Výzkum na ÚFALu

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Seminář TMC (CESNET)

Osnova

- Počítačová lingvistika, formální popis jazyka.
- Lingvistické problémy.
- Lingvistické nástroje a data.
- Lingvistické aplikace.
- Široká škála přístupů.
- Náměty ke spolupráci.

Na hranici oborů...

čeština,
angličtina, němčina ...

matematika

počítače

Počítačová lingvistika

Formální popis češtiny

zákony

udělejte

pro

lidi

Formální popis češtiny

Morfologická rovina:

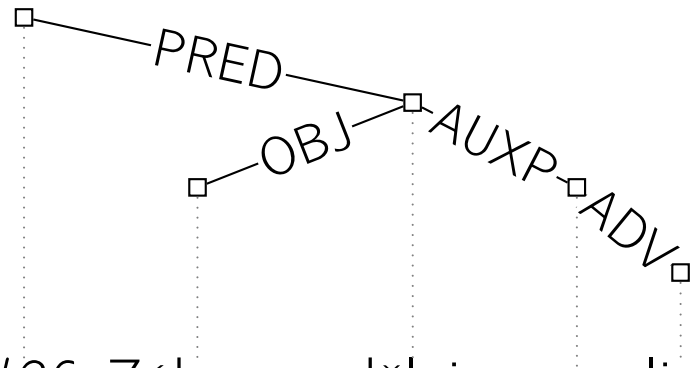
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Formální popis češtiny

Morfologická rovina:

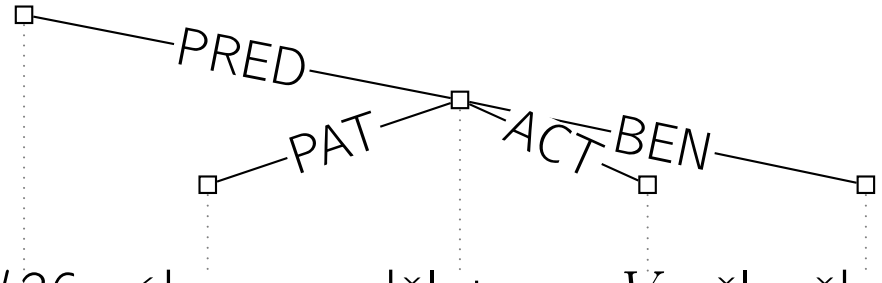
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Analytická rovina (povrchová syntax):



#36 Zákony udělejte pro lidi

Tektogramatická rovina (hloubková syntax):



#36 zákon$_{Pl}$ udělat$_{imp}$ Vy člověk$_{Pl,pro}$

Lingvistické problémy

- Víceznačnost a význam slov.

  Spal celou Petkevičovu přednášku. Ženu holí stroj.

- Bohatost slovních tvarů.
  - Čeština má 7 pádů, 3 čísla a 4 rody.
  - V angličtině se používá ∼50 morfologických značek, v češtině 4000.

- Koordinace a apozice:

  Předseda vlády, Petr Nečas , a Martin Lhota přednesli příspěvky o...

Vstup	We have both countries inside and outside the Eurozone.
Reference	Máme tu země eurozóny a země stojící mimo eurozónu.
Hypotéza	Máme obě země uvnitř a vně eurozóny.

- Zájmena.

Ukázka bohatosti jazyka

Kolik je správných překladů následující věty?

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

Ukázka bohatosti jazyka

Příklady ze 71 tisíc správných překladů anglické věty:

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdobná.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podobná.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

Lingvistické nástroje

- Identifikace kódování dokumentu a jazyka.
- Rozpoznání hranic vět a slov:

  Švejk 12. prosince dorazil na král. Vinohrady s dopisem.

  ajskrím → I scream / icecream.
- Morfologická analýza.
- Povrchový a hloubkový větný rozbor.
- Identifikace pojmenovaných entit:

  Bílý dům se nechal slyšet.

  Rice University ≠ univerzita rýže
- Koreference (mj. identifikace, co zastupují zájmena).

