Organizační úvod

Poznámka
První týden bude předtermín. Pokud to stihneme, tak místo přednášky.

Přednáška bude spíše informativní, jednoduchá. Zkouška bude snad prezenčně a písemná.

Úvod

Poznámka
Na pomezí mezi lingvistikou, informatikou, umělou inteligencí, psychologií, logikou a matematikou.

Poznámka (Podobory)
Hlavní rozkvět s nástupem statistiky do lingvistiky. Dnes už zase spíš ústup kvůli strojovému učení.

- Rozpoznávání (daleko těžší) a generování mluvené řeči (jednoduché).
- Fonetika (zkoumá zvuky, fóny, třídí je a klasifikuje – nauka o *tvorbě* hlásek).
- Fonologie (zabývá se pouze těmi zvukovými rozdíly, které nesou význam, základní jednotkou je foném, je to nauka o *funkci* hlásek).
- Morfologie (tvarosloví).
- Syntaxe (skladba).
- Sémantika (význam).
- Překlad.
- Formalismy (syntaktické, umělé jazyky, např. Chomského hierarchie).
- Korpusová lingvistika (datové sady – potřebujeme ohromné množství dat pro analýzu, učení, …, mohou obsahovat mnoho dalších věcí, např. popis syntaxe a tak dále).
- …

Poznámka (Problémy s významovou ekvivalencí)
Ekvivalence je závislá na znalostech (např. Kulaťák = Vítězné náměstí), na logických souvislostech (prodává = kupuje se od něho).

Naopak někdy můžou být v podstatě stejné, ale ne ekvivalentní (nakrájet salám ≠ nakrájet ze salámu, Česky se mluví na Moravě ≠ Na Moravě se mluví česky.).

Poznámka (Problémy s víceznačností a vágností)

- Kritika brazilského delegáta byla ostrá.
- V místnosti stojí zelené skříně a židle.
- Na recepci se dostavil i ředitel banky roku.
- Vysoká škola lesnická v Trutnově otevřela novou fakultu.
- Často loví tlouště na višni. (Často vs. často, oni vs. on, na višni vs. na višni, ty vs. toho tlouště.)
- Když slepice málo snáší, tak se vejce špatně shání.
- Páry vycházejí z lesa.
- Včera jsem viděl Frantu v tramvaji. (V tramvaji se vztahuje k podmětu / předmětu).
- Dědeček se rozložil na gauči.
- Závodnice se před závodem se soupeřkami oddávala sexu.
- Dálnice postavená Ruskem stála 10 miliard. (Rusko stavělo, nebo prošla Ruskem. Ne, postavil to pan Rusko)

  Stát se skvělou vědkyní jde i se dvěma dětmi.
- 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.
- Otec Emmons bude trénovat Australany. (Otec koho čeho, ne Otec kdo.)
- Padlým námořníkům se znovu rozsvítilo.
- V hotelu Corradu se za jeho nejslvanější éry scházely prostitutky. Často tam bydlely špičky ČSSD jako Miloš Zeman, Jiří Paroubek a …
- Každému osmému Čechovi hrozí chudoba, všichni v EU jsou na tom ještě hůř.
- Děti za korunu.
- Ochutnejte našeho řezníka.
- Inspektor Barnaby řeší případ vraždy otrávením úředníka sociální správy.
- Bramborové a švestkové knedlíky.

- Muž venčící psa, který v Praze pokousal dítě, se přihlásil na policii.
- Králem Tour je počtvrté Froome, dorazil se šampaňským.
- V lese viděli kolemjdoucí mrtvolu.
- Sarah Palinová řekla Ne, Obamovi se nepostaví.

Poznámka (Reklamní vložka)
Erasmus Mundus Language and Communication Technologies (EM LCT). 2 roky v cizině.

1 Morfologie

Poznámka (Historie)
Nejstarší odvětví lingvistiky (cca 4. stol. př. n. l. v Indii).

