Nedostatky bezkontextové gramatiky

- Běžná gramatika nezachytí schodu podmětu s přísudkem.
- Lze řešit přidáním zvláštních neterminálů pro jednotné číslo, množné číslo... → Velký nárust počtu neterminálů

Rozšířené přechodové sítě

- ATN - Augmented Transition Networks

Gramatika:

S → NP VP
NP → Det N
VP → V [NP]

Síť:

→ S_0 → seek NP(Prohledání grafu NP) → S_1 → seek VP →
→ NP_0 → cat Det(vyzvednutí členu ze slovníku) → NP_1 → cat N → NP_2 →
→ VP_0 → cat V → VP_1 → seek NP/jump → VP_2 →

Systémy Q

- Použit mechanizmus grafového analyzátoru
  - Slova tvoří ohodnocení hran.
  - Postupně tvoříme hrany ohodnecené levou stranou pravidel z gramatiky. Hrany které byli překlenuty označíme tečkou.
  - Opakujeme, dokud existují pravidla.
  - Následuje fáze čištění:
    1. Odstraníme všechny hrany odstraněné tečkou.
    2. Ponecháme pouze hranu vedoucí z počátku do konce.
    3. Odstraníme paralelní hrany se stejným ohodnocením.
  - Analýzu celé věty můžeme rozložit do několika fází.
- Nevýhodou může být složitý zápis gramatiky

Unifikační gramatiky

- speciální datový typ - Sestava rysů(Feature Structure FS)

$$\begin{bmatrix} \text{věta:} & \text{oznamovací} \\ \text{podmět:} & \begin{bmatrix} \text{slovní druh:} & \text{N} \\ \text{číslo:} & \text{S} \\ \text{rod:} & \text{F} \end{bmatrix} \\ \text{přísudek:} & \begin{bmatrix} \text{slovní druh:} & \text{V} \\ \text{číslo:} & \text{S} \\ \text{rod:} & \text{F} \end{bmatrix} \end{bmatrix}$$

Nezachycuje shodu, proto vylepšíme:

$$\begin{bmatrix} \text{věta:} & \text{oznamovací} \\ \text{podmět:} & \begin{bmatrix} \text{slovní druh:} & \text{N} \\ \text{shoda:} & 1 - \begin{bmatrix} \text{číslo:} & \text{S} \\ \text{rod:} & \text{F} \end{bmatrix} \end{bmatrix} \\ \text{přísudek:} & \begin{bmatrix} \text{slovní druh:} & \text{V} \\ \text{shoda:} & 1 \end{bmatrix} \end{bmatrix}$$

- Unifikace:

$$\begin{bmatrix} \text{číslo:} & \emptyset \\ \text{rod:} & \text{F} \end{bmatrix} \cup \begin{bmatrix} \text{číslo:} & \text{S} \\ \text{rod:} & \emptyset \\ \text{pád:} & 7 \end{bmatrix} = \begin{bmatrix} \text{číslo:} & \text{S} \\ \text{rod:} & \text{F} \\ \text{pád:} & 7 \end{bmatrix}$$

- Lze kombinovat syntaxi i sémantiku
- Velmi průhledné a obecné(lze použít i na morfologické, syntaktické a sémantické úrovni)
- Při špatném návrhu struktury můžeme skončit u slovesa s pádem nebo podobných věcí. Lze zlepšit zavedením typů.

Automatické rozpoznávání cizích slov

Vlastnosti českého jazyka

- foném = grafém, fonologické písmo. Vyskytuje se jen u jazyků, do nichž to bylo zavedeno reformou.
- Slovenština - rytmický zákon - po dlouhé slabice následje krátká
- Obojetné souhlásky jsou původem měkké
- Problémy - napětí, napnul, napjetí
- Vývoj

  - stahování - bojati se → bát se, ale zůstalo zachováno bojím se
  - depalatizace - odstraňování měkčidel
  - diftongizace - ú → ou
  - změny hlásek - slovanské g → h, ó → uo → ů

