
            - to je náročné, jsou různé přístupy (algoritmická pravidla/statisticky/kombinace), nejlépe to umí čistě statistické metody → úspěšnost až 96%.
    - **Částečná morfologická desambiguace** založená na pravidlech
        - Oliva a Petkevič vytvořili pravidla, která platí bez výjimek
        - pomocí spolehlivých pravidel redukuje počet značek, odstraňuje nevhodné, ponechává všechny, které nelze spolehlivě odstranit
        - pokud se u některého slovního tvaru vyškrtá vše, znamená to, že ve větě musí být gramatická chyba
            - koupil MS - kontrola gramatiky ve Wordu
            - funguje rychle, poté se musí chyba najít
        - seznam pravidel, díky postavení ve větě a kontextu dokázali s téměř 100% přesností vyškrtnout několik značek

        - stále několik zůstalo, ale i tak lepší pro statistiku
    - \+ statistika → jen zanedbatelné zlepšení přesnosti asi o 0,2 %
        - → statistická metoda v podstatě stejně úspěšná
- **Lemmatizace**
    - proces výběru správného základního tvaru (lemmatu), ze kterého byl odvozen daný vstupní tvar
    - někdy nepotřebujeme značku, stačí základní tvar/y
    - klíčová operace pro vyhledávání v textech
    - u nás má úspěšnost 99,9 %.
- **Stemming**
    - odříznutí koncovky
    - na rozdíl od lemmatizace je základním tvarem kmen slova
    - populární je tzv. Porterův stemmer
        - tváří se univerzálně, ale pro češtinu nelze (alternace kmene)
    - pokud se něco projevuje stejně, chceme, aby to ukazovalo na stejný základ (je jedno, jestli lemma nebo kmen)
- **Generování**
    - proces výběru správného slovního tvaru, pokud známe lemma a příslušnou kombinaci gramatických kategorií

- **Kontrola překlepů jako aplikace morfologie**
    - vznikl pro to velký slovník - na konci 80. let, prof. Hajič
        - 800 000 slov
        - příliš velký pro tehdejší počítače a textové editory
            - → slova se stáhla ke společnému základu
            - snížil se počet lemmat (pravidla - něco vem, to k tomu připlácni… (vznikala tak i neexistující slova))
            - vešlo se na harddisk i s text. editorem a ještě zbylo místo
    - Požadavky: (v praxi velmi obtížné splnit, 4-6 je priorita)
        - najít a opravit všechny překlepy
        - přezkoušet kontextové podmínky korigované verze
        - neznámá slova se nemají hlásit jako chybná
            - jako chybu říct jen to, o čem jsme si 100% jistí
        - nedávat falešná chybová hlášení
        - maximálně automatická korektura
        - krátký čas zpracování
    - **2 základní metody**:
        - **Porovnávání řetězců se slovy ve slovníku**
            - Buď se slovníkem všech možných slovních tvarů daného jazyka (wordlist) nebo se slovníkem lemmat a provádíme morfologickou analýzu.
            - Výhoda: Je to spolehlivé a jednoduché
            - Nevýhoda: Pomalé, náročné na kvalitu slovníku, místo, nerozezná to chybná slova od neznámých a každé zlepšení musí zařídit autor či uživatel.
        - **Porovnávání skupin znaků** (dvojic, trojic) a hledání nedovolených kombinací znaků.
            - Výhoda: Je to nezávislé na slovníku a rychlé.
            - Nevýhoda: Velmi neúplné a neodhalí překlepy ve slovech, která se skládají jen z vhodných kombinací znaků.
            - Spíše jako doplněk než jako zdroj
    - **Možná vylepšení:**
        - vzít v úvahu okolnosti vzniku chyb (např. blízké klávesy)
        - zohlednit statistiku chyb
        - zohlednit možné pravopisné chyby (mně x mě, jsem x jsme)
        - různé heuristiky na oddělení chyb a neznámých slov
        - zapojení syntaxe a sémantiky
        - pracovat s kontextem (např. porovnávat s korpusy)

- **Komunikace s uživatelem**
  - velká vzdálenost nabízeného slova → nenabídnout možnosti, rovnou opravit
  - jeden kandidát výrazně silnější → nabídnout jen jeho



- **<u>Systém ASIMUT</u>**
  - Automatická Selekce Informací Metodou Úplného Textu
  - Vznikl 1990 Králíková, Panevová
  - Ještě v době, kdy neexistovaly slovníky
  - Metoda uměla vyskloňovat podst. jm a příd. jm., u sloves nefungovala tak dobře
  - Nekontrolovalo se, zda jsou přid. jm. a podst. jm v ustáleném spojení ve stejném pádu
  - Text předzpracovaný - vytáhly se příd. jm. a podst. jm. a k jejich seznamu se dodalo, kde se vyskytují (paragraf, věta, …)
  - **2 základní moduly**: jazykový a vyhledávací
  - **vyhledávací modul**
    - Sloužil pro automatické vyhledávání ohýbaných slov v textu na základě parametrů
    - Vstup: výrazy (složené z podst. a příd. jm.) v základním tvaru doplněné o sadu operátorů
      - (! vyskloňovat, -1- obě slova musí být bezprostředně za sebou, obecně -n-)
      - př. *vzdálenost!, rodinný! -1- domek!*, *vzdálenost!, odstup! -3- rodinný! -1- domek!*
    - neřeší, jestli vyhledává slova v gramaticky správném tvaru
    - předpokládá členění textu na slova, věty, odstavce
  - **jazykový modul**
    - Modul pro automatické skloňování českých slov. Pro dané vstupní slovo vrátí všechny jeho možné tvary.
    - Zajímavější
    - Využívá retrográdní slovník dr. Slavíčkové (1975)
      - žádný rozsáhlý slovník
      - slovník seřazený dle písmen slov odzadu (podle konce lemmat)
      - 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)
      - slovník nevyužívá přímo, ale na jeho základě byl vytvořen klíč pro určování vzorů slov (seznam pravidel dle konců slov)
    - **Algoritmus**:
      - Porovnávají se jednotlivé znaky zákl. tvaru slova odzadu (háček a čárka jsou zvláštní znaky), dokud není možné (až na výjimky) určit, jak slovo skloňovat. Poté slovnímu základu (event. základům v případě změn v kmeni) přidáme všechny vhodné pádové koncovky.
      - Umí i základní alternace
  - v té době bylo obtížné zachytit češtinu - spec. kódování háčků a čárek a kroužků
  - nebyl vyžadován velký slovník, jen slovník výjimek
  - množina koncovek zahrnuje životné i neživotné - vznikají patvary

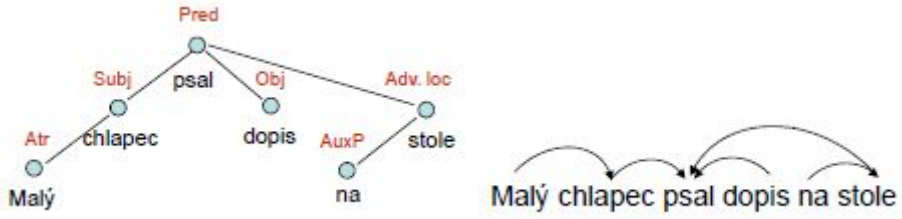
- po vybudování rozsáhlých topologických slovníků už se nepoužívá - slovníky fungují lépe
- **Problémy**
    - ne vždy lze jednoznačně určit vzor (právník i trávník mají stejnou koncovku, ale liší se v životnosti)
    - problém přegenerování (systém vygeneruje i neexistující tvary) - příliš hrubá klasifikace, pádové koncovky mají varianty
    - malý rozsah retrográdního slovníku (je tedy nutno přidávat výjimky)
    - pro slovesa už nefunguje tak spolehlivě (přílis velká víceznačnost koncových segmentů zákl. slovesných tvarů)
- Další pojmy:
    - **Negativní slovník**
        - Obsahuje ta slova, která nejsou při dotazování (vyhledávání v textu) důležitá (spojky, citoslovce), tato slova jsou odstraněna při předzpracování textu.
    - **Konkordance**
        - Všem slovům mimo negativní slovník byla přiřazena adresa a frekvence výskytu v textu, používaná pro účely urychlení hledání. Slova z negativního slovníku dostala jen adresu (kvůli počítání vzdáleností mezi jednotlivými významovými slovy v textu). Samotné vyhledávání pak neprobíhalo na původním textu, ale na této konkordanci.

