Úvod do počítačové lingvistiky

(za případné chyby se omlouvám, nejsou úmyslné :o)

ASIMUT (Automatická selekce informací metodou úplného textu)

Vytvořen na MFF v 90. letech minulého století.

Jeho úkolem je vyhlédávání informací v úplném textu (tj. textu, který není předem nijak upravován).

Tvoří jej 2 části:

1. vyhledávací část (dotazovací modul)
2. jazykový modul

Text v databázi byl předem upraven - jednotlivé dokumenty byly rozděleny na sekce, části, odstavce, věty a slova.

Důležitou součástí systému byl také **negativní slovník**. Obsahoval slova, které nenesou žádnou informační hodnotu nebo se v textu opakují příliš často (typicky spojky, zájmena apod.).

Při vkládání nového textu do databáze se vytvářela tzv. **konkordance** (v podstatě speciální index, přes který se poté vyhledávalo).

Vyhledávací část

Příklad dotazu: *vzdálenost!, odstup! -3- rodinný! -1- domek!*

Operátory určující vzdálenost slov (**distanční operátory**):

- *-1-* slova musí být vedle sebe
- *-2-* mezi slovy můžou být nejvýše 2 další slova
- *-3-* slova musí být ve stejné větě
- *-4-* slova musí být ve stejném odstavci

Další operátory:

- *,* (čárka) operátor disjunkce
- *!* (vykřičník) operátor skloňování

Jazykový modul

Prošel několika verzemi - v té definitivní umožňoval uživateli pouze zadat slovo (podstatné nebo přídavné jméno) a bez dalších doplňujících informací jej uměl vyskloňovat.

Metoda použitá v jazykovém modulu byla založena na vlastnostech jednotlivých koncových segmentů základních tvarů - jako podklad posloužil **retrográdní slovník** (hesla v retrográdním slovníku jsou řazená abecedně od posledního znaku slova). Proto místo rozsáhlého slovníku základních tvarů byly zapotřebí jen **seznamy vyjímek pro jednotlivé koncovky** (ty nejdelší měly několik desítek položek).

Tam, kde je to nutné, systém dokáže odvodit potřebné odvozené kmeny (*patro/pater*).

Pozor! Ne vždy lze spolehlivě určit skloňování - např. *právník/trávník* (mužský živ. vs neživ. tvar). Proto systém **přegenerovává** (generuje i nesprávné tvary - ty se ale většinou v textu neobjevují).

MOSAIC (Morphemics Oriented System of Automatic Indexing and Condesation)

Metoda vytvořena v 70. letech minulého století na MFF UK (Z. Kirschner a kol.)

Cílem byla metoda pro automatickou indexaci nezkracených dokumentů, avšak metoda MOZAIKA (jak název napovídá) umožňovalá také kondenzaci textů.

Vstupem metody je text, nejlépe bohatě strukturovaný.

Myšlenka: tvoření slov vykazuje jistou pravidelnost - zakončení může nést sémantickou informaci.

V angličtině např.:

- *-er, -or* - původce děje, osoba vykonávající činnost
- *-ity, -ness* - vlastnost

V češtině např.:

- *-ík, -ič, -čka, -ér* - původce děje, osoba vykonávající činnost
- *-graf, -skop, -metr* - přístroj
- *-ace*, *-kce*, *-áž*, *-ní*, *-za* - proces, činnost

Algoritmus

1. **Lematizace** - určení základního tvaru, ze kterého vznikl tvar nalezený v textu.

   U metody MOZAIKA byla lematizace spojená s morfologickou analýzou - rozpoznávaly se i morfologické charakteristiky příslušného slovního tvaru (pád, rod, číslo atd.)

2. **"Selekce"** - nalezená lemata jsou podrobena kontrole - existuje několik omezení:

   a) Některé koncové segmenty se můžou spojovat s kmeny slov, které do dané tematické oblasti nepatří.

   b) Kromě omezení na na určité kmeny existují i omezení na určité slovní tvary (některá slova se můžou v odborném textu objevit pouze v určitém pádě nebo čísle).

   c) Také je možné formulovat omezení určitých kombinací znaků, které se nesmí objevit v základním tvaru (zejména u těch koncových segmentů, jenž jsou víceznačné).

   d) Lze použít i negativní slovník pro speciální vyjímky.

3. **Zjednodušená syntaktická analýza** - soustřeďuje se zejména na analýzu jmenných skupin.

   Příklad: Termín *operační zesilovač TESLA KC 415* charakterizuje text více než samostatný termín *zesilovač*.

4. **Výpočet vah** jednotlivých termínů.

   Jednotlivým slovním tvarům jsou přiřazeny váhy během morfologické analýzy. Pozdější jsou na základě synktatické analýzy přepočítány (přiřazeny celým složeným termínům).

   Váhy se určují na základě:

   - Pozice slova v textu (nadpis, podtitul, shrnutí etc.)
   - Délky slovního spojení (delší spojení mají větší váhu)
   - Vztahu k jiným výrazům (zejména na principu inkluze - jeden termín je zcela obsažen v jiném termínu). Toto ma svůj význam, protože často je delší termín na určitém místě textu (typicky v následující větě) zastoupen pouze kratší verzí.

5. **Normalizace** - je důležita pro porovnávání dokumentů různé délky.

Morfologie, morfologická analýza, morfém

Morfologie

Zabývá se tvořením tvarů slov a jejich významem, v širším smyslu i tvořením nových slov. Různé tvary slov vznikají skloňováním (deklinace) a časováním (konjugace).

Morfologická analýza

Úkolem **morfologické analýzy** je najít základní tvar slova a nalezení charakteristik gramatických morfémů.

Morfém

Nejmenší jednotka jazyka, která nese význam. Dělí se na:

a) lexikální (kmen slova) - určuje význam slova: za-**hrad**-ní

b) gramatický - určuje morfologickou charakteristiku slova (slovesný tvar): za-hrad-**ní**

Alternace a alomorfy

Alternace je změna hlásek uvnitř kmene. **Alomorf** je soubor různých tvarů (alternací) jednoho kmene.

Lematizace

Generování

Generování je opačný proces k morfologické analýze (lematizaci). Vyžaduje slovník lemat a slovník koncovek (a taky soubor pravidel pro skloňování).

Kontrola pravopisu, kontrola překlepů

Kontrola pravopisu

Různé přístupy:

1. **word list** - slovník korektních slov daného jazyka (použitelný pouze pro jazyky bez skloňování)
2. slovník lemat + koncovky + skloňování (vyžaduje generátor)
3. zakazané kombinace hlásek

Kontrola překlepů

Levenshteinova metrika ("vzdálenost dvou slov") je definována jako počet operací nutných k transformaci jednoho slova na druhé. Těmito operacemi jsou:

- vložení písmene
- nahrazení písmene
- smazání písmene

Je základní metodou pro vytváření nabídky oprav.

Ontologie

Ontologie je nauka o významu.

Syntaxe přírozeného jazyka - valence, závislostní a složkové stromy

Syntax

Syntax je lingvistická disciplína zabývající se vztahy mezi slovy ve větě, správným tvořením větných konstrukcí a slovosledem.

Syntaktická analýza

Synktatická analýza se zabývá stavbou věty a vztahy vět v souvětích.

Závislostní a složkové stromy prezentují dva různé datové typy pro znázornění syntaxe přirozeného jazyka.

Valence

"Valencí rozumíme v lingvistice schopnost lexikální jednotky, především slovesa, vázat na sebe jiné výrazy a mj. tak zakládat větné struktury."

Slovesný valenční rámec

Slovesný valenční rámec je tvořen (v širokém smyslu) všemi doplněními, které mohou dané sloveso v daném významu rozvíjet. V užším smyslu je **slovesný valenční rámec** tvořen *aktanty* a *obligatorními volnými doplněními* slovesa. Každé sloveso má alespoň jeden valenční rámec, často má ovšem válců více.