Hledání cizích slov

- Jiná seskupení písmen, nerespektují pravidla češtiny
- Cizí grafémy
  - x,q,w
  - f (kromě "foukat", "doufat", "zoufat")
  - g - původní g se změnilo v h
  - ó - české přešlo v ů
- Až na vyjímky nezačínají slova na samohlásky
- Existuje pouze jedna dvouhláska - ou, zbytek jsou cizí slova, slováci mají dvouhlásek víc.
- Nosovky - rekomando, konstanta, ...
- Při přejímání slov se redukují zdvojené souhlásky
- ph → f
- Můžeme používat cizích předpon a přípon
- Tvrdá samohláska se po měkké souhlásce mění na měkkou samohlásku.
- Tvrdá souhláska se před měkkou samohláskou měnína měkkou souhlásku.
  - *é* je tvrdé, *ě* je měkké, ale *e* je obojetné
- metateze likvid - ((E—O)R)—((E—O)L)→(R(E—A))—(L(E—A))
  - ert → ret, Elbe → Labe
- Slova končící na *ý* jsou přídavná jména - vyjímky prý, úterý, čehý
- *-ost* - většinou podstatné jméno odvozené zpřídavného
- Nerozlučné skupiny souhlásek
  - zd/st, žd/št
- Polosamohlásky - l/r

Automatický překlad

Motivační příklad

Budeme překládat větu:

Kageru-to mugade hala jo deka medsene seno gejay!

kageru	sejít se, shromáždit se
mugade	hlava
hala	skupina
jo	jeden, jedna
deka	pět
medse	hodina
seno	velký
jay	muž

Sejít se hlava skupina jeden pět hodina velký (?muž?)

Chybí nám morfologie

≤sloveso≥+to	budoucí čas
přípona -ne	časová předložka v
předpona ge	místní určení

Sejde se hlava skupina v jeden pět hodina velký u muž.

Pravidla domorodého pravopisu

- V oznamovacích větách pevný pořádek slov
- Systém číslovek se chová jako římský
- Přídavná jména rozvíjejí následují větný člen

Sejde se hlava skupina ve čtyři hodiny u velkého muže.

Ustálená spojení

mugade hala	rada starších
seno jay	náčelník

Rada starších se sejde ve čtyři hodiny u náčelníka.

Další problém - hodiny se počítají od začátku dne.

Potřebujeme tedy ještě vědět, jak to v patřičném prostředí chodí.

Překlad může záviset na situaci.

Historie

Zdrojový text → analýza → transfer(překlad, strukturální změny ve stromě, ...) → generování → cílový text

Ideální by bylo vynechat transfer a provádět analýzu až do interliguy a odtud přímo generovat.

- 1946 - A. D. Booth - Slovníkový překlad slovo od slova.
- 1948 - R. M. Richens - pracuje s morfémy.
- 1950 E. Reifler - zavádí pre- a post- editing
  - controled language - preediting - člověk píše jednoduše s ohledem na překlad
- 1952 - První konference o strojovém překladu na MIT
  - L. E. Doster - pivotní jazyk - přirozený jazyk hrající roli interlinguy(ta byla formálním jazykem)
- 7. 1. 1954 - Georgetownský experiment
  - 45 vět, 250 slov, 6 syntaktických "zákonů" ...
- 1956 - První mezinárodní konference
- 1957 - N. Chomsky - Standart theory
- 1960 - Y. Bar Hillel - "Vysoce kvalitní plně automatický překlad nemůže být nikdy dosažen"
- Začína se hovořit o kategoriích

FAHQMT	plně automatický
HSMT	člověk pomáhá počítači
MAHT	počítač pomáhá člověku
	slovníky, translation memory(hledá podobné věty u nichž z
HT	překládá člověk

Současné trendy

Statistické metody

- Měření kvality pomocí referenčního překladu
- Využívají již přeložené texty

Nástroje podporující překlad

- Používají "překladovou paměť"
- Nabízejí překladateli to, jak to přeložil minule
- IBM Translation Manager, Deja Vu, TRADOS, SDLX

České systémy

APAČ

- Překlad z angličtiny do češtiny
- Vytvořen ve shodném formalismu jako METEO
- Transdukční slovník - odborná slova přejatá z cizích jazyků(řevctina nebo latina) do ENG i CZ se překládají dobře(-ation → -ace(industrialization → industrializace, -ic → -ický)

