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.

3 Významné teorie

Definice 3.1 (Deskriptivismus)

Bloomfield 1933, dále Ch. Hockett a Z Harris – Chtěli zachránit jazyky před vyhynutím (indiáni).

Jazyková fakta klasifikuje a registruje, ale nevysvětluje. Zpracovává zejména povrchovou větnou strukturu.

Definice 3.2 (Analytická syntax)

Jespersen 1937 – matematičtější přístup.

Definice 3.3 (Logický přístup)

Ajdukiewicz 1935 – z toho vznikla kategoriální gramatika

Definice 3.4 (Povrchová/hloubková syntaktická struktura)

Jedna povrchová reprezentace (věta) může odpovídat více hloubkovým. Nebo naopak, jednu hloubkovou (význam) lze vyjádřit více povrchními.

Definice 3.5 (Transformační gramatika)

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

3 komponenty: báze (soubor bezkontextových pravidel generujících složkové stromy, tzv. frázové ukazatele), transformační komponenta (transformační pravidla operující na celých frázových ukazatelích), fonologický komponent (obsahuje regulární přepisovací pravidla, zaměřuje se na zvuk).

Množina přijatelných vět daného jazyka je vytvářena generativní procedurou, pomocí konečného počtu přepisovacích pravidel. Jde v podstatě o bezkontextovou nebo kontextovou gramatiku. (Ta ale není schopna zachytit vztahy mezi variantami vět, např. tázací vs. oznamovací.)

Na to je transformační složka, která obsahuje transformace, které jsou definovány strukturním indexem řetězců a strukturní změnou (rozřízne strom a prohodí podstromy, např transformace věty z aktivní na pasivní).

Transformační komponenta se však ukázala jako příliš silná (jazyky typu 0), takže se pokračovalo spíše v ústupu od této komponenty: Government-binding theory (GB) – teorie založená na obecných principech univerzální gramatiky (společné pro všechny lidi/jazyky/národnosti) a parametrech platných pro jednotlivé jazyky. Teorie minimalismu (znovu Chomsky) – obsahuje pouze dvě roviny, logickou (LF) a fonetickou (PF), sloužící jako rozhraní mezi zvukem (PF), reprezentací jazyka a významem (LF).

Definice 3.6 (Tree Adjoining Grammars)

Polovina 70. let – Joshi, Levy, Takahashi

Každému slovu je přiřazena elementární struktura (strom), kde se za další listy dá dosazovat (ale kořen dosazeného stromu musí mít stejnou kategorii jako list). Odpovídají bezkontextovým gramatikám, takže si musíme něco doplnit, aby zvládly i třeba neprojektivní struktury.

Definice 3.7 (Lexical Functional Grammar)

Rozlišuje 2 základní typy struktur: c-struktury (spojování slov do frází), f-struktury (reprezentují funkční vztahy ve větě (např. vazby sloves))

Definice 3.8 (Kategoriální gramatiky)

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

Kategorie mají formát a/b nebo $a \setminus b$, kde lomítko určuje pozici argumentu b, tedy zda je vpravo nebo vlevo od a. V čisté kategoriální gramatice se používají pouze dvě pravidla $X/Y\ Y \to Y\ X \setminus Y \to X$. (Vlastně kategorie znamenají, co je potřeba doplnit, aby věta fungovala.) Například Dexter likes Waren má kategorie $(NP, (S \setminus NP)/NP, NP)$ a skonči na S. Absolutně se nehodí na jazyky s volnou větnou skladbou.

Definice 3.9 (Unifikační gramatiky)

Popis vlastností objektů: Objekt je reprezentován množinou vlastností (jednoduchých rysů). Popis každé vlastnosti je dvojice název a hodnota.

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

Příklad
Slovo books: [graphematic_form: books, POS: noun, gender: neutral, number: singular].

Slova se skládají tzv. Unifikací, což je sjednocení množin. Pokud tam je nějaký spor ve vlastnosti, tak je nelze spojit.

Sestavy rysů mohou obsahovat i proměnné (doplnění (při unifikaci) pak probíhá na více místech).

Problémem je, že je možné unifikovat vlastnosti, které spolu nijak nesouvisejí. Druhým problémem je zoufalá časová neefektivita.

Definice 3.10 (Typované sestavy rysů)

Problém s unifikací nesouvisejících vlastností lze vyřešit tak, že sestava má i typ a máme někde poznamenáno, které typy mohou mít které vlastnosti.

Definice 3.11 (HPSG)

Jeden z nejvyspělejších unifikačních systémů. Slovo má 2 rysy (PHON (zvuk, fonetická forma) a SYNSEM (syntaktické a sémantické informace)).

3.1 Nástroje pro syntaktickou analýzu

Definice 3.12 (Augmented Transition Networks (Woods, 1970))

Reprezentován pomocí sítí, kterou se procházelo (CAT – hledání kategorie, JUMP – přechod do dalšího stavu bez hledání kategorie (nepovinné větné členy), SEEK – přechod k podsíti).

Definice 3.13 (Q-systémy (Alain Colmerauer 1969))

Poprvé nasazen v Kanadě při překladu meteorologických zpráv mezi angličtinou a francouzštinou.

Formalismus pro transformaci grafů (stromů). Grafy jsou linearizovány.

Atomy, stromy, seznamy a operátory a proměnné. Podle instrukcí přidává hrany (spojuje fráze). Následně čistí použité hrany a hrany vedoucí do vrcholu, ze kterého už nic jiného nevede.

Q-systém*y*, protože se počítá s tím, že se budou postupně aplikovat různé systémy (rozpoznání shodných přívlastků, analýza vět, analýza celého souvětí, …).

Reálná gramatika (např. RUSLAN = překlad čeština → ruština) je příšerně složitá.

Definice 3.14 (Funkční generativní popis)

Zakladatel Petr Sgall (1967), později E.Hajičková, J. Panevová. Navazuje na Pražskou lingvistickou školu.

Zkoumá mimo jiné i změny významu při změně slovosledu.

Základní vlastnosti: Stratifikační teorie (pracuje s 5 (původně) rovinami – fonetická, fonologická, morfématická, povrchová a tektogramatická (ekvivalent hloubkové)). Formy a funkce – jednotka na vyšší rovině reprezentuje funkci jednotky na rovině nižší (tektogramatická rovina je nejvyšší). Závislostní reprezentace. Teorie valence (vazby vyžadované nebo povolené řídícími slovy, zejména slovesy.)

Definice 3.15 (Valence)

Dva základní druhy závislých členů na TG rovině: Aktanty (Konatel, Patient, Adresát, Origo, Efekt, každý může být ve větě zastoupen pouze jednou) a Volná doplnění (mohou se vyskytovat vícekrát).

Další dělení je na obligatorní a fakultativní: Obligatorní aktant nesmí ve větě chybět (může chybět na povrchové rovině, pokud ho známe např. z kontextu).

Používá se tzv. dialogový test, tj. pokud se zeptáme na aktanta a nemůžeme odpovědět „Nevím.“, tak je aktant obligatorní (např. Moji přátelé přijeli. „Kam?“ Naopak „Odkud?“ už může být „Nevím“.)

Definice 3.16 (Kontrola gramatické správnosti)

Problém je s tím, že spousta vět může gramaticky dávat smysl. Např: „Sportovci věnovaly plyšáka.“ (podmět nevyjádřený) „Tatínek šli do práce.“ (Kdo, co, kam odnesl? Šle ≈ mašle.)

Definice 3.17 (RFODG)

TODO

Založeno na měkkých (lze je porušit) a tvrdých (nelze je porušit) pravidel. Nejdříve se zkusilo vše tvrdě, pak se teprve některá pravidla změkčila.

Zjistilo se, že je téměř nemožné vybudovat takovou gramatiku.

Definice 3.18 (LanGR)

TODO

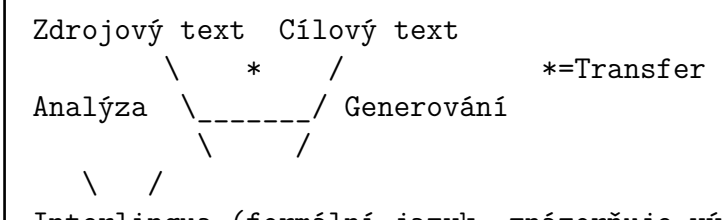
4 Automatický překlad

Poznámka
Slovník nám nestačí (Sejít se hlava skupina jeden pět hodina velký muž.) Je třeba i tvarosloví, to však také nestačí (Sejde se hlava skupina jeden pět v hodině velký u muže). Je třeba i syntaxe[a], to už skoro je, ale (Sejde se hlava skupina ve čtyři hodiny u velkého muže). Chybí ustálená spojení – idiomy (Rada starších se sejde ve čtyři hodiny u náčelníka). To však stejně nemuselo být všechno. Zvykem (jednou z tzv. reálií) totiž v tomto smyšleném příběhu je počítat hodiny od východu Slunce, věta:překlad – 1:0.