Aktanty jsou doplnění slovesa charakterizována dvěma podmínkami:

1. nemohou se vyskytovat více než jedenkrát (bez apozice nebo koordinace)
2. jejich kombinace je charakteristická pro jednotlivá slovesa

Závislostní strom (D-strom - dependency tree)

Formální prostředek pro znázornění struktury věty - modelování pomocí závislostí. Závislostí lze vyjádřit jazykový jev označovaný jako podřízenost (existují i jiné synktaticky rozlišitelné jevy: koordinace a apozice).

Za **závislý člen** se považuje ten člen, který lze vynechat, aniž by věta ztratila synktatickou správnost.

Neformální definice: Každé slovo ve větě, až na jedno, je (přímo) závislé na některém jiném slově. Tyto vztahy představují částečné uspořádání, které lze znázornit jako strom. Uzly stromu odpovídají jednotlivým výskytům slovních tvarů - nesou v sobě informaci o slovním tvaru, o jeho pořadí ve větě a o uzlu, na němž tento uzel závisí. Uzel, který nezávisí na žádnem jiném uzlu, tvoří kořen stromu.

Koordinace - např. *Adam a Eva*.

Apozice - např. *Hamlet, princ dánský.*

Projektivita, neprojektivita

Složkový strom (C-strom - constituent tree)

Zavedl jej Noam Chomsky.

Lze na něj pohlížet jako na derivační strom bezkontextové gramatiky. Složkový strom postupně dělí větu na části (složky) a to bez ohledu na závislosti (podřadné, řídící složky) nebo souřadné složky.

Q-systémy

Autor: Alain Colmerauer (1959)

Jedná se o **formalismus pro transformaci grafů** (přesněji linearizaci grafů).

Příklad výsledku: S(NP, VP(V, NP))

Základní vlastnosti:

- Grafový analyzátor (*chart parser*)
- Tři typy objektů: atom, strom a seznamy stromů
- Tři typy (implicitních proměnných):
  1. atom (konstanta) - písmena za začátku abecedy (A-J)
  2. strom - písmena ze středu abecedy (L-N)
  3. seznam - písmena z konce abecedy (U-Z)
- Operátory: *-DANS-*, *-HORS-*, *-ET-*, *-OU-*, *-NON-*, =, *''*

Příklad: S(NP,VP(V,NP)) může být popsáno jako: A*(U*), nebo S(NP, L*) nebo M* apod.

***** signalizuje, že se jedná o proměnnou.

Pozn.: Q-systémy byly použity např. u českého systému RUSLAN.

Gramatiky

Gramatika je struktura popisující formální jazyk.

Transformační gramatika

Navazuje na předválečnou americkou lingvistiku, snahu o explicitní popis jazykových pravidel. Transformační gramatiku zavedl Noam Chomsky v knize **Syntactic Structures** (1957).

Tvoří jí 3 základní komponenty:

1. **Báze** - soubor bezkontextových pravidel, generující složkové stromy - tzv. frázové ukazatele (*phrase markers*)
2. **Transformační komponenta** - soubor transformačních pravidel operujících na celých frázových ukazatelích. Dělí se na obligatorní a fakultativní (tj. povinné a volitelné).

   Transformace jsou definovány strukturním indexem řetězců a strukturní změnou - příklad:

   Pravidlo: NP_1 - V - NP_2 => NP_2 - was - V+en - by - NP_1

   Důsledek: John chose a book. => A book was chosen by John.

   Transformace jsou velmi silné (odpovídají Turingovým strojům).
3. **Fonologická komponenta** - soubor regulárních přepisovacích pravidel přidělující řetězům morfémů fonetické interpretace.

Generativní procedura - soubor konečného počtu přepisovacích pravidel, vytváří množinu správných vět daného jazyka. Jde v podstatě o kontextovou nebo bezkontextovou gramatiku. Není schopna zachytit vztahy mezi variantami vět (např. mezi větou tázací a oznamovací).