Poznámka (Předmět morfologie)
Předmětem morfologie je studium vnitřní struktury slov. (Na rozdíl od lexikologie – studuje slova jako jednotky slovní zásoby – a lexikografie – sestavuje slovníky).

Také studuje způsoby skloňování (deklinace) a časování (konjugace).

Definice 1.1 (Morfém)

Nejmenší znaková jednotka jazyka nesoucí význam. Existují 2 druhy: lexikální (nese význam slova jako takovéůo) a gramatické morfémy (určuje gramatickou roli).

Definice 1.2 (Dublety, alternace, alomorfy, autosémantická a synsémantická slova)

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

- Alternace – změna hlásek uvnitř kmene (vůz – vozu, švec – ševce, prkno – prken, …).
- Alomorfy – varianty kmene odvozené od stejného slovního základu (matka – matce – matek – matčin).
- Autosémantická (plnovýznamová) slova.
- Synsémantická (pomocná) slova.

Poznámka (Morfologická typologie jazyků)
Například by šlo dělit jazyky podle způsobu vyjádření rozdílu mezi jednotným a množným číslem v různých jazycích (např. jazyky mají: Japonština – nic, Tagolog – funkčním slovem, Turečtina + Svahilština – afixem = příponou, Angličtina + Arabština – zvukový rozdíl, Malajština – reduplikace = zopakování slova).

Dělíme jazyky na: Analytické (slovo = morfém, také izolační), syntetické (slovo > morfém, také flektivní (morfém = spojení gramatických významů) a aglutinační (morfém = 1 význam)), polysyntetické (slovo = věta, také polysyntetické). Jazyky však přebírají charakteristiky z ostatních jazyků, tedy hranice se dosti smývají.

Definice 1.3 (Morfologie založená na morfémech)

Vidí slovo jako řetízek morfémů (jako korálky na niti).

Definice 1.4 (Morfologie založená na lexémech)

Vidí slovo jako výsledek aplikace pravidel, která slovo mění a tím vytváří nový tvar.

Definice 1.5 (Morfologie založená na slovech)

Nám nejbližší. Centrální význam mají vzory. (Generujeme tvary podle toho, že víme, ke kterému vzoru patří.) Hodí se i tam, kde předchozí dva postupy selhávají (např, kde jeden morfém reprezentuje více gram. kategorií).

Definice 1.6 (2-level Morphology)

Systém zpracování morfologie vyvinutý dvěma Finy na začátku 80. let. Založený na automatech, společný byl mechanismus morfologie. Dvou úrovňový, protože jedna úroveň je generování slov, druhá je analýza již hotových (převratná byla pravě druhá úroveň). Pravidla se aplikují paralelně, nikoli sekvenčně (nezáleží na pořadí). Lexikální vyhledávání (= prohledávání slovníku = většinou prohledávání trie) a morfologická analýza (= uplatňování pravidel) probíhá současně.

Na češtinu moc nefunguje.

Definice 1.7 (Česká morfologie)

Vyvíjena od roku 1989 zejména prof. Hajičem. Využívá pozičních značek (značka = jednoznačný význam), značky jsou 15místné (2 jsou rezerva), ale rozeznává se 13 kategorií:

- POS (# = 10) – slovní druh
- SUBPOS (# = 75) – slovní poddruh
- GENDER (# = 8) – rod
- NUMBER (# = 4) – číslo, včetně duálu
- …

Např nejnezajímavější = AAFP33N (adjective, regular, feminine, plural, dative, no poss. gender, no poss. number, no person, …)

Poznámka (Činnosti využívající morfologii)
Morfologická analýza (výsledkem je seznam lemmat a značek popisujících jednotlivé kombinace gramatických kategorií spjatých s daným vstupním slovním tvarem).

Morfologické značkování (vybírání správné značky v daném kontextu – statistika).