„Lidský překlad:“ Rada starších se sejde v 11 hodin dopoledne u náčelníka.

[a] VSO = Sloveso + podmět + předmět (pořadí větných členů).

Poznámka (Prehistorie)
1933 první patenty pro překladové stroje (George Artsruni a P. P. Smirnov-Trojanskij). Trojúhelník (Vopuázův):



Dále 1946 (A. D. Booth) – idea automatického dvojjazyčného slovníku, text zpracováván slovo od slova.

1948 (R. M. Richens) – ve slovníku nejsou zachycená celá slova, ale předpony, kmeny a přípony zvlášť.

1950 (E. Reifler) – zavádí pre-a post-editing (člověk, ne nutně lingvista / překladatel, který upraví text, aby se lépe překládal (zpřesnit výrazy atd.) / aby dával větší smysl (po překladu)).

1952 – První konference o strojovém překladu na MIT, (L. E. Dostert) – pivotní jazyk pro překlad více jazyků (např. angličtina).

7. 1. 1954 – Georgetownský experiment (do roku 1956) – 45 vět s 250 slovy, 6 syntaktických „zákonů“, jednoduché oznamovací věty bez negací, slovesa ve 3. osobě, málo předložek. Ruština -> Angličtina.

1956 – První mezinárodní konference (12 vědeckých skupin na amerických univerzitách).

? – zpráva ALPAC (má se více investovat do lingvistiky – vedlo k zatrhnutí veškerých financí v USA).

1857 – N. Chomsky.

1960 (Y. Bar Hillel) – „Vysoce kvalitní plně automatický překlad nemůže být nikdy dosažen.“ („The box was in the pen.“)

1976 – METEO (TAUM, 1976) – překlad meteorologických zpráv $A \to F$. Dobře definovaná a výrazně syntakticky i sémanticky omezená množina. Systém umí poznat, že něco neumí přeložit.

Další významné systémy: SYSTRAN (překlad dokumentů EU, přímý překlad mezi 20 páry, ovšem uspokojivá kvalita pouze u nejstarších párů (A-F-N), problémy řešeny ad hoc), EUROTRA (projekt EU v 80. letech, 72 jazykových párů, nezvládnutá modularita, do jisté míry podobný negativní efekt jako zpráva ALPAC), VERBMOBIL (německý nástupce EUROTRy, překlad mluvené řeči, předváděn na Světové výstavě v Hannoveru, od té doby se o něm příliš nepíše)

Poznámka (Současné trendy)
Statistické metody (využívají paralelní značkované korpusy, rychle se zlepšují, pro měření kvality však používají technické metody založené na referenčních překladech – BLEU score)

Definice 4.1 (Současné trendy, statistické metody, překladová paměť)

Používají se systémy, které jen rozpoznávají shodu v textu a nabízejí překladateli informace. Zejména se hodí na manuály, dokumentace, atd. Protože tam se mění jen část, tedy se překládá pořád to samé.

První IBM Translation Manager, Dejá Vu (pouze překladatelská tabulka neznámých frází bez kontextu, ale mnohokrát využitá fráze se přeloží jen jednou), TRADOS Translator's Workbench (překlad přímo textu, tedy s kontextem).

Definice 4.2 (České systémy)

První věta přeložena v r. 1957 na počítači SAPO.

„The consonants have not by far been investigated to the same extent as the vowels."

„Souhlásky zdaleka nebyly prozkoumány do stejné míry jako samohlásky."