V Syntactic Structures také Chomsky zavedl následující dělení jazyka (od kterého později postupně ustoupil):

1. Povrchová (analytická) rovina (S-structure) - pracuje s vztahy mezi slovy v rámci věty
2. Hloubková (tektogramatická) rovina (D-structure) - pracuje s významem slov

Rozdíl mezi hloubkovou a povrchovou rovinou (dle Chomského): hloubková rovina je stejná pro každý jazyk, kdežto povrchová rovina je pro každý jazyk různá (protože se liší syntaxe jazyka).

FGD (Functional Generative Description)

Funkční generativní popis češtiny se rozvíjí od 60. let 20. století (zejména J. Sgall).

FGD je **závislostní** typ formalismu, který byl navržen pro účely teoretického popisu struktury českých vět. Základní charakteristikou FGD je **stratifikační**[1] **přístup** k popisu jazyka - popis jazyka je rozdělen do několika rovin. Každá z rovin je množinou zápisů vět, každá má svou syntax.

1. Rovina podkladové reprezentace (tektogramatická rovina)
2. Rovina povrchové syntaxe
3. Morfematická rovina
4. Fonologická rovina

Jak již bylo zmíněno, FGD využívá závislostního formalismu. Na tektogramatické a povrchové rovině je věta zachycena jako **závislostní strom**. K popisu věta na nižších rovinách už není třeba strom, stačí řetězec.

DG (Dependency Grammar)

Gramatika pracující se závislostními stromy.

TAG (Tree Adjoining Grammars)

Vznik: polovina 70. let minulého století

Autoři: Joshi, Levy, Takahashi

Specifickým rysem tohoto formalismu je, že nepracuje s řetězci, ale se stromy. Jsou **silnější než bezkontextové** gramatiky, ale **slabší než kontextové**.

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

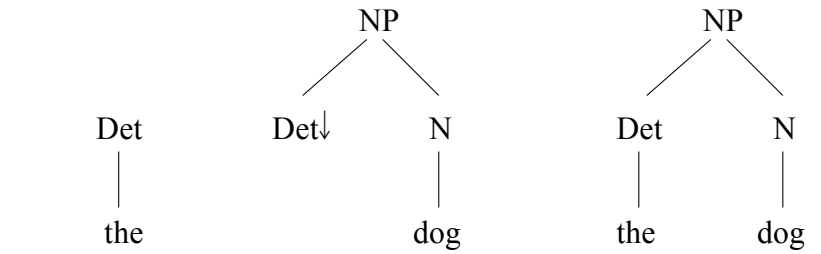
V pozdějších letech byla definována rozšíření TAG:

1. **LTAG** - každý elementární strom musí mít alespoň jeden list ohodnocený termínálním slovem (každý strom nese nějaký lexikální význam).
2. **FTAG** - každý vrchol elementárního stromu má přiřazeny dvě proměnné top a bot - tyto hodnoty jsou později kontrolovány při substituci a připojování (díky tomu lze zajistit např. shodu v pádě a čísle).

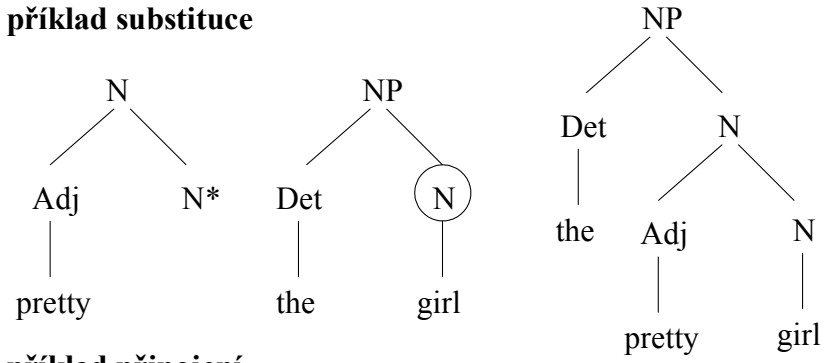
1 Stratifikace = rozvrstvení.