Částečná morfologická desambiguace založená na pravidlech – pomocí (100% platných pravidel) spolehlivých pravidel redukuje počet značek, odstraňuje nevhodné, ale ponechává všechny, které nelze spolehlivě odstranit. (Tohle ve skutečnosti pomohlo o 0.05% oproti statistice, ale vedlo to k vytvoření softwaru na opravu syntaxe (pokud vypadnou všechny možnosti, kterou značku by mohlo slovo mít, je zde syntaktická chyba), který koupil MicroSoft a používá ho.)

Lemmatizace (výběr správného základního tvaru, ze kterého byl odvozen daný vstupní tvar. Klíčová operace pro vyhledávání v textech.)

Semming (odříznutí koncovky, na rozdíl od lemmatizace je základním tvarem kmen slova.)

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

TODO!!!

Definice 1.8 (Systém ASIMUT)

Jazykový modul: Neobsahuje žádný rozsáhlý slovník. Je založen na retrográdním (seřazen podle abecedy, ale podle slov pozpátku) slovníku dr. Slavíčkové (1975). (Slova, která jsou poblíž mají podobné vlastnosti, jelikož mají stejné koncovky.)

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

Problémy: Špatně se hledá, která slova chybí. Větší problém je, že některé slova se nedají rozeznat podle dokud je celé nepřečtete (speciálně životné vs. neživotné: právník vs. trávník, lazebník vs. sazebník), takže většinou je algoritmus nastaven tak, aby se generovali i neexistující tvary (používá se spíše ke kategorizaci slov). A moc nefunguje pro slovesa. Byl tedy brzy převálcován algoritmy používající slovníky.

Další pojmy: Negativní slovník (obsahuje nepodstatná slova při dotazování (spojky, citoslovce apod.)), Konkordance (předzpracování, přidělovali se adresy a frekvence pro porovnávání podobnosti významů).

Definice 1.9 (Systém MOZAIKA)

Morphemic Oriented System of Automatic Indexing and Condensation (systém pro indexaci dokumentů). Na rozdíl od standardního přístupu, kdy existuje slovník klíčových slov a podle něj se indexuje, využívá toho, že řada přípon a koncovek nese význam (v ANJ -er, -or, -tion, -ity, -ness, v češtině -ič, -ač, -čka, -ér, -or, -dlo, -metr, -graf, -fon, -skop, -ace, -kce, -áž, -ní, -za, -ost, -ita, -nce, -aný, -ený, -ací, -ecí).

Algoritmus - Vstupem je nijak nepředzpracovaný text, u kterého je zachována typografie. Lematizace a morfologická analýza poskytnou lemata a morfologické značky. Nalezená lemata jsou profiltrována negativním slovníkem a podle délky. Syntaktická analýza pomůže odhalit několikaslovné termíny. Váhy se přiřadí i podle umístění (nadpisy, první a poslední odstavce, první a poslední věty, atd.). Nakonec se vezme 10 výrazů s nejvyšším skórem, které se normalizuje (dělí se vahou nejdůležitějšího výrazu).

Výhody: není nutné vytvářet slovníky, pouze koncovek, lokální syntaktická analýza umožňuje větší flexibilitu při hledání termínů.

Nevýhody: pracné vytváření slovníků pravidel, neobsahuje řešení odkazů (= především zájmena)

2 Syntax

Definice 2.1 (Reprezentace)

Závislostní stromy – na rozdíl od střední školy mají všechny „tokeny“ (včetně interpunkce) vrchol a kořenem je pouze přísudek. Používá se spíše méně. Dá se zapsat šipkami nad textem. Zdaleka ne vše se s ním dá zachytit.

Složkový strom – slova/fráze (= složky) se po dvou skládají do frází, dokud nezbude jen jedna fráze. Trochu vhodnější pro pevné pořadí slov ve větě, proto se používá více. Dá se zapsat pomocí závorek. Odpovídá derivačnímu stromu bezkontextové gramatiky

Definice 2.2 (Neprojektivní konstrukce)

Při projekci slov svisle dolů protneme mnoho hran. Např. „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.“.

Každá cca. 7. věta je neprojektivní.