Jak se dělá strojový překlad

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

Obsah prezentace

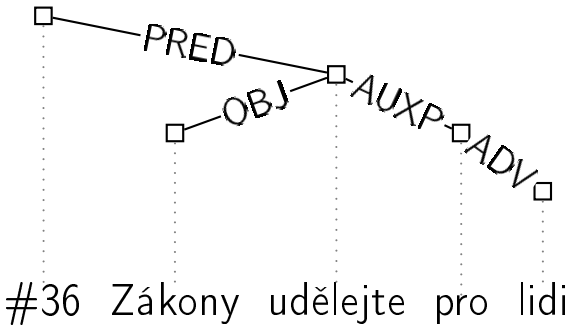
- Hlavní zaměření ÚFALu:
  - Formální popis přirozeného jazyka (čj, aj, arabština, ...),
  - Jazyková data (korpusy, slovníky),
  - Nástroje pro automatické zpracování jazyka.
- Dva přístupy ke strojovému překladu:
  - Frázový (mj. Google),
  - Stromečkový (mj. ÚFAL).
- Proč studovat na MFF (a ÚFALu).

Formální popis češtiny

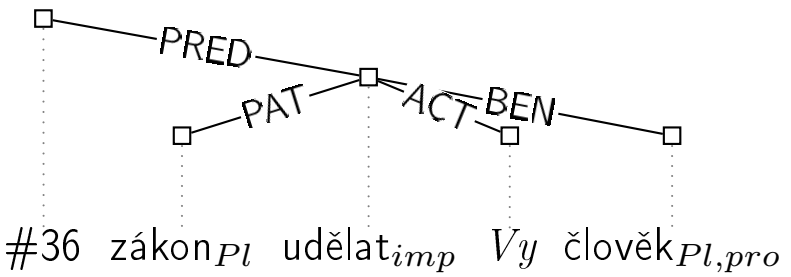
Morfologická rovina:

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):



Tektogramatická rovina (hloubková syntax):



Lingvistická data, nástroje a aplikace

Jazyková data

- **Korpusy** jsou (velké) sbírky textů:
  - Texty typicky označkované nebo včetně větných rozborů (1 mil. vět ručně).
  - Některé vícejazyčné: CzEng (7 mil. vět, 50 mil. slov, ∼1-2 tis. knih).
- **Slovníky** na ÚFALu jsou strojově čitelné:
  - Např. morfologický slovník říká, že *kočka* je české slovo a *kočke* ne.
  - Valenční slovník říká, že *rodiče přijali Petra* je správně a *rodiče přijeli Petra* není.

Nástroje pro práci s texty či nahrávkami

- Identifikace jazyka, rozpoznání hranic vět a slov, ...
- Automatická morfologická analýza a větný rozbor.
- Identifikace pojmenování, ...

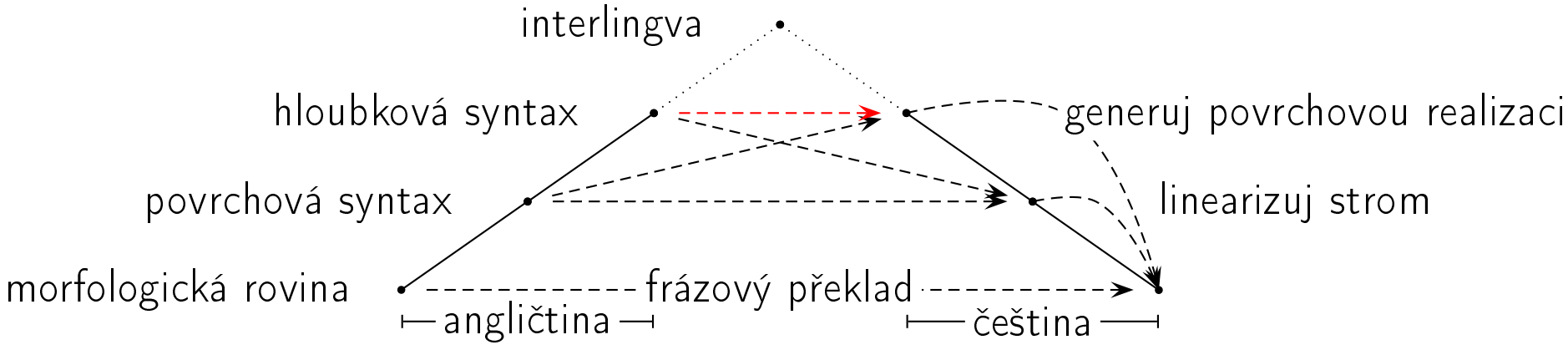
Aplikace: Vyhledávání na webu, syntéza/rozpoznávání řeči, sumarizace, strojový překlad...

Strojový překlad

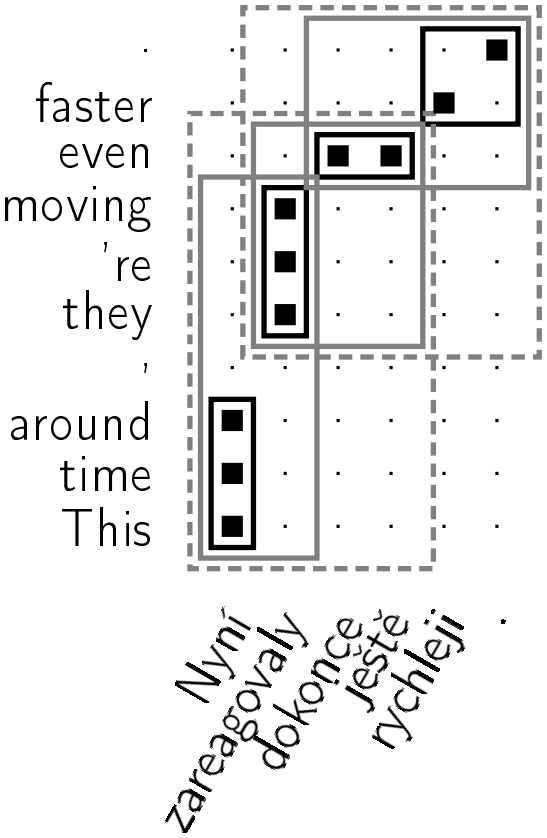
Strojový překlad zajímavý akademicky i komerčně:

- Hřiště pro testování užitečnosti mnoha dílčích nástrojů zpracování jazyka.
- EU utrácí ročně 1 000 000 000 eur za překlady.

Přístupy ke strojovému překladu:



Frázový překlad



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data:

