Jazykové technologie

Karel Oliva
Ústav pro jazyk český
Akademie věd ČR

Změněná situace jazyka jakožto komunikačního prostředku

- komunikace mezi lidmi (navzájem)

- komunikace mezi lidmi a stroji

  => změna schopnosti užívání jazyka

  => cíl jazykových technologií

Cíle jazykových technologií

produkty, které využívají přirozený jazyk (a znalosti o něm):

- **interakce člověk-počítač (povely, výsledky)**
- **strojový překlad, překlad s podporou počítače**
- **vyhledávání informací (plné texty, databáze, Internet) a rešeršní služby**
- **podpora kreativity (př.: korektor pravopisu)**
- **...**

Příklady:

- hlášení v metru/tramvajích/autobusech …
- ovládání servisních zařízení

  (např. v automobilu)
- předčítání psaného textu

  (pomůcka pro nevidomé)
- rozpoznávání rukopisu
- přepis mluveného slova (“diktát”)
- prediktivní psaní (T9, iTap, …)

Základy jazykových technologií

- **Aplikace matematické lingvistiky (teorie)**
  - **studium jazyka matematickými prostředky**
- **Aplikace komputační lingvistiky**
  - **pomezní věda: CompSci, AI, jazykověda, kognitivní psychologie a psycholingvistika, zpracování akustického signálu**
- **Validace na korpusech**

Potřeba "inteligentních" nástrojů

- **Příklad: "korektor pravopisu"**

  Psi štěkaly. **OK**

- **Příklad: vyhledávání na *www.atlas.cz***

  Vyhledání odkazů na dokumenty obsahujících slovo "maso":

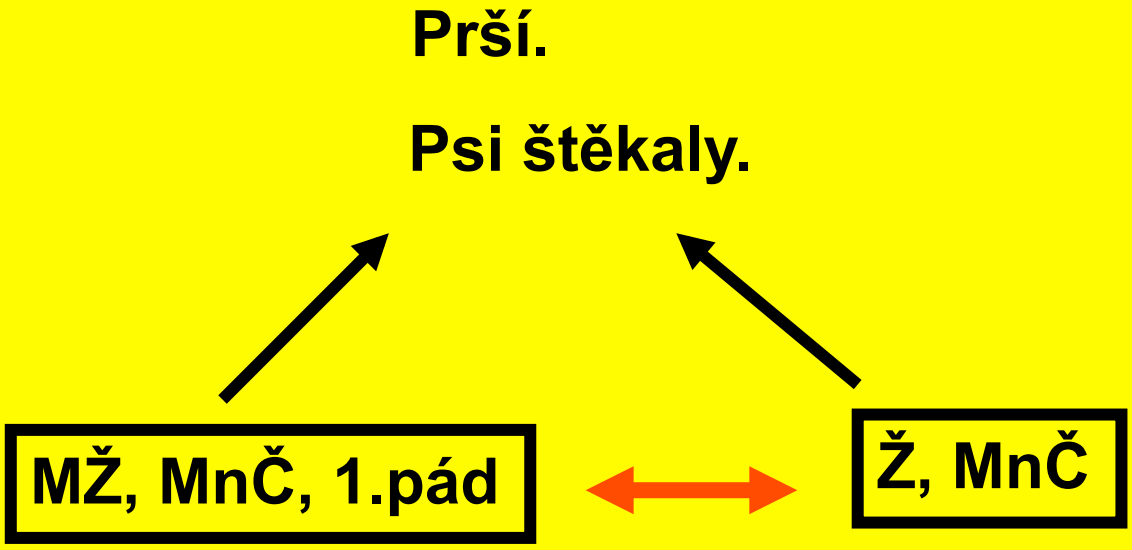
  Olympus útočí na masy s mini-mju digital.

Příklad aplikace:
Softwarový nástroj
„Kontrola české gramatiky“

- **Segmentace textu na slova, čísla, interpunkci**
  …. a na věty
- **Analýza správnosti věty**
- **Vyhledání a ohlášení chyby**
- **Návrh opravy**

Chyba, minimální chybová konfigurace

- **Chyba („definice“): porušení jazykového pravidla**
- **Minimální chybová konfigurace**:



Problém 1:
reálné konfigurace nejsou minimální

- **Psi štěkaly.**
- **Psi** včera **štěkaly.**
- **Psi** včera na dvoře **štěkaly.**
- **Psi** včera na dvoře dlouho **štěkaly.**
- …
- Naši staří **psi** včera na dvoře dlouho **štěkaly** na kočku.