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

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



příklad substituce



příklad připojení

Unifikační gramatiky

Unifikační gramatiky jsou v jistém smyslů protipólem transformačních gramatik. Transformační gramatiky popisovaly synktatickou strukturu věty tak, že hlavní součásti popisu byla informace o způsobu vzniku této struktury a struktura samotná byla pak spíše jen vedlejším výsledkem. Oproti tomu **unifikační gramatiky** jsou založeny na popisu (statické) synktatické struktury pomocí informace získané na základě pozorování. Celková informace se pak skládá z informací dílčích a základním kamenem takové informace je popis jedné pozorované informace - **rys**. Formalní definici rysu tvoří dvě části: název vlastnosti a hodnota.

Feature structure (sestava rysů)

Množina vlastností (rysů) daného objektu je definována jako **sestava rysů**. Platí pro ní přirozené předpoklady, jako např. jednoznačnost jmen vlastností, nevýznamnost pořadí rysů atd.

Jméno vlastnosti je identifikátor, hodnota může být atomická nebo komplexní (další sestava rysů nebo seznam hodnot).

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

$$\begin{bmatrix} \textit{grafématický\_zápis} : \textit{knihou} \\ \textit{slovní\_druh} : \textit{podstatné\_jméno} \\ \textit{rod} : \textit{ženský} \\ \textit{číslo} : \textit{jednotné} \\ \textit{pád} : 7 \end{bmatrix}$$

Unifikace sestav rysů

Je obdobou operace sjednocení množin. Výsledkem unifikace dvou sestav rysů patřících k jednomu objektu je jedna sestava rysů obsahující rysy obou unifikovaných sestav. Je partné, že může dojit k situaci, že unifikované sestavy rysů nesou informace, jenž jsou spolu **nekompatibilní**. Výsledek takové unifikace se označuje jako sestava nesoucí "příliš mnoho" informace a značí se $\perp$.

Použití unifikace v přepisovacích pravidlech

Pokud pozměníme přepisovací pravidla tak, že místo neterminálů dosadíme sestavy rysů, můžeme místo operace identifikace při odvozování používat operaci unifikace a to tak, že přepisovací pravidlo lze použít vždy, když výsledkem unifikace je platná sestava rysů.

LFG (Lexical Functional Grammar)

Lexical = (not transformational) bohatý lexikální slovník, který obsahuje např. informace o vztazích mezi různými tvary sloves.

Functional = (not configurational) gramatické pojmy jako podmět a předmět jsou primitivními typy, t.j. nejsou definovány přes konfiguraci frázových složek.

LFG dělí jazyk do několika struktur, dvěma hlavními strukturami jsou:

1. Gramatická reprezentace (**f-structure** neboli feature structure) - používá matice (atribut × hodnota)
2. Synktatické složky (**c-structure**) - používá stromy.

GPSG (Generalised Phrase Structured Grammar)

Vznik: konec 70. let minulého století

Autoři: Gerald Gazdar (dále také Ivan Sag, Geoffrey Pullum)

HPSG (Head-driven phrase structured grammar)

Sémantika

Sémantika je nauka o významu jednotlivých slov, morfémů a jiných znaků, případně též o jejich vztahu ke skutečnosti, kterou označují.

Fregova koncepce (funkcionální koncepce sémantiky)

Cílem Fregovy koncepce je znázornění rozdílu mezi **významem** a **pravdivostní hodnotou**.

Příklad: Výroky "Jan Švejnar je prezidentem." a "Václav Klaus je prezidentem" mají velmi blízký význam (mají stejný predikát "být prezidentem", liší se pouze subjekt). Pravdivostní hodnoty jsou ovšem různé - druhý výrok platí, kdežto první ne. Ovšem (z definice) by mohl existovat jiný svět, kde by Jan Švejnar byl prezidentem a Václav Klaus ne, tedy pravdivostní hodnoty by byly obráceny, ale význam by samozřejmě zůstal beze změn.

WordNet

WordNet je **sémantický slovník** anglického jazyka. Jeho vývoj byl započat v roce 1985 na Princeton University.

WordNet rozděluje slova do množin synonymů (tzv. *synsets*), obsahuje stručné definice slov a k tomu i informace o sémantických vztazích mezi jednotlivými množinami. Byl vytvořen mj. za účelem automatické analýzy textů a pro podporu aplikací pracujících s umělou inteligencí.