Ruslan

- Překlad z češtiny do ruštiny
- Nebyl dokončen
- Transdukční slovník
- Q-systémy
- Záchranná pravidla pro případ problémů při analýze

Systém Česílko

- Lokalizace velkých softwarový systémů
- Překlad z češtiny do slovenštiny a polštiny
- Použije se překladová paměť a český překlad se automaticky přeloží do slovenštiny a polštiny a vrátí se do paměti
- Rozdíly Češiny a Slovenštiny
  - Shodná syntaxe (problém - bude-li → ak bude)
  - 3 slovníky - Český a Slovenský morfologický slovník a překladový slovník
  - Většinou shodné pořadí ve ětě
  - Naprosto odlišné tvarosloví
  - Používá se tagger - určuje v jakém kontextu se slovo vyskytuje
  - Možno využívat dalží slovníky

PC Traslator

Problémy

- Rozpoznání vecí, které se nepřekládají - Rozpoznávání pojmenovaných entit - Named-entity recognizer
- Zachovává anglický pořádek slov ve větě

Korpusová lingvistika

Nejcennější je to, že korpus je označkovaný. Nejčastěji morfologická alanlýza, ale je dobré vědět i pády, rody, čísla a podobně. Důležitým rozhodnutím je, co do korpusu dávat.

Značkování

- Základem je morfoligická analýza
- Další nejdůležitější je syntaxtické označkování(může pomoct při ověřování hypotéz ohledně jazykových pravidel)

Historie

Brownův korpus

- W. N. Francis a H. Kučera na Brown University
- 1 milion slov textů v americké angličtině vytištěných v roce 1961
- 15 druhů textu, 500 textů, každý asi kolem 2000 slov
- Snaha o získání reprezentativního vzorku(dnes spíše snaha o získání co největšího množství dat)
- Označkován pouze morfologicky

PennTreebank

- Články z Wall Street Journal
- Syntakticky značkovaný korpus
- Závorkovací systém.

Český Národní Korpus

- Spolupráce UK, MU a Ústavu pro jazyk Český
- Sbírá text od začátku devadesátých let - zpočátku hlavně novinové texty
- Morfologicky značkováno
- V současnosti 500milionů slov, 100milionů slov dostupných veřejnosti
- Neznačkuje se ručně, používá se tagger - je potřeba výsledky pročistit
- Jsou vytvářena gramatická pravidla pomáhající odhalit chyby způsobené taggerem

- Složení - 60% - novinové texty, 15% literatura(11% fikce), 25% technické a specializované texty
- Průměrně 4.29 značek na slovo
- Průměrná přestnost taggeru - 95%

Pražský závislostní korpus

- Inspirován Penn Treebankem, ale mnohem složitější
- Snaha vybudovat značky na několika úrovních
  - morfologickái
  - syntaktická -analytická
    * skoro povrchová - přiřazení analitické funkce
    * definováno pomocí příkladů
    * tvoří strom - uzly jsou funkční jednotky(slova, interpunkční znaménka, ...)
    * každá věta má jeden technický uzel
  - tectogrammatická(hloubková)
    * Musí tam být vše podstatné(doplnění nevyjádřenéhopodmětu)
    * Oproti předchozí úrovni něco ubyde(např. předložky určující pád a občas i něco přibyde(povinné aktanty(slova spadající do valenčního rámce slovesa)
    * Už zbudování struktury a rohodnutí která slova se budou značkovat může být záležet na úhlu pohledu
    * Čtyři podstruktury
      1. závislostní struktura, funktory
      2. topic/focus - jádro a ohnisko věty a slovosled
      3. koreference - co zastupují zájmena, kontext
      4. zbytek
    * Věty chápányv kontextu
- Během anotace byli vyvynuty i manuály pro anotaci(potřeba sjednotit anotátory
- Podmnožina Českáho národního korpusu(novinové texty)

Pražský Arabský závislostní slovník

- Na araby se dobře scháněj peníze ;-)
- Ukázka, že tectogrammatická rovina je dobrá pro všechny roviny, lépe se bude překládat na tectogrammatické úrovni, otázka je ovšem náročnost dostání se na tectogrammatickou úroveň

PCEDT

- Anotován automaticky
- Česko-Anglický
- Data z Wall Street Journal

Anotační pomůcky

- NetGraph
- TrEd