Poté se moc nedělo, až když byly přivezeny Q systémy. Pak vznikl systém APAČ (překlad z angličtiny do češtiny, autor Dr. Zdeněk Kirschner, vytvořen ve stejném formalismu jako METEO, překlad textů z oblasti vodních pump, transdukční slovník (-action => -ace, -ic => -ický, …)), systém RUSLAN (překlad manuálů k operačním systémům sálových počítačů, překlad 1 věty trval cca 4 minuty, transdukční slovník (fungoval jen na cizí názvy i mezi slovanskými jazyky), minimální transfer, výzkum zastaven v roce 1990 těsně před provozními zkouškami), nakonec systém Česílko (chtěl překládat ručně do češtiny a z ní pak do všech slovanských jazyků (stejná syntaxe), první pokusy se slovenštinou, už tehdy se přišlo na to, že je potřeba plný dvojjazyčný slovník a že je úplně jiné tvarosloví).

5 Sestavování korpusů

Definice 5.1 (Korpus)

Sbírka textů doplněná o nějaké informace. (Podle informací můžeme mít různé druhy korpusů.)

Chceme po nich: výběr vzorků a reprezentativnost, konečnou velikost (aby spočítané výsledky platily, lze vyřešit verzováním), strojem čitelná forma, standardní reference (aby sloužil širšímu publiku musí značkovat tak, aby se v tom lidé vyznaly, tedy nechceme vyrábět svoje vlastní značky).

Například
První korpus Brown?

První a nejznámější syntakticky anotovaný korpus PennTreebank, který měl texty z burzy.

Poznámka
Na začátku jsme si říkali o anglických, německých a dalších korpusech.

Definice 5.2 (Český národní korpus)

Výsledek společného úsilí UK, MU, ÚJČ. Ústav českého národního korpusu existuje od 1994 na FF UK.

ČNK je značkován na morfologické úrovni. Stovky milionů slov. Ze začátku 15 % literatury, 60 % noviny, 25 % technické a odborné texty.

Definice 5.3 (Pražský závislostní korpus)

Malá podmnožina ČNK. Syntakticky anotovaný.

Vznikal cca 10 let, paralelně vznikala příručka, jak anotovat syntaxi.

Založen na teorii Funkčního generativního popisu profesora P. Sgalla.

Úrovně anotace: Morfologie, analytická rovina (povrchově syntaktická) i tektogramatická (závislostní struktura, jádr / ohnisko a hloubkový pořádek slov, korference, vše ostatní).

Poměrně jedinečný, většina syntaktických korpusů je dělána automaticky.

Definice 5.4 (Universal Dependencies)

Pokus o sjednocení způsobu anotací v korpusech. (Švédsko, Upsala.)

6 Pravděpodobnostní a statistické metody

Definice 6.1 (Vyhlazování)

Pokud chceme studovat pravděpodobnost věty podle pravděpodobnosti trigramů (delší gramy už jsou nemyslitelné), tak selžeme na tom, že nemáme dostatečně velká data, takže nám často vyjde pravděpodobnost věty (součin pravděpodobnosti trigramů) rovná 0.

Opravou je tzv. vyhlazování – nahrazení nul miniaturními čísly. (Tím však ztratíme možnost říct, že je něco 100% špatně.)

Definice 6.2 (Statistický překlad)

Základní myšlenka = použít paralelní korpus jako trénovací data. (Např. korpus dokumentů EU.)

Definice 6.3 (Metoda zašumněného kanálu)

Hledáme pravděpodobnostní model, podmíněnou pravděpodobnost libovolné např. anglické věty, máme-li např. francouzskou.

Francouzskou větu bereme jako anglickou, která přišla „zašumněným“ kanálem a hledáme anglickou, která byla nejpravděpodobněji „zdrojem“.

Definice 6.4 (Metrika BLEU)

Oceňuje kandidáty na překlad podle referenčních překladů vytvořených lidmi. Nejprve hledá v kolika unigramech, pak bigramech, pak trigramech nakonec 4gramech se překlad shoduje s nějakým (libovolným) referenčním překladem. Celkové skóre je pak čtvrtá odmocnina ze součinu poměrů počtu nalezených a nenalezených n-gramů vynásobená penalizací za menší počet slov?

Rozbíjí se na jazycích s velkou flexí, jazycích s volnou větnou stavbou, …

Je třeba cca 4 překlady a 1000 různých vět.

Nehodí se na porovnávání systémů, lze ji používat jako vývojovou metriku (tj. metriku na porovnávání, jak jde vývoj – rychlé zhodnocení).