Korpusy

Textový korpus je strukturovaný soubor digitálně uložených textů. Většinou je označkovaný (anotovaný) na základě předchozí morfologické analýzy.

Brown corpus of standart American English

První moderní elektronický korpus. Byl vytvořen na Brown University v Providence. Autory jsou W. N. Francis a H. Kučera.

- 1 milion slov textů v americké angličtině vytištěných v roce 1961
- 15 druhů textů, 500 textů (každý po cca 2000 slov) - v různých kategoriích různý počet textů:
  - novinové reportáže: 44 textů
  - humor: 9 textů
  - krásná literatura: 75 textů

PennTreebank

První a nejznámější synktaticky anotovaný korpus. Byl vytvořen na University of Pensylvania.

- Obsahuje cca 1 milion slov.
- Obsahuje 2499 článků ze souborů 98 732 článků z Wall Street Journal nasbíraných během 3 let (na přelomu 80. a 90. let) - jeho obsahem je proto spíše "burzovní" angličtina.

Český Národní Korpus (Czech National Corpus)

Akademický projekt zaměřený na budování rozsáhlého počítačového korpusu především psané češtiny. Je vytvářen společným úsilím UK Praha, MU Brno a ÚČJ AV ČR.

- Je anotován na morfematické úrovni.
- V současnosti obsahuje cca 500 miliónů slov, z toho 100 miliónů je dostupných jako SYN2000.
- SYN2000 je založen na:
  - 15% literárních textů
  - 60% novinových článků
  - 25% odborných textů

Morfologický analýzátor pro CNC:

- obsahuje přes 700 tis. lemat
- rozpoznává přes 15 mil. slovních tvarů
- používá poziční značkovací systém
  - každá značka má 15 pozic (znaků), průměrně je použito 4,29 značek na slovo
- využívá statistické modely

- automaticky se učí nová pravidla na základě kontextu (jako základ byl použit ručně anotovaný korpus obsahující cca 1,2 mil. slov)
- přesnost je cca 94%

Pražský Závislostní Korpus (The Prague Dependency Treebank)

Projekt ÚFAL-u MFF UK Praha. Podrobný značkovací model použitelný pro různé typy jazyků.

Cíl: vytvoření víceurovňového analyzátoru pro morfologickou, analytickou a tektogramatickou rovinu.

Jako teoretický základ byl zvolen Funkční generativní popis (FGD) od P. Sgalla.

Roviny anotace:

1. Morfologická - cca 2 mil. slovních jednotek
2. Analytická (povrchově synktatická) - cca 1,5 mil. slovních jednotek
3. Tektogramatická - 0,8 mil. slovních jednotek

N-gramy, n-gramové modely

N-gram (obecně) je n prvková podposloupnost dané posloupnosti. V lingvistice jsou prvky posloupnosti slova, posloupností je věta nebo celý text.

N-gramový model je pravděpodobnostní model, kde pro n-prvkový řetězec slov W je pravděpodobnost jeho výskytu P(W) definována jako:

$$P(W) = P(w_n | w_{n-1} \dots w_1) * P(w_{n-1} | w_{n-2} \dots w_1) * \dots * P(w_1)$$

Nejčastěji se používají n-gramové modely pro n = 1, 2, 3 nebo 4.

Vyhlazování

I použití n-gramového modelu je velmi náročné na vypočetní kapacity. Pokud máme slovník V o 40000 slovech, pak pro n = 3 je velikost modelu $|V|^3 = 6{,}4 \times 10^{13}$. Typická velikost trénovacích dat je v řádu 100 mil. slov. V modelu je proto příliš mnoho nulových pravděpodobností, ale některé z nich zastupují existující kombinace.

Vyhlazování je pokus o řešení této situace, které spočívá v nahrazení nulových pravděpodobností nějakou velmi malou hodnotou.