Řešení:

- **teorie „minimálních nesprávných vět“**
- **teorie „rozšíření minimálních nesprávných vět“**

Problém 2: některé neminimální konfigurace nejsou negramatické

Při rozšiřování negramatické konstrukce nemůžu přidat „cokoliv, co mě napadne“

<u>Řešení</u>:

• lingvistická teorie „vždy nesprávných vět“

Problém 3:
lexikální a tvarová homonymie češtiny

- **Jeho čelisti klapaly naprázdno.**
- **Jeho čelisti klapali naprázdno.**
- **Jeho čelisti vycházeli po koncertě ze sálu.**

Problém 3:
lexikální a tvarová homonymie češtiny

- Jeho ~~čelisti~~ klapaly naprázdno.
- Jeho ~~čelisti~~ klapali naprázdno.
- Jeho ~~čelisti~~ vycházeli po koncertě ze sálu.

Problém 3:
lexikální a tvarová homonymie češtiny

- **Jeho dásně klapaly naprázdno.** ☺
- **Jeho dásně klapali naprázdno.** ☹
- **Jeho dásně vycházely po koncertě ze sálu.** ☺

Hráli si tam nějaké hezké dívky?
Hráli si tam nějaké hezké hry? ***(1./4. pád, MnČ, Ž)***

<u>Řešení:</u>

- **Vytvoření přehledu homonymie v češtině**
- **Vytvoření metodiky pro odstraňování homonymie**

Odstraňování homonymie

bigram **~ uspořádaná dvojice (abstraktních) slov**

Příklad: *[Adjective, Noun]* je bigram

negramatický (nemožný) bigram

~ bigram, který nelze najít v žádné gramaticky správně utvořené větě

Příklady negramatických bigramů

- "*bych/bys/ ...by/abych/.../kdybych/...„* + V-praes

  ... aby vinou(~~V-praes~~, Noun) ....

- "*a/ale*" + reflexivní příklonka

  ——— *... a se(Refl, Prep) ...*

- "*lze/nelze*" + V-fin(≠ "*(ne)bylo/(ne)bude„*)

  ... ~~lze~~ jí(Pron, V-fin) ...

Generalizace na rozsáhlejší konfigurace

- ***Trigramy, tetragramy, pentagramy***, ...
  důležitá "ideální" slova (symboly)
  ZAČÁTEK_VĚTY a **KONEC_VĚTY**

Příklad:

ZAČÁTEK_VĚTY

+ N-nom

+ V-fin(≠ "*být/bývat/slout*")

+ N-nom

+ **KONEC_VĚTY**

Ústava-nom *zaručuje* *rovnost*-~~nom~~/acc

Generalizace na rozsáhlejší konfigurace

- **Konfigurace *neomezené* délky**

Základní postřeh:

konfigurace pevné délky lze rozvolnit přidáním materiálu mezi členy konfigurace (při zachování negramatičnosti)

Příklad: **Prep + V-fin**
Prep + Adv + V-fin
Prep + Adv + Adv + V-fin
Prep + Adv + Adv + Adv + V-fin
...

Hledání chyb (ČNK)

Negramatický bigram: ***Prep + Vfin***

- Před bránou … připoutali vojáci Květuši
  k zábradlí **před pokladou**, a kdyby ...
- Budiž pokoj **v přehradí** Tvém a upokojení
  na paláci Tvých (Bible kral. 1613)
- … před mnoha lety v Ťienově zahradě
  v Su - Čou ...

Problém 4: návrh opravy konkrétní chyby

Psi pokousaly hyeny.

Má se opravit na

- **Psi pokousali hyeny. (chyba ve slovesném tvaru)**

nebo na

- **Psy pokousaly hyeny. (chyba ve tvaru podst. jména)**

Řešení:

- **„váhy“ minimálních nesprávných vět**

Vývoj systému – přehled:

- Obecná lingvistická teorie „nesprávných vět“
- Obecná lingvistická teorie odstraňování homonymie
- Popis rozsáhlé oblasti „nesprávných vět češtiny“ („ne-gramatika češtiny“)
- Popis odstraňování homonymie v češtině
- Empirická studie frekvence chyb
- Efektivní softwarová podpora:
  - Specializovaný programovací jazyk
  - Implementace

Řešení problému s *mini-mjú digital*

Olympus útočí na masy s mini-mju digital.

Předložka *NA* se pojí se 4. nebo 6. pádem

Tvary podst. jm. *MASO* ve 4. nebo 6. pádě:

maso, mase, masu, masa, masech

Atraktivita oboru

- **Kombinace inovativního výzkumu**
  v "klasické humanitní" disciplíně
  s "matematickou přesností"
- **"Cutting edge research"**
  - špička vyhledávacího výzkumu
- **Perspektiva vývoje oboru**
- **Užitečné aplikace**