Lingvistická data

- **Korpusy** jsou (velké) sbírky textů:
  - Texty typicky označkované nebo včetně větných rozborů.
    Pražský závislostní korpus (PDT): 1.5 mil. slov.
    Pražský čj-aj závislostní korpus (PCEDT): 50 tis. vět.
  - Některé vícejazyčné: CzEng (15 mil. vět, 220 mil. slov, odpovídá ∼50 metrům knih, ty tvoří však jen čtvrtinu).
- **Slovníky** na ÚFALu jsou strojově čitelné:
  - Morfologický slovník říká, že *kočka* je české slovo a *kočke* ne.
  - Valenční slovník říká, že:

    Rodiče přijali Petra. → je správně

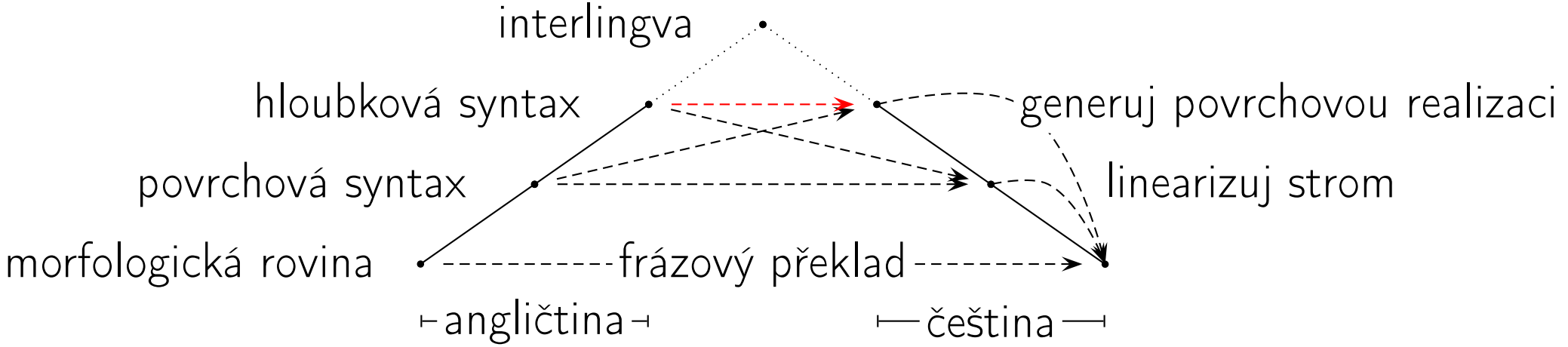
    Rodiče přijeli Petra. → není správně
  - Slovník subjektivity obsahuje hodnotící výrazy.

Lingvistické aplikace

	Děláme	
	Dobře	Aktuálně
Kontrola překlepů	**	*
Kontrola pravopisu	**	
Vyhledávání dokumentů	partneři	Khresmoi
Dolování informací z textu	*, partneři	Khresmoi
Automatická sumarizace textů	partneři	Khresmoi
Syntéza a rozpoznávání mluvené řeči	*	**
Dialogové systémy	*	Vystadial
Strojový překlad	**	MosesCore, Faust
Strojový překlad mluvené řeči	*	AMALACH
Analýza smýšlení (sentiment)	*	**

Široká škála přístupů

Jako příklad poslouží úloha strojového překladu:



- Čím víc vstup rozeberu, tím snazší by měla být fáze transferu.
- Hypotetická interlingva zachycuje čistý význam.

- Statistické systémy se natrénují "samy" podle ukázek.
- Pravidlové systémy ručně píší lingvisté-programátoři.

Nejlepší je přístupy kombinovat.

Střípky ke spolupráci

Hotovo, stačí spustit:

- „Zahuštění“ textových dat:
  - Převedení slov na základní tvary.
  - Identifikace plnovýznamových slov, odstranění pomocných.
    ⇒ Snazší detekce klíčových slov, tématu textu.
- Rozbor vět: podmět, přísudek, předmět. Vyřešení zájmen.
- Detekce pojmenovaných entit (*U tří slunců*), kolokací (*bílý kůň*).
- Shlukování dokumentů (politika, sport, e-mail, ...).

Lákavý výzkum (ztřeštěné nápady):

- Monitoring relevantních sociálních sítí (např. soom.cz).
- Identifikace pisatele textu.

Odkazy

http://ufal.mff.cuni.cz/

→ Research → Prague Czech-English Dependency Treebank 2.0

→ Data: Ukázkové české a anglické povrchové i hloubkové rozbory

→ Video Recordings

LINDAT: Repozitář lingvistických dat, možnost „anotací na zakázku“:

http://ufal.mff.cuni.cz/lindat/

Ukázky strojového překladu:

http://ufal.mff.cuni.cz/tectomt/