Strojový překlad (MT - machine translation)

Transfer - přenos zanalyzované věty z jednoho jazyka do druhého (slovosled, morfologie). K překladu je použité esperanto nebo interlingua (umělé jazyky).

ALPAC (American Language Processing Advisory Committee) vydal v 1969 r. zprávu, jejímž (nechtěným) důsledkem bylo zastavení vývoje strojového překladu v USA.

Historie

TAUM - METEO (1976)

První komerčně úspěšný sýstem, překládal meteorologické zprávy z angličtiny do francouzštiny (dobře definovaná a výrazně synktaticky i sémanticky omezená podmnožina jazyka). Systém sám rozpoznal text, jehož překlad mu dělal potíže a předal jej ke zpracování lidskému překladateli. Používal se až do 90. let minulého století.

SYSTRAN

Systém pro překlad dokumentů EU. Přímy překlad mezi cca 20 páry jazyků, ale uspokojivé kvality dosahoval pouze u nejstarších párů (A-F-N). Data byla oddělena od programu.

EUROTRA

Oficiální projekt EU v 80. letech. Byl megalomanský (72 párů jazyků). Nepovedl se z hlediska modularity.

VERBMOBIL

Německý nástupce systému EUROTRA, sloužil pro překlad mluvené řeči a jeho použití bylo tématicky omezené - domlouvání schůzek mezi obchodními partnery.

Současné trendy

1. Statistické metody
   - využívají paralelní označkované korpusy
   - pro měření kvality používají technické metody založené na referenčních překladech - tzv. **BLEU score**
   - více níže
2. Nástroje podporující překlad
   - systémy využívající již dříve přeložených textů - princip tzv. **překladové paměťi**

České systémy pro strojový překlad

APAČ (80. léta)

Systém pro překlad z angličtiny do češtiny, byl založen na stejném formalismu jako systém METEO. Byl zaměřen na překlady z oblasti textů o vodních pumpách.

Používal **transdukční slovník** (s cca 1500 výrazy):

- -ation => -ace (industrialization => industrializace)
- -ic => -ický (static => statický)

Ruslan (1985-1990)

Byl zaměřen na překlad manuálů k sálovým počítačům z češtiny do ruštiny. Také využíval transdukční slovník (cca 8500 slov). Gramatika byla zapsána pomocí Q-systémů.

Česílko (1998 - ?)

Je výsledkem snahy o vytvoření systému pro překlad mezi blízkými jazyky s minimálním lidským podílem na překladu. Zdrojový text musí být nejdříve (člověkem) přeložen z angličtiny (resp. jiného jazyka) do češtiny a až poté je strojově zpracováván a překládán do dalších (slovanských) jazyků.

- FAHQ (vysoce kvalitní, plně automatický překlad)
- jsou použity plné morfologické slovníky
- statistická analýza češtiny

PC Translator 2003

Český komerční systém.

Statistický strojový překlad

Prvotní idea je spojována se jménem Warrena Weavera (1949).

Základní myšlenkou je použití paralelních (bilinguálních) korpusů jako tréninkové sady různých překladů.

Noisy Channel Model

Skládá se ze dvou komponent:

1. *P(e)* - jazykový model (the language model)
   - pravděpodobnosti výskytu jednotlivých vět v jazyce
   - může jím být např. trigramový model vypočítaný na základě libovolných dat
2. *P(f|e)* - překladový model (the translation model)
   - pravděpodobnosti překladu jednotlivých vět z jazyka *e* na jednotlivé věty z jazyka *f*
   - je vypočítán na základě paralelního korpusu

Pozn.: Je nutné si uvědomit, že různé věty můžou nést identickou informaci (*Mám hlad.* vs *Jsem hladový.*), ale nemusí si být rovnocenné z hlediska četnosti použiti v daném jazyce.

BLEU score

BLEU skóre je metrika, která určuje "věrnost" statistického strojového překladu. Je nutné si uvědomit, že je to spíše metrika určená pro porovnání dvou překladů stejného textu, ne pro absolutní kvalitativní řazení různých překladů, protože se vždy počítá vzhledem k určitým referenčním překladům.