- **<u>Systém MOZAIC</u>**
    - Morphemic Oriented System of Automatic Indexing and Condensation
    - 70's MFF Kirschner a kol.
    - Systém pro indexaci dokumentů, tvoření souhrnů, seznamů klíčových slov.
    - Vyvinut pro automatické indexování a kondenzaci textu
    - Podobně jako ASIMUT nepoužívá rozsáhlé slovníky klíčových slov, ale lingvistické poznatky.
    - Standardní přístup k indexaci - slovníky klíčových slov, dokumenty indexovány těmito slovy, v úvahu se bere četnost výskytu
    - Využívá toho, že řada přípon a koncovek nese význam (Aj: -er/-or konatel děje, -tion činnost, -ity/-ness vlastnosti; Čj: -ič/-ač/-čka/-ér/-or/-dlo/-metr/-graf/-fon/-skop přístroje a nástroje, -ace/-kce/-áž/-ní/-za procesy nebo činnosti, -ost/-ita/-nce vlastnosti, -aný/-ený jsou výsledky procesů, -ací/-ecí účel, atd.)
        - založeno na lexikální sémantice přípon a koncovek
    - Pro pokrytí tématické oblasti elektrických obvodů stačilo 800 přípon, technickou terminologii by pokrylo cca 2000 přípon.
    - Testován na elektrotechnických textech
    - Syntaktická analýza - Chceme dostat k sobě víceslovné výrazy, hledání víceslovných termínů
    - **Algoritmus** indexování textu:
        - Na vstupu je čistý (nijak nepředzpracovaný) text se zachovaných typografickým členěním.
        - Lematizace a morfologická analýza → získáme lemata a morfologické značky.
        - nalezená lemmata jsou profiltrována a jsou odstraněna ta, jejichž kmen nemá vztah k dané tematické oblasti (k tomu se využívá malý negativní slovník - řádově desítky slov), či jsou příliš krátká nebo obsahují nevhodné kombinace hlásek.
        - Syntaktická analýza jmenných skupin pomocí jednoduché gramatiky v jazyce Systémů Q pomůže odhalit tématicky významné několikaslovné termíny (*zesilovač* obsah textu charakterizuje mnohem méně než termín *operační zesilovač TESLA KC 415*), u nich se započítávaly i termíny v nich obsažené.

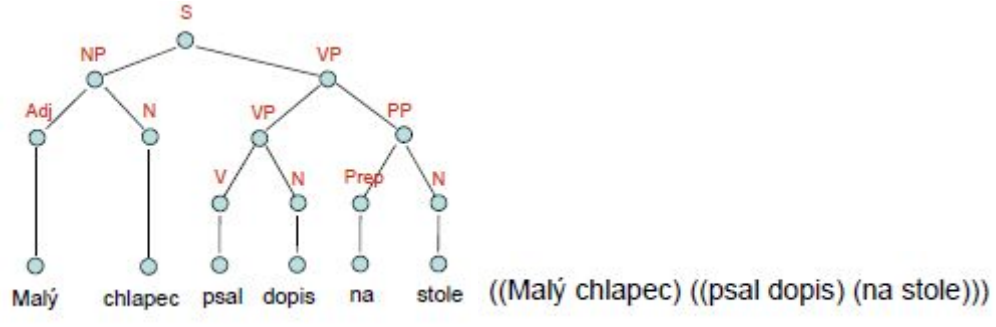
    - Vážené ohodnocení termínů podle důležitosti. Záleží na tom, v jak důležité části textu jsou (nadpis, první/poslední odstavec, první/poslední věta). Váhy jsou exponenciální.
    - Normalizace vah vzhledem k délce dokumentu. (Nejčastější termín získá 100 bodů, zbytek poměrně.) Umožňuje porovnávat relevanci různě dlouhých dokumentů.
    - Výstupem je 10 nejvýznamnějších termínů, seřazených podle četnosti výskytu.
- **Výhody:**
    - Není nutné vytvářet slovníky odborných termínů, pouze množiny relevantních koncovek a přípon, 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.

SYNTAX

- **Syntax** (skladba) se zabývá vztahy mezi slovy ve větě, tvořením větných konstrukcí, slovosledem.
- Základní datový typ: **strom**
  - pro zápis syntaxe věty
  - **závislostní strom**
  - **složkový strom**
- **Závislostní strom**

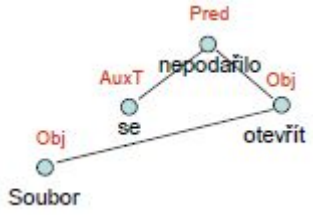


  -
  - velmi dobře a přehledně zachycuje vztahy mezi jednotlivými větnými členy
  - nedává návod, jak strom získat (tj. strom nezachycuje postup výpočtu)
  - zdaleka ne všechny vztahy ve větě jsou přirozeně popsatelné jako závislost, zdaleka ne vždy je jasné, co na čem závisí (např. koordinace, předložky apod.)
  - velmi se podobá větnému rozboru ze základní školy
  - kořen je jediný a obsahuje přísudek, uzly jsou právě slova věty, závislosti jsou orientované hrany
  - pamatuje si původní slovosled věty
  - přirozenější pro slovanské jazyky
  - šipkování od závislého k řídicímu
  - mají další možnosti pro zaznamenávání následujících jevů:
    - **Koordinace**
      - různé větné členy se stejnou sémantickou rolí
      - např. *Jan a Marie*; *černý nebo bílý*
    - **Apozice**
      - různé větné členy se stejnou syntaktickou rolí, shodnou gramatickou kategorií (tzn. gramaticky kongruentní)
      - např. *Matematicko-fyzikální fakulta (MFF)*; *Ivo Truchlivý, učitel matematiky*
    - **Parenze (vsuvka)**
      - věta či větný člen, který syntakticky nesouvisí s okolím, ale upřesňuje, o čem se v okolí mluví
      - *Mohl bych, prosím, zavřít okno?*
- **Složkový strom**



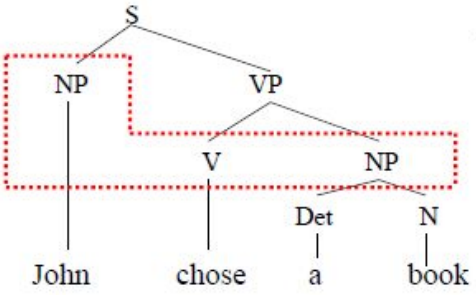
  -
  - odpovídá derivačnímu stromu bezkontextové gramatiky
    - tedy větu rozdělí do částí, které se zase rozdělí do částí, atd.
  - slova věty (tokeny) odpovídají listům stromu
  - je méně přehledný, obsahuje mnohdy velké množství nadbytečných uzlů
  - přirozené jazyky nebývají bezkontextové
    - neprojektivní konstrukce mu činí problém

- dá se znázornit pouhým uzávorkováním věty, kde uvnitř závorky jsou vždy právě dva prvky, kde prvek je buď jiný uzávorkovaný výraz, či samotné slovo
- vidíme terminály a neterminály
- hlavně anglo-americká tradice
- bezprostředně související slova jsou daleko od sebe (např. *chlapec* a *psal*)

- **Neprojektivní konstrukce**
  - závislost mezi dvěma slovy ve větě oddělenými slovem třetím, které (ani nepřímo) nezávisí na žádném z nich
  - např. *Soubor se nepodařilo otevřít. Vánoční nadešel čas. Která děvčata chtěla dostat ovoce? Tuto knihu jsem se mu rozhodl dát k narozeninám.*
  - Závislostní strom s tím nemá problém (jen hrany se v něm jakoby kříží). Složkový ano.
  - 

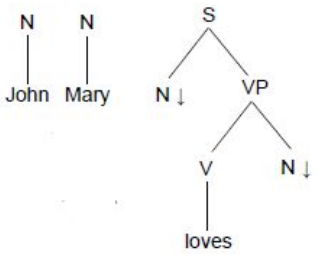
  - V češtině jsou běžné, ale jsou i v jiných jazycích.
    - v jedné české větě může být teoreticky nekonečně neprojektivit
      - prakticky od zhruba 4 není větě rozumět
    - v holandštině hodně neprojektivit, ale z hlediska konstrukce snazší než české

- **Transformační gramatika**
  - navazuje na předválečnou americkou lingvistiku, na snahu o explicitní popis jazykových pravidel
  - **předchůdci:**
    - **Deskriptivismus** (1993 Bloomfield)
      - Jazyková fakta klasifikuje a registruje, ale nevysvětluje
      - Zpracovává zejména povrchovou větnou strukturu
    - **Analytická syntax** (1937 Jespersen).
    - **Logický přístup** (1935 Ajdukiewicz) – kategoriální gramatika.
  - využívá (už tehdy existující) koncept **povrchové** (surface) a **hloubkové** (deep structure) **syntaktické struktury**
    - povrchová struktura řeší spíše zápis, hloubková význam
    - jedné povrchové reprezentaci může odpovídat více hloubkových (jedna věta je významově víceznačná) nebo naopak (více možností, jak vyjádřit stejný význam)
    - už zhruba ve 30. letech se začalo pracovat s těmito 2 koncepty
  - Jazyk však nebyl dosud popsán formální matematickou strukturou. Spíše se popisovalo, než že by se vysvětlovalo. Míchala se syntax a sémantika. Syntaktické jevy se popisovaly pomocí sémantiky apod.
  - **Noam Chomsky** 1957 Syntactic Structures (revoluce v popisu přirozeného jazyka)
    - popsal **3 základní komponenty**:
      - **Báze**
        - Soubor bezkontextových pravidel. Tato pravidla generují složkové stromy, tzv. frázové ukazatele (phrase makers). S → NP VP, kde NP je noun phrase, VP je verb phrase.
      - **Transformační komponenta**
        - Soubor transformačních pravidel operujících na celých frázových ukazatelích.
        - Z původních podkladových frázových ukazatelů vytváří povrchovou strukturu věty.
        - **2 typy** transformačních pravidel:
          - **Obligatorní** – transformace musí být provedena (pokud je to možné).

        - **Fakultativní** – transformace je volitelná.
    - **Fonologická komponenta**
        - Soubor regulárních přepisovacích pravidel. Řetězcům morfémů přidělují fonetickou interpretaci a význam. (Fonetika i fonologie zkoumají zvukovou stránku jazyka. Fonetika zkoumá, jak se hlásky v těle tvoří a vnímají, zatímco fonologie zkoumá funkci hlásek a zvukové rozdíly, které mají v jazyce nějakou funkci.)
- Množina přijatelných vět daného jazyka je vytvářena **generativní procedurou**, souborem konečného počtu přepisovacích pravidel
    - Jde v podstatě o bezkontextovou nebo kontextovou gramatiku:
    - $\mathbf{VP} \rightarrow \begin{Bmatrix} V^{sg}_{intr} & Adv \\ V^{sg}_{tr} & NP \end{Bmatrix} \; / \; NP^{sg} \_$
    - Generativní metoda není schopna zachytit vztahy mezi variantami vět, např. mezi větou tázací a oznamovací.
- **Transformace** (v transformační komponentě) jsou definovány **strukturním indexem** řetězců (řez stromem, výraz se matchuje na množinu vrcholů) a **strukturní změnou** (co se má s namatchovanými vrcholy provést).
    - 

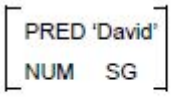
    - $NP_1\text{-}V\text{-}NP_2 \Rightarrow NP_2\text{-was-}V\text{+-en-by+}NP_1$
- Pravidla mohou být bezkontextová. Pak má tato složka sílu Turingova stroje, což je moc. V dalších verzích byla tato složka oslabena.
- **Vývoj transformační gramatiky**:
    - 1965 Aspects of the Theory of Syntax (N. Chomsky) - **Standard Theory**
    - 1968 **Extended Standard Theory**
    - 1980's **Government-binding Theory (GB)** – založená na obecných **principech** univerzální gramatiky a **parametrech** platných pro daný jazyk
    - 1990's **Teorie minimalismu**
        - obsahuje jen dvě roviny – opět Chomsky:
            - Rovina logické formy (LF) – reprezentace jazyka a významu
            - Fonetická rovina (PF) – zvuková stránka jazyka, rozhraní mezi zvukem

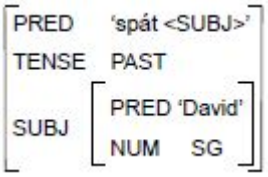
- **Tree Adjoining Grammars (TAG)**
    - pol. 70. let - Joshi, Levy, Takahashi
    - **Substituce stromů**
    - Elementární struktury jsou stromy
        - Pracuje se složkovými stromy
    - Šipka ↓ označuje, který uzel je možné substituovat



- Formalismus pro popis gramatik
- Drží se myšlenky přepisování kompletních **stromů** (ne řetězců)
    - např. X miluje Y=Emil, tak za X se dá dosadit „Milan“, ale i strom „Milan, Ferda a Dežo“)
- Při substituci se musí oba neterminály (kořen, list stromu) shodovat
- Pořadí substitucí nehraje roli
- Proces končí, když už nelze žádný neterminál nahradit
- Generativní síla odpovídá bezkontextovým gramatikám, po modifikacích mohou být i silnější (kontextové)
- Typy základních stromů:
    - základní (initial) strom: Udává valenční vztahy a strukturu věty
    - pomocný (auxiliary) strom: Pomocí těchto se tvoří rekurze ve stromu
- Typy změn:
    - Substituce – list stromu je nahrazen pomocným stromem, jehož kořen je značený stejně, jako list původního stromu.
    - Adjungace – vnitřní uzel je nahrazen pomocným strom, kořen opět značen stejně jako list původního stromu.

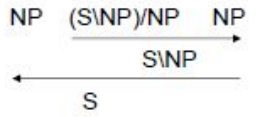
- **Lexical-Functional Grammar (LFG)**
    - sada dvojic atribut-hodnota
    - Rozlišuje **2 základní typy struktur**:
        - **c-strukture** (constituent structure)
            - Spojuje slova do frází
            - Datový typ je složkový strom.
        - **f-strukture** (functional structure)
            - Reprezentuje funkční vztahy ve větě (např. vazby sloves)
            - Používá datový typ matice atribut-hodnota

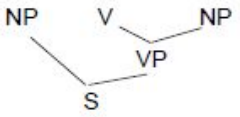




    - Hodnotami atributů mohou být i množiny
    - Každá c-struktura se spojuje pouze s jednou f-strukturou. Opačně jich může být i více.
- **Kategoriální gramatiky**
    - Každému vstupnímu slovnímu tvaru je přiřazena **kategorie**, která fakticky reprezentuje popis syntaktických vlastností dané slovní formy (je to vlastně množina syntaktických vlastností daného slova)
    - Např. 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 pouhá 2 pravidla:
        - X/Y Y → X
        - Y X\Y → X.

- Příklad: Dexter likes Waren





  - **Combinatorial Categorial Grammar**
    - Mark Steedman
- **Unifikační gramatiky**
  - navazují na f-structure z LFG
  - **Popis vlastností objektů**
    - Objekt je reprezentován množinou vlastností (jednoduchých **rysů**)
    - Popis každé vlastnosti je dvojice **<nazev_vlastnosti>:<hodnota_vlastnosti>**
    - Neuspořádaná množina vlastností = **sestava rysů** (feature structure)

$$\begin{bmatrix} \text{graphematic\_form} & : \text{books} \\ \text{POS} & : \text{noun} \\ \text{gender} & : \text{neutral} \\ \text{number} & : \text{plural} \end{bmatrix}$$

    - Vlastnosti mohou být např. grafémický zápis, slovní druh, rod, číslo, pád, atd.
  - **Unifikace**
    - Spojování dvou sestav rysů popisujících stejný objekt

$$\begin{bmatrix} \text{POS} : \text{verb} \\ \text{person} : \text{third} \\ \text{number} : \text{plural} \end{bmatrix} \cup \begin{bmatrix} \text{gender} : \text{masc animate} \\ \text{number} : \text{plural} \end{bmatrix} = \begin{bmatrix} \text{POS} : \text{verb} \\ \text{person} : \text{third} \\ \text{gender} : \text{masc animate} \\ \text{number} : \text{plural} \end{bmatrix}$$

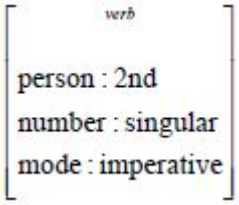
    - Povolena pouze tehdy, pokud hodnoty všech rysů z určité sestavy neodporují nějaké hodnotě stejného rysu z jiné sestavy
    - Pokud dvě sestavy rysů obsahují rozporné informace, potom je výsledkem unifikace vnitřně rozporná sestava rysů obvykle označovaná jako $\perp$
  - **Sestavy rysů**
    - Základní datová struktura unifikačních gramatik
    - Obsahují kombinaci rysů, která popisuje určitý jev (např. shodu apod.)
    - Hodnotou vlastnosti (rysu) může být také sestava rysů nebo proměnná.

$$\begin{bmatrix} subject : \begin{bmatrix} person : 2 \\ gender : fem \end{bmatrix} \\ predicate : \begin{bmatrix} person : 2 \\ gender : fem \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} subject : \boxed{1} \begin{bmatrix} person : 2 \\ gender : fem \end{bmatrix} \\ predicate : \boxed{1} \end{bmatrix}$$

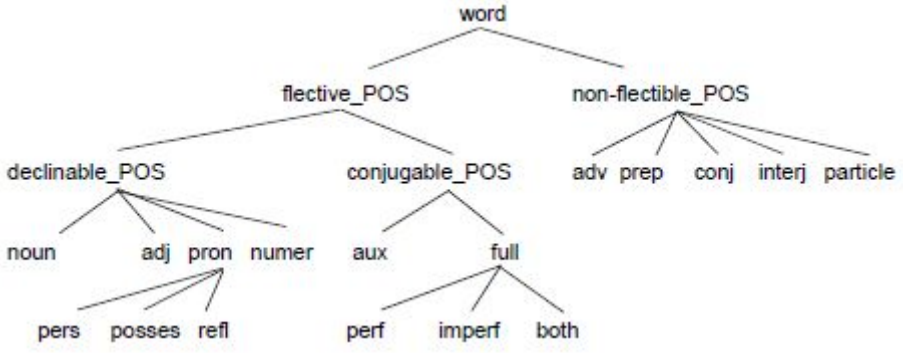
    - Problém:
      - Lze unifikovat i vlastnosti, které spolu nijak nesouvisejí (třeba pád podmětu a způsob přísudku)

$$\begin{bmatrix} \text{case} : \text{acc} \end{bmatrix} \cup \begin{bmatrix} \text{mode} : \text{ind} \end{bmatrix} = \begin{bmatrix} \text{case} : \text{acc} \\ \text{mode} : \text{ind} \end{bmatrix}$$

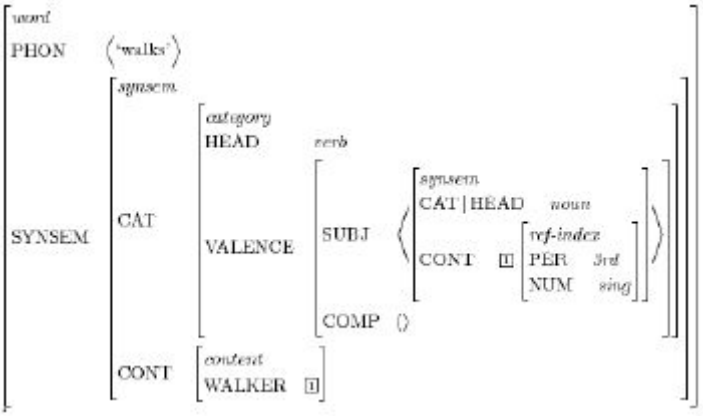
  - **Typované sestavy rysů**
    - Nakonec převládly, to jediné, co dávalo smysl
    - Typ sestavy určuje její vlastnosti



- Využívají toho, že některé typy objektů mají společné vlastnosti
  - slovesa - osoba, číslo, čas, způsob atd.
- Typy jsou obvykle organizovány hierarchicky
  - Slova se dělí na ohebné a neohebné druhy. Ohebné zase na časované a skloňované. Atd.

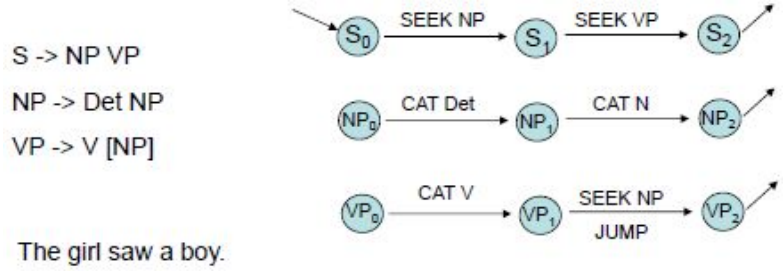


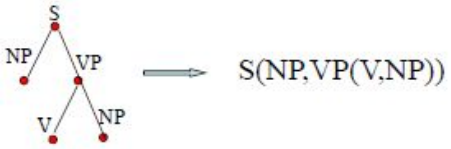
- **Head Driven Phrase Structure Grammar (HPSG)**
  - Pollard a Sag (1985)
  - Zahrnuje principy, gramatická pravidla a slovníkové položky (tříděné, dle různých kategorií).
  - Slovník je bohatě strukturován, položky nesou řadu informací.
  - Základním typem je **znak (sign)**.
  - Slova a fráze jsou dva různé podtypy znaku.
  - Slovo má **dva základní rysy**:
    - **[PHON]** (zvuk, fonetickou formu)
    - **[SYNSEM]** (syntaktické a sémantické informace)
    - tyto rysy jsou dále děleny.

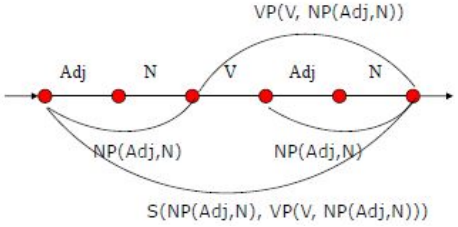


  - Umožňuje kompletně popsat každé jednotlivé slovo
  - Někdy dlouhé čekání - dříve velmi neefektivní
    - další implementace už trochu lepší
- **Functional Unification Grammar (FUG)**
  - Funkční unifikační gramatika
  - Martin Kay
- **Generalized Phrase Structure Grammar (GPSG)**
  - Gerald Gazdar, Geoffrey Pullum, Ivan Sag, Ewan Klein (1985)

- Dnes už je ale převálcovaly statistické metody.

- **Nástroje pro syntaktickou analýzu**
  - **Augmented Transition Networks** (Woods, 1970)
    - primitivní, ale úspěšnější než další systém
    - Rozšířené přechodové systémy.
    - tři typy hran:
      - CAT (přechod do stavu, nalezne-li příslušnou kategorii)
      - JUMP (přechod do stavu bez hledání kategorie)
      - SEEK (přechod k podsíti)
    - 

  - **Q-systémy** (Alain Colmerauer – otec prologu, 1969)
    - Formalismus pro tansformaci grafů
    - Grafy (stromy) jsou linearizovány
    - 

    - Grafový analyzátor (chart parser)
    - Linearizace ne řezem, jako u Chomského
    - Vada - cesta zpátky není jednoznačná - pořadí uzlů není zachováno
      - řešení: dodat pomocné uzly - např. # na pozici kořenového uzlu
    - 3 typy objektů (+ implicitní typy proměnných v původní verzi)
      - atomy (konstanty) - písmena z počátku abecedy A-J
      - stromy - střed abecedy, písmena L-N
      - seznamy stromů - konec abecedy, písmena U-Z
    - operátory -DANS- -HORS- -ET- -NON- -OU- =
    - př.
      S(NP,VP(V,NP)) může být popsáno jako:
      A*(U*) nebo S(NP,L*) či M*
      (* signalizuje, že se jedná o proměnnou)
    - 

    - vstupní graf:
    - pravidla:
      Adj + N ==> NP(Adj,N)
      V + NP(U*) ==> VP(V, NP(U*))
      NP(U*) + VP(V*) ==> S(NP(U*), VP(V*))
    - u pravidel záleží na pořadí
    - hrany na levé straně úspěšně použitého pravidla se zapamatují
      - na konci se smažou, byly překlenuty novou hranou

        - systém smaže také hrany, které na konci nejsou součástí cesty ze začátku do konce
    - jednotlivé malé gramatiky lze řetězit za sebe
        - každá může dělat jen část
        - v mezifázi se graf vždy vyčistí
    - nevýhoda - reálná pravidla dost nečitelná
    - nesmírně efektivní ve srovnání s ostatními v 80. letech, běželo rychle
    - dobrý mechanismus, ale neuchytil se, protože se o něm psalo jen francouzsky (muselo se)
    - Používal je např. RUSLAN, ale jinak byly oblíbené téměř jen ve francouzštině (jinde byly oblíbenější rozšířené přechodové systémy)
- **Funkční generativní popis** (Sgall, 1967)
    - později E. Hajičová, J. Panevová
    - projekt z MFF
    - teorie použitelná na jazyky jako čeština (volný slovosled, vysoká flexe)
    - navazuje na Pražskou lingvistickou školu
    - kniha: The Meaning of the Sentence in its Pragmatic Aspects, 1986
    - Stratifikační teorie – chápe popis jazyka jako popis 5 rovin (každá má nějakou funkci):
        - **fonetická**
        - **fonologická**
        - **morfématická**
        - **povrchová**
        - **tektogramatická** - sémantika slov
        - předpoklad, že roviny jsou nějak spojené
    - Formy a funkce - jednotka na vyšší rovině reprezentuje funkci jednotky na rovině nižší (TG je nejvyšší)
    - Na vyšších úrovních (povrchová a tektogramatická) se jazyk popisuje závislostní reprezentací, typicky závislostními stromy
    - **Teorie valence** (vazby, vyžadované nebo povolené řídicími slovy, zejména slovesy)
        - existuje již od 60. let
        - 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
        - schopnost některých slov (především sloves) „vyžadovat" jiné větné členy a tvořit s nimi věty
        - v rámci FGP:
            - **slovesa** - Panevová (1974-1975), Hajičová (1979), Panevová (1980) a (1994)
            - **substantiva** - Novotný (1980), Panevová (2000)
            - **adjektiva** - Piťha (1982), Panevová (1998).
        - **Vallex** - Lopatková, Žabokrtský 2007
        - 2 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** - obligatorní aktant nesmí ve větě chybět (může ovšem chybět na povrchové rovině, pokud ho známe např. z kontextu)
        - **fakultativní**
        - (na TG rovině)
        - dialogový test pro rozlišení obligatorních a fakultativních

Dialogový test

Moji přátelé přijeli.		Moji přátelé odjeli.	
Kam?	Odkud?	Odkud?	Proč?
*Nevím	Nevím.	*Nevím	Nevím.

    - **valenční rámec**
        - seznam aktantů (i fakultativních) a obligatorních volných doplnění
- **Kontrola gramatické správnosti**
    - problémy specifické pro češtinu:
        - shoda podmětu s přísudkem
        - interpunkce
        - neprojektivní konstrukce
        - zájmena (mě/mně)
    - Jak kontrolovat?
        - **Chybové vzorky**
            - vhodné hlavně pro jazyky s pevným slovosledem, kde se chybné konstrukce spíš vyskytují v lokálním kontextu (nerozlézají se daleko po větě)
        - **Gramatika**
            - nelze ale rozeznat, kdy je konstrukce chybná pouze vzhledem k (neúplné) gramatice a kdy je opravdu špatně
    - **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
                - umožnilo se něco porušit - povolila se jen 1 věc
            - negativní neprojektivní
                - povolí se vše - chyby i neprojektivity
                - velmi pomalé → metoda není velmi použitelná
        - snaha o co nejplynulejší fázování výpočtu
            - zlepšil Tom Holan, 2001 (disertační práce)
        - gramatika ručně psaná → nekompletní
            - to, co nebylo v gramatice, se hlásilo jako chyba
            - neúspěšný experiment
    - **LanGR**
        - P. Květoň 2003
        - 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
        - každé pravidlo má 4 části: kontext, desambiguační část, report a akce
        - pravidla tvořena speciálně pro češtinu (pro jiný jazyk by byla téměř úplně jiná)
        - používá desambiguaci na to, že když se odstraní všechny tagy, tak víme, že je něco špatně, v tu chvíli se ale musí určit, co je špatně a jak to opravit (a toto samozřejmě neopraví všechny chyby)

- Neřeší to ten problém, že věta může být správně, ale až v dalekém kontextu přes jiné věty. (Tatínek šly do práce.)
- Obecně používá tuto přípravu (klasický postup při zpracování psaného textu): segmentace (rozseká na věty) → tokenizace (rozseká na slova) → morfologická analýza (každému tokenu dá seznam dvojic lemma – tag) → morfologická desambiguace (každému tokenu vybere ideálně jeden token) → syntaktická analýza (větný rozbor) → sémantická analýza (rozbor významu věty).
- používá MS Word

STROJOVÝ PŘEKLAD

// under construction

KORPUSOVÁ LINGVISTIKA

- Zabývá se metodami, jak popisovat velké množství nasbíraných dat, studuje data v korpusech.
- **Korpus** je rozsáhlý soubor textů (v digitální podobě, el. databáze) v daném jazyce, většinou anotovaný (označkovaný) na základě přechozí morfologické a někdy i syntaktické analýzy. Je to cenný soubor dat, ale někdy se chybně považuje za reprezentativní vzorek či rovnou celý jazyk.
    - V současné době obsahují mnoho milionů slov běžného textu, jehož vlastnosti mohou být analyzovány pomocí značek (tagů) (doplněných informací identifikující a klasifikující slova nebo jiné útvary) a konkordačních programů.
    - Nejsou čisté - obsahují cizí slova, gramaticky nesprávné věty, …
- V korpusech musí být značky, abychom mohli pokládat dotazy typu: najdi 3 předložky za sebou
- **Charakteristika moderních korpusů**
    - Výběr vzorků a reprezentativnost
        - jazyk je nekonečný - korpus konečný...
        - reprezentativnost je důležitá, ale jde stranou
            - (knihy chráněné autorskými právy, nelze použít moc literatury)
    - Konečná velikost
        - s výjimkou tzv. monitorovacích korpusů (data jsou stále přidávána)
        - umožňuje kvantitativní výzkum
    - Strojem čitelná forma
        - snadné prohledávání, rychlá manipulace, snadno doplnitelné
    - Standardní reference
        - aby korpus mohl sloužit širšímu publiku, musí dodržovat určité standardy
- **Brown Corpus of Standard American English**
    - první moderní elektronický korpus
    - 1961 W.N.Francis a H.Kučera
    - 1 milion slov textů v amer. angličtině z roku 1961
    - 15 druhů textu (novinové reportáže, humor, krásná literatura, …), dohromady 500 textů, každý cca 2000 slov
    - Texty byly vybírány schválně náhodně. Celé to bylo pečlivé, ale milion slov není moc. A texty nebyly anotované.
- **Penn Treebank**
    - první a nejznámější syntakticky anotovaný korpus
    - 1990's Univerzita v Pensylvánii
    - cca 1 milion slov
    - 2 499 článků ze souboru článků Wall Street Journal (WSJ) v průběhu 3 let (burzovní angličtina)
        - dosti omezující
        - články různě dlouhé
    - autoři: Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz a Ann Taylor
    - syntaktická analýza využívala složkové systémy, tedy anotace pomocí uzávorkování a různých značek
    - snaha ho přeložit do češtiny (PCEDT), což se podařilo, ale s obtížemi
        - vyžadovalo to lidi, kteří by uměli dobře česky i anlglicky a vyznali se v prostředí burzovních textů z 90's. (Nejen překladatel, ale i ten, kdo to reviduje.)
        - motivace pro překlad včetně podobného značkování byla taková, aby se nějaký statistický program mohl učit rozdíly.
- další anglické: British National Corpus, American National Corpus, Corpus of Contemporary American English, Oxford English Corpus
- **Český národní korpus** (CNC)
    - Od 1994 společně UK, MU a Ústav pro jazyk český
    - Morfologicky označkovaný („otaggovaný“)
        - tedy na morfologické úrovni
        - ne ručně, ale automatickými nástroji pro morfologickou analýzu
    - Současně obsahuje asi 600 miliónů slov a je složený z převážné části z novinových článků, dále z literatury a odborných textů.

- část (100 milionů slov) uvolněno pro veřejnost - SYN2000
    - 15 % literatury, 60 % novinové texty, 25 % technické a odborné texty
- (čísla už dnes neplatí…)
- Morfologická analýza používá 15-ti poziční značky. V korpusu je rozeznáno 700 000 lemat, 15 miliónů slovních forem a po stochastické desambiguaci zůstane u každého slova průměrně 4,29 tagů. Používá statistické metody. Na učení se využívá ručně označkovaný korpus s 1,2 milióny tokenů (slov). K automatickému učení používá kontextová pravidla (asi 11 000 pravidel). Automaticky určuje váhy. Dosahuje rychlosti 200 tokenů za sekundu a výsledná úspěšnost je přes 94%.

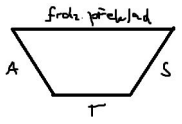
- **Pražský závislostní korpus** (Prague Dependency Treebank)
    - Propracované anotační schéma aplikovatelné na jazyky různých typů
    - Anotovaný automatickou metodou (přesnost asi 70 %)
        - Pak se stromečky opravovaly ručně
    - Data jsou podmnožinou ČNK
    - Založené na teorii Funkčního generativního popisu prof. P. Sgalla
    - 100 000 vět, 1.25 milionu běžných slov
    - **úrovně anotace:**
        - morfologie
        - analytická rovina (povrchově syntaktická)
            - přiřazují se funkce - podmět, přísudek, ...
        - tektogramatická rovina (4 poroviny):
            - závislostní struktura, (detailní) funktory
            - jádro/ohnisko (topic/focus) a hloubkový pořádek slov
            - koreference (většinou pouze gramatická)
            - vše ostatní (gramatémy):
                - detailní funktory
                - hloubkový rok, číslo, …
    - vznikl anotační manuál - některé jevy složité, bylo potřeba anotovat jednotně
        - 1 500 stránek
    - díky ručnímu zanotování (po atom. metodě) - lepší než většina treebanků, spolehlivější
    - projekt na 10 let
    - grant až 300 mil. Kč
    - Je anotovaný (opět automaticky) na několika rovinách (na některých rovinách jsou anotovány jen jeho části):
        - Slovní rovina (w-rovina). Pouze surový text bez anotace, ovšem včetně členění.
        - **Morfologická rovina** (m-rovina).
            - Každému slovu ve větě přiřadí několik atributů (lemma, tag (15-ti poziční značka), jednoznačné id využité při propojování rovin, odkaz do slovní roviny, atd.).
            - Anotace probíhala dvoufázově: nejdříve anotoval automatický morfologický analyzátor → a pak dva lidští anotátoři na sobě nezávisle vybírali správná lemmata a tagy z výsledků automatického → nakonec třetí lidský anotátor vybral nejlepší možnost z předchozích dvou.
        - **Analytická rovina** (a-rovina).
            - Každá věta je reprezentována stromem orientovaným do kořene s ohodnocenými hranami mezi uzly. Uzly jsou právě prvky morfologické roviny, hrany jsou ohodnoceny podle závislostních vztahů uzlů, či určují další jevy (koordinace – s předchozí větou, apozice, interpunkce). Každý uzel si i pamatuje své pořadí ve větě kvůli grafickému znázornění.
            - Byl použit automatický parser na předzpracování textu a dále automatický nástroj, který na základě pravidel určoval ohodnocení hran, ale výstup byl často chybný či neúplný, tedy museli nastoupit ruční anotátoři. Následně byly provedeny automatické kontrolní testy (např. slovesný jmenný predikát závisí na být) a porušení byla ručně opravena.
            - Nakonec byla provedena společná revize morfologické a analytické roviny (např. shoda v pádě, rodu a čísle závislého a nadřízeného uzlu, atd.).
        - **Tektogramatická rovina** (t-rovina).
            - Opět je každá věta reprezentována stromem. Nicméně jeho uzly už nemusí být právě prvky morfologické analýzy (některé prvky zde nejsou (např. předložky) a některé uzly tu jsou navíc (např. nevyjádřený podmět)). Zachycuje hloubkovou strukturu věty. K některým uzlům jsou připojeny gramatémy poskytující o uzlu informaci, kterou nelze jinak odvodit. K uzlům reprezentujícím sloveso či některé typy podstatným jmen je přiřazen valenční rámec (odkaz do vallexu). Dále nějaké koreference.

- **Anotační a vyhledávací nástroje**
    - **Netgraph**
        - prohledávání stromů
        - víceuživatelská internetová aplikace cleint-server
    - **TrEd**

        - tree editor
        - původně anotační nástroj
        - postupně přibily i vyhledávací funkce a další
- **Universal Dependencies**
    - společný formát korpusů
    - výsledek mezinárodní standardizace
    - dá se např. vyhledávat ve více korpusech najednou - lze porovnávat jazyky
        - např. v našem repozitáři LINDAT

Pravděpodobnostní a statistické metody

- Motivace: víme, že existují 3 překlady pro dané slovo. Je těžké určit, který je pro danou situaci vhodný. Nicméně mohl by nám k tomu pomoci kontext okolních slov. Statistické překladové metody v podstatě zkoumají, jakou mají různé kombinace slov v daném jazyce pravděpodobnost – a dle toho se rozhodují o překladu.
    - Potřebujeme k tomu velké množství událostí - při dostatečně dlouhé sérii pokusů se relativní četnost jednotlivých výsledků začne blížit jejich pravděpodobnosti.
    - Můžeme zkoumat (paralelní) korpus
        - četnost výskytu “in” v anglickém a tří možností ve francouzském
- Pravděpodobnost výskytu slova w v textu T je P(w) = počet výskytů slova S v textu T / počet slov textu T.
- **Základní vzorce**:
    - P(A,B) = P(A)P(B) (A, B nezávislé)
    - Bayesův: P(A|B) = P(B|A)P(A) / P(B)
    - p(A,B) = p(A|B) p(B) = p(B,A)
    - p(A,B,C) = p(A|B,C)p(B|C)p(C)
    - p(A|B)=p(A) a tedy p(A,B)=p(A)p(B) (A, B nezávislé)
- **Modelování jazyka** (rozpoznávání) je technika, která se snaží předpovídat, co bude následující slovo na základě předchozího kontextu (historie).
    - Dříve čistě akusticky, později (v 90. letech) se přidala statistika - zlepšení
    - Předpovídání - násl. morfologické značky, násl. slova, apod.
        - např. v angličtině po členu nebude sloveso…
    - Pomocí podmíněné pravděpodobnosti
    - Nechť jsme před slovem w. Označme h dosavadní historii (text před slovem w). Pak nás zajímá P(w|h). Což z Byesovy věty spočítáme jako P(w|h) = P(h|w)*P(w)/P(h). Díky větě o úplné pravděpodobnosti pak můžeme počítat pravděpodobnost celé věty W jako:
        - $p(W) = p(<w_i>_{i=1..n}) = p(w_n|<w_i>_{i=1..n-1}) * p(w_{n-1}|<w_i>_{i=1..n-2}) * p(w_{n-2}|<w_i>_{i=1..n-3}) * \ldots * p(w_2|w_1) * p(w_1)$
        - pravděpodobnost n-tého slova v kontextu předchozích
    - Jelikož příliš dlouhá historie by byla výpočetně náročná a zároveň by mnohé pravděpodobnosti byly příliš malé (kombinace dlouhých sousloví nejsou příliš pravděpodobné), tak se historie omezuje - v reálném případě počítáme s úseky délky 3 (trigram) nebo 4 (kvadrigram)
    - trigramový model:
        - p(W)= p(w3|w2w1)*p(w2|w1)* p(w1)
    - Termín **n-gram** znamená n-tice slov za sebou (lépe by však bylo upřesnit „slovní n-gram“, jindy se n-gramem totiž myslí n-tice písmen).
- **Vyhlazovaní.** Ve velkém slovníku je příliš mnoho nulových pravděpodobností (kombinací trigramů je hodně, ale v daných textech se jich vyskytne jen malá část). To se řeší tak, že nulové pravděpodobnosti se nahradí nějakými malými hodnotami.
    - Zkoumání by jinak přestalo být opřené o reálná dataa
    - Bohužel ale setře rozdíl mezi nesmyslnými kombinacemi a těmi málo pravděpodobnými
- **Statistický překlad**
    - Použije se se paralelní překladový korpus jako trénovací množina příkladů dobrého překladu
    - **Paralelní korpusy**
        - existují jak pro dvojice jazyků, tak i pro větší množiny (korpus dokumentů EU)
        - musí být spárované - po větách, větných členech, atd. (oproti práci překladatelů)
        - takový korpus bude většinou velmi malý (oproti dříve probraným jednojazyčným)
    - V posledním desetiletí převládající metoda
    - **Fázový překlad**
        - skládat překlady pouze z paralelních dat
        - zavedl Google
        - kolem r. 2000 pracovali s frázemi až do délky 9



    - bez ohledu na lingvistiku, na pravidla - čistě mechanické
        - spoustu různých variant

- **Metoda zašuměného kanálu**
    - Chceme překládat z F do A. Hledáme pravděpodobnostní model P(A|F), který vyjádří pravděpodobnost libovolné anglické věty a, máme-li francouzskou větu f. Parametry se nastaví podle tréninkového korpusu.
    - Bayesův vzorec: P(a|f) = P(f|a)P(a)/P(f)
        - tím se vlastně otočil směr překladu
        - hledáme 2 modely:
            - překladový model P(f|a)
            - jazykový model cílového jazyka P(a)
        - P(f) není podstatné, nějaká konstanta, větu máme
    - Předstíráme, že francouzská věta je výsledkem nedokonalého přenosu přes zašuměný (nespolehlivý) kanál a hledáme její správný originál.
    - Zkoumáme dobrou metodu překladu (akorát opačným směrem) a současně se díváme, jak dobrou (správnou, hezkou) větou v angličtině je hypotéza
    - Jazykový model P(A) - kvalitní, větší korpus
        - může být trigramový model založený na mnohem rozsáhlejším korpusu cílového jazyka, řádově stamiliony slov
    - Překladový model P(F|A) je založen na mnohem menším paralelním korpusu (miliony slov)
    - Překlad probíhá obráceně
    - Jazykový model odfiltruje nepodařené překlady, vyrovná chyby překladového modelu
        - problém: často se jim po cestě ztratí negace, to jazykový model nespraví (původní i negace jsou stejně přijatelné)
    - Jazykový model vybírá pouze “hezké” věty, nemá vztah k originálu
    - Hledání překladových hypotéz (dekódování) je obtížným problémem samo o sobě
- **Evaluace systémů automatického překladu**
    - Potřebujeme zpětnou vazbu, jestli se po změně systém zlepšuje, nebo ne
    - Jak měřit kvalitu překladu? To je obtížná záležitost i ručně, natož automaticky. V roce 2002 vznikla míra Bleu.
    - **BLEU**
        - metrika, standard
        - máme sadu referenčních překladů (daný text kvalitně přeložený, nejlépe několik variant)
        - skóre - vezmou se unigramy až kvadrigramy a porovnává se s referenčními větami (jestli se v některém z ref. překladů vyskytují)
            - je jedno, v jakém ref. překladu se vyskytuje, hlavně že v nějakém
        - penalizace za stručnost (např. pokud systém přeloží jen prvních pár slov věty ale kvalitně)
            - tendence favorizovat krátké věty, jejichž všechny n-gramy by existovaly v referenčních příkladech, přestože by ref. překlady byly výrazně delší
        - celkové skóre:
            - BLEU = BP * $(p_1 p_2 p_3 p_4)^{1/4}$
            - penalizace za stručnost vynásobená geometrickým průměrem n-gramové přesnosti pro n=1..4
            - výsledkem je vždy číslo mezi 0 a 1 (“přesnost překladu v %”)
        - určené pro použití během vývoje systémů - měl by porovnávat ten samý systém
        - potřeba více ref. překladů a více testovacích vět (~1000)
        - problémy:
            - vysoce flektivní jazyky - jen u špatného pádu velmi nízké skóre
            - jiný slovosled může způsobit velmi špatné výsledky v této metrice

        - nebere v úvahu morfologii, tedy pouze chybná koncovka (ale správný význam) pokazí skóre stejně jako úplně špatný překlad
        - hodně náročné na velikost trénovacích dat, proto se lépe překládá mezi „velkými jazyky“, kde se data lehko shání
- pro nepatrné zlepšení potřeba výrazně větší trénovací data
    - Google - pro o 0,05 lepší BLEU skóre potřebuje 2x větší data

SÉMANTIKA

- **Sémantika přirozeného jazyka**
  - Pomocí syntaxe můžeme rozlišovat gramaticky správně a nesprávně utvořené věty. Nicméně nic to neříká o jejich pravdivosti.
  - Je nutno rozlišovat mezi **významem** a **pravdivostí** věty.
    - Pravdivost je dána kontextem, není obsažena v jazyce. Jsou k ní potřeba různá pravidla a předpoklady světa, ze kterého vycházíme.
    - I nepravdivá sdělení mohou mít svůj význam.
    - U některých sdělení zase není možno ověřit pravdivost.
  - Je těžké obecně říct o větách, jestli mají stejný význam. (Pozorovali ho dobrovolně. X Byl jimi pozorován dobrovolně.)
  - **Vyplývání**
    - z pravdivé věty často vyplývají různé další skutečnosti (na základě obecných pravidel a zákonitostí), nicméně tyto zákonitosti nejsou stoprocentní, mohou mít výjimky, které nás předem nenapadnou. (Tučňáci jsou ptáci. => Tučňáci mají křídla a létají.)
    - to, že je věta pravdivá, mívá důsledky - věta nese víc informací
      - př. Kare prodal auto sousedovi. => Karel měl auto, už ho nemá, soused je od něj koupil a teď ho má.
  - Sémantika formálních jazyků často spojuje pravdivost s významem, pro přirozené jazyky je nutné zvolit jiné teorie.
  - **Fregeho princip kompozicionality** (Gottlob Frege, 1848-1925).
    - Význam složeného výrazu je jednoznačně určen významy jeho částí a způsobem jejich kombinace.
    - Tedy např. význam textu je určen významy jednotlivých vět a jejich poskládáním. Obdobně význam vět je určen významem jejich slov, atd.
    - Buduje se odspoda - význam slov a spojek (vztah mezi klauzulemi/slovy)
- **Lexikální sémantika**
  - **Význam slov** můžeme popisovat zase pouze pomocí nějakého (meta)jazyka, ten může být:
    - formální - např. vycházející z něčeho již vystavěného
      - vhodný matematický nebo logický kalkul (predikátová logika) či soustava sémantických rysů, sémů
    - přirozený (ten stejný nebo jiný)
  - nebo se dá pro popis v reálném světě využít kombinace jazyka a situace (předmětu)
    - Toto je křída.
  - Problémy:
    - Význam slova závisí i na kontextu okolních slov a vět (např. Střílení poslanců ohrožuje naši demokracii).
    - Význam slov není jednoznačný (oko, list, kohoutek, štěně, hlava…).
  - **Význam slov** můžeme nezávisle na kontextu popisovat pomocí významových (sémantických) tříd (rysů)
  - **Ontologie** je množina tříd objektů, která představuje klasifikaci objektů universa U, např.:
    - fyzické objekty,
    - kvantity,
    - vztahy,
    - vlastnosti,
    - akce (činnosti),
    - živé bytosti, atd.
    - třídy lze dále zjemňovat: slovesa pohybu, péče o tělo, změny, komunikace apod.
  - Dané slovo (objekt) pak popíšeme pomocí příznaků ke každé třídě: + (patří do ní), - (nepatří do ní), 0 (nezávisí na ní).
  - Ontologie jsou buď doménové (domain) (někde jsem našla, že zpracovává jen jednu doménu – obor; jinde že to je množina názvů oborů) či vrcholové (upper) (Top Ontology – prý množina nejzákladnějších výrazů, nezávislých na jazyku).
  - **Popis významu slov**

- **ve slovnících**
    - pomocí synonym,
    - pomocí definic,
    - pomocí množiny vybraných primitivních výrazů daného přirozeného jazyka
        - výkladový slovník - omezená množina slov, odpovídá zhruba charakteristickým rysům
    - pomocí speciálního metajazyka: sémantických rysů
        - význam slova se snažíme popsat posloupností charakteristik (+ operátory)
        - muž = HUM, MASK, ADU
        - dívka = HUM, FEM, -ADU
- pomocí **sémantické sítě**
    - forma sémantické sítě lépe zachycuje víceznačnosti
    - je možné pracovat s hierarchií pojmů
    - výhodnější pro počítačové zpracování
    - umožňují určit různé vztahy a směry vztahů mezi pojmy, tj. nejen hierarchii sémantických tříd, ale i vztahy napříč nimi
    - zabývají se vztahy jako hyperonimie (slovo nadřazené) a hyponimie (slovo podřazené), synonymie (ekvivalentní význam ale jiná forma) a antonymie (slova protikladná), meronymie (býti částí) a holonymie (obsahovat)
    - osvědčily se - jednoduché, přirozené a do určité míry to funguje
        - jen do určité míry, protože evropské vznikly překladem anglické
            - synonymita ale není stejná v různých jazycích…

- **WordNet**
    - 1993, George A. Miller z Princetonu
    - rozsáhlá lexikální databáze anglických slov
    - obsahuje podstatná a přídavná jména, slovesa a příslovce
        - seskupená do množin synonym, tzv. synonymických řad neboli **synsetů**
    - každý synset vyjadřuje určitý koncept (všechna slova v synsetu mají stejný význam)
        - mezi sebou jsou navzájem propojeny sémantickými a lexikálními relacemi
    - sítí je možno procházet v prohlížeči, nicméně vznikala hlavně ručně
    - veřejný, možné stáhnout
    - dnes ještě víc hesel, než j uvedeno na slidu
        - na slidu - ve verzi 3.0 - téměř 155 000 hesel a 117 000 synsetů
    - Z příkladu mi přijde, že to vypadá jako takový výkladový slovník – k danému slovu to vypisuje určité jeho významy. Pro dané významy to vypíše slovní popis, příklad a seznam synonym. Co z toho dělá sémantickou síť je asi to, že to tvoří vnitřní strukturu (ostatní pojmy fungují jako odkazy), kterou lze procházet a jsou tam zaznamenány i různé relace nadřazenosti/ podřazenosti/ býti částí apod.
- **EuroWordNet**
    - 1997, prof. Vossen z Amsterdamu
    - rozšíření WordNetu do více jazyků
        - nejdříve přidány holandština, italština a španělština
        - později francouzština, němčina, čeština a estonština)
    - oproti původnímu WordNetu zavedeny změny:
        - byla zavedena vrcholová ontologie (množina 63 nejzákladnějších výrazů (konceptů), nezávislých na jazyku)
            - protože tam bylo více jazyků
        - ke každému jazyku pak bylo vybráno 1000 základních konceptů tvořících jádra sítí slov, jazykově závislé
        - v Aj WordNetu získal každý synset jednoznačný identifikátor, díky kterému vznikl mezi-jazykový index (Inter-Lingual Index, ILI), jazykově nezávislý soubor indexů. Pak byly na sebe různojazyčné WordNety navázány a vznikly vztahy ekvivalence (EQ-relations)
- **Aplikace WordNetu**
    - Automatický překlad - může fungovat jako slovník

        - Jednak může pomáhat v počítačem asistovaném překladu (Computer Aided Translation). Překladatel si v něm může hledat významy slov, jejich synonyma, antonyma, příklady použití, slova odvozená apod.
        - Jednak společně s morfologickou a syntaktickou analýzou může díky tomu, že ukládá valenční rámce pro slovesa, sloužit k automatickému strojovému překladu (2 paralelní WordNety)
    - IE - Extrakce informací
        - umožňuje pracovat se sémantickými vztahy (zejména synonymie)
        - může sloužit při vícejazyčném vyhledávání
    - Určování významů slov (Word Sense Disambiguation)
        - zachycuje více významů slov - zdroj dat pro rozpoznávání jednotlivých významů
        - lze zavést klasifikační funkci pro rozhodnutí, o který jde, nebo podle příkladů rozhodnout
    - Reprezentace znalostí, odvozování využívající významů slov, vztah k sémantickému Webu
    - Vyhodnocování kvality překladu (zlepšení automatických metrik typu BLEU)
        - BLEU selhává u použití synonym, která nejsou v referenčním překladu
- **Problémy sémantických sítí**
    - cizojazyčné WordNety vznikaly především jako překlad toho anglického, tedy nezachycují typické vlastnosti jazyků
    - také vzniklo mnoho chyb a nekonzistencí
    - projekty přestaly být financovány a přestaly se rozvíjet
    - podobných výsledků (a lepších, rychlejších, s méně usilím) lze dnes dosahovat pomocí statistických metod
    - obecně v době Googlu apod. některé projekty jako je WordNet už nemají tak dobrý smysl
- **Reprezentace významu věty**
    - **predikátová logika 1. řádu**
    - + se přidávají nové vlastnosti
    - konstruuje logické formule z jednotlivých výrazů věty na základě principu kompozicionality
        - jednotlivým složkám věty náleží odpovídající části sémantického zápisu
    - Alík skáče jump(Alík), $\exists x\, x = Alík\ \&\ jump(Alík)$
    - Všichni psi skáčou. $\forall x\, dog(x) \rightarrow jump(x)$
    - Každý student podepsal petici. $\forall x\, student(x) \rightarrow \exists y\, petition(y)\ \&\ sign(x, y)$
    - Petici podepsal každý student. $\exists y\, petition(y)\ \&\ \forall x\, student(x) \rightarrow sign(x, y)$
    - problémy:
        - modalita, čas a postoj - kvůli nim jsou potřeba nové operátory, které mají jako argumenty formule
            - *possible*(F), *necessary*(F), *believe*(x,F), *true_at_some_time_in_the_future*(F)
        - presupozice - předpoklad, který musí být pravdivý, aby celá věta vůbec měla pravdivostní hodnotu
            - Jupiterův měsíc ma oranžové pruhy. - Jupiter musí mít právě jeden měsíc.
        - neurčitost (fuzziness) - nevystačíme s T/F hodnotami, potřebujeme jemnější dělení
            - Pavel je mladý. Většina špičkových sportovců dopuje.
    - dobře popsaný systém, umíme s ním pracovat, ale neadekvátní
    - popíše dobře jen jednoduché věty
- jakmile začneme predikátovou logiku rozšiřovat, musíme rozlišovat mazi funkcí a její hodnotou
    - Cena Big Macu je 20 Kč. X Myslím, že cena Big Macu je 20 Kč.
    - Vlevo můžeme nahradit výraz “Cena Big Macu” jeho hodnotou 90 Kč, dostaneme FALSE, ale vpravo nelze nahrazení provést (není to ekvivalentní tvrzení).
- **intenze výrazu**
    - samotný popis, charakteristika
    - intenzí pojmu čtverec je pravoúhlost a stejná délka stran
    - je to výraz sám o sobě
- **extenze výrazu**
    - souhrn věcí, které pod pojem spadají

    - větší množina, kterou výraz reprezentuje
- **Základní přístupy k sémantice**
    - **Modelově-teoretická sémantika**
        - pracuje s pravdivostními podmínkami vztaženými k určitému modelu.
        - reprezentantem je montagueovská gramatika:
            - syntaktické kategorie odpovídají sémantickým typům.
            - obsahuje základní (lexikální) výrazy a jejich interpretaci, syntaktická a sémantická pravidla.
    - **Kompozicionální sémantika**
        - vychází z principu kompozicionality.
        - používá různé reprezentace
            - sémantické rysy a jejich skládání
            - koncepty a převod (překlad) ze syntaktické reprezentace
            - logickou reprezentaci a zjišťování pravdivosti
- **Montagueovská gramatika**
    - původně Universal Grammar
    - americký logik Richard Montague (1930-1971), srozumitelně vyložená Barbarou H. Partee
    - teorie sémantiky přirozeného jazyka
    - založena na formální logice, zvláště na lambda kalkulu a teorii množin
    - používá pojmy intenzionální logiky a teorie typů
    - vychází z předpokladu, že neexistuje žádný zvláštní rozdíl mezi sémantikou přirozených a formálních jazyků (článek “The Proper Treatment of Quantification on Ordinary English” (1973))
    - první rozsáhlý pokus systematicky popsat sémantiku přirozeného jazyka
        - logici před Montaguem považovali přirozený jazyk za příliš mnohoznačný a nestrukturovaný pro formání logickou analýzu, zatímco lingvisté měli pocit, že formální jazyky nejsou schopny zachytit strukturu jazyků přirozených
    - sémantická pravidla úzce svázaná se syntaktickými (ale je to sémantická teorie, ne syntaktická (i když má v názvu slovo gramatika))
    - obsahuje sadu syntaktických kategorií
        - kategorie tvaru X/Y
        - postupným budováním kategorií lze použít i na velmi složité věty
            - nekonečně mnoho možných kategorií - lze použít libovolný počet lomítek pro nové kategorie
        - pro každou kategorii má množinu konkrétních slov
        - dále obsahuje syntaktická pravidla pro slova z těchto kategorií.
    - podobné teorii valenčních rámců
- **TIL**
    - Transparentní intenzionální logika
    - reaguje na fakt, že predikátový kalkul 1. řádu, který stále mnoho teorií používá k popisu významu jazykových výrazů, nedostačuje - intenzionální logika je vhodnější
    - je založen na modifikaci typovaného lambda kalkulu
    - nemá vlastní logické spojky, kvantifikátory apod.
    - je transparentní systém - formální aparát reprezentující způsoby, jakými jsou konstruovány objekty, nejsou pro TIL předmětem studia, pouze prostředkem ke studiu těchto konstrukcí
    - nepreferuje jistá vybraná slova jako tzv. logická slova (logické spojky, kvantifikátory apod.), jež by určovala charakter logiky
    - TIL aplikována na analýzu přirozeného jazyka se stává sémantikou založenou na pojmu **možných světů**
    - univerzum je v TIL chápáno jako množina společná všem možným světům, kromě možných světů se neuvažuje o tzv. možných individuích
    - v příkladu - nálepka individua (Alena), individuální koncept (ministr zahraničí)
    - $\tau\omega$ … reference k danému času a světu - $\Lambda\tau$ … možný čas, $\Lambda\omega$ … možný svět
    - populární v Brně - pomocí TIL se tam snaží propojit syntaktickou a sémantickou analýzu
- **Rozpoznávání vztahů v textu**

- Pochopení smyslu textu je ještě těžší než smyslu věty. Problém je, že věty v textu na sebe navazují a odkazují se (např. nevyjádřeným podmětem).
- **Anafora**
  - (slovo anafora má dva různé významy, třeba rozlišovat)
  - výraz, jehož interpretace závisí na kontextu
  - nemůžeme pracovat se samostatnými slovy, ale i s celou větou
  - **Exofora** - odkazování mimo text, zájmeno poukazuje k mimotextové situaci či skutečnostem
    - *Vidíš ho? Dejte mi, prosím, tyhle tři.*
  - **Endofora** - odkazování v rámci textu
    - **Anafora** - zpětně
      - *Petr se seznámil se sympatickou dívkou. Pozval ji do kina.*
      - *Petr vyzradil tajemství. To neměl dělat.*
    - **Katafora** - dopředu
    - nepříliš časté, v románech k vybudování napětí
      - Když se zlobí, není s Petrem žádná řeč.
      - Věřte tomu nebo ne, máme schodkový rozpočet.
      - *Vyšel jsem z domu. Věděl jsem, že jsem sledován. Když jsem se zastavil, zastavil se i on. Když jsem se ohlédl, dělal, že lelkuje. Měl na sobě stejný šedý kabát jako vždycky. Už ho důvěrně znám, estébáka Jiřího.*
  - anaforický vztah **předchůdce - následník**
  - typy anaforických vztahů:
    - **zájmena** a **"nulové výrazy"**
      - nevyjádřený podmět nebo jiný větný člen
      - s těmi se paradoxně pracuje lépe
        - ze syntaktického stromu poznáme, že chybí podmět → *sáhneme do předchozí věty*
      - *Petr si koupil vstupenku. Vsunul ji do kapsy. Byla děravá.*
    - určité **jmenné skupiny**
      - *Elektronický zesilovač Tesla* vs. *Toto zařízení…*
      - v systému MOSAIC nejsou tyto vztahy zachyceny
      - museli bychom vědět, že zesilovač je zařízení
    - **elipsa**
      - vypuštěné části výrazů na základě paralelismu s předchůdcem
      - vynechání části věty obsahující informaci, která je příjemci známa a bez níž větu dokáže pochopit
      - jmenná vs. slovesná
      - *Včera jsem šel pěšky. Kam? Domů.*
      - *Petr přinesl dva stoly. Dřevěný a kovový.*
      - *Petra půjde do kina. Jirka taky.*
    - textové **spojovací výrazy**
      - výrazy vyjadřující mezivětné souvislosti v textu
      - souřadicí a podřadicí spojky, výrazy jako *například*, *na jedné straně - na druhé straně*, *jednak - jednak*, *nejdříve - potom* apod.
      - je třeba závorkovat, odkud kam to patří k výrazu
  - důležitost pro aplikace:
    - získávání informací z textu
      - *Škoda představila nový model Octavie. Jde o pětidvéřové kombi, které má ...*
    - automatický překlad
      - *Otevřenou tabulku upravte podle potřeby. Uložte ji pomocí ikony v panelu nástrojů.*
    - dialogové systémy

            - Kdy jede nejbližší vlak do Ostravy? Má jídelní vůz?
        - řešení anafory:
            - je nutné využít celou řadu informací:
                - morfologické značky
                    - např. u zájmen musí být shoda v rodě
                - syntaktická struktura věty
                    - pomůže určit vhodné kandidáty na předchůdce
                    - valenční informace umožní doplnit elipsu
                - statistické přístupy
                    - pravděpodobnost výběru některého z určených kandidátů
                - aktuální členění
                    - témata zmíněna v základu a v ohnisku věty (na začátku a uprostřed) jsou odkazována různými způsoby - využito v algoritmu Zásoby sdílených znalostí
                - rozsáhlé pomocné znalosti
                    - ontologie, sémantické sítě, tezaury apod.

- **Zásoba sdílených znalostí**
    - od 80. let
    - prof. hajičová a prof. Vrbová (?)
    - Stock of Shared Knowledge
    - modeluje zásobu znalostí, o které mluvčí předpokládá, že ji sdílí s posluchačem
        - tato zásoba se mění v souladu s tím, co je “v centru pozornosti” v daném časovém okamžiku
    - každá věta má vliv na tuto “hierarchii sdílení”, avšak ne každý zmíněný objekt má stejný účinek
    - několik jednoduchých pravidel pro určení stupně sdílení
    - čím nižší číslo, tím aktivovanější
        - v ohnisku - dostane 0
        - v jádru - 1
    - pokud není slovo zmíněno ani asociováno v další větě, postupně zapomínáme (+2) až úplně zapomeneme
    - př. *The school garden was full of children. …*
        - *parents* asociováno s *children*, dostane 0 + 2 = 2 (0 za *chiildren*)
    - slabiny
        - množina asociovaných termínů
            - vybrané jen některé, může jich být více
            - nikdy nevíme dopředu, jaké termíny se objeví
        - nevíme, jestli zrovna +2 je správné (proč ne třeba +1, +3?)
    - dalo práci vymyslet příklad, na kterém vše dobře sedělo
    - dobrá myšlenka, ale špatné automatické zpracování
        - nějakou dobu na tom pracoval Tom Holan

SKRYTÉ MARKOVSKÉ MODELY

- **Hidden Markov Model - HMM**
- aplikace
    - analýza řečových signálů - 1. využití
    - morfologické značkování
    - rozpoznávání značek aut
- máme přirozenou sekvenci - znaků, slov, …
    - závislé, nějak uspořádané
- Markovova hypotéza
    - pracuje jen s omezeným kontextem - stačí pro dostatečně dobré výsledky
        - max. bigramy nebo trigramy
    - ve skutečnosti mohou být související slova od sebe libovolně daleko
- proč skryté? Snažíme se používat na jevy, které nevíme...
- jednoduchá realizace
    - pravděpodobnostní konečné automaty
- pravděpodobnosti přechodu mezi stavy
    - součet čísel vycházející z jednoho vrcholu = 1
- 3 úkoly:
    - určení s jakou pravd. mohla být pozorována nějaká značka
    - dekódování - hádání nejpravd. posloupnosti skrytých stavů
        - př. hádání nejpravděpodobnější posloupnosti morfologických značek nad slovy ve větě
            - nejprve učení (máme daný model, který ho naučíme)
    - naučení se statistického modelu
- učení - # skrytých stavů - všechny možné stavy
    - trénovací množina
        - např. slova + konkrétní morfologické značky
            - počet možných stavů dán počtem všech možných značek
    - najde pravděpodobnosti přechodů a jednotlivých stavů
- dekódování
    - graf - hledání nejpravd. posloupnosti je vlastně hledání nejkratší cesty v grafu (pravd. jsou ohodnocení hran)
    - matice - # stavů * # pozorování
        - se všemi možnými přechody a jejich pravd. (někde může být i pravd. 0)
- angličtina vhodnější než čeština - méně morfologických značek
    - u př. značky z PennTreebank - stavy jsou značky, pozorování jsou slova
    - (pro češtinu se u korpusu využívá něco přes 1000 různých značek)
- Viterbiho alg. - pro bigramy - v kvadr. čase v závislosti na velikosti vstupu
    - poměrně efektivní
    - pomocí něj se hledá nejpravděpodobnější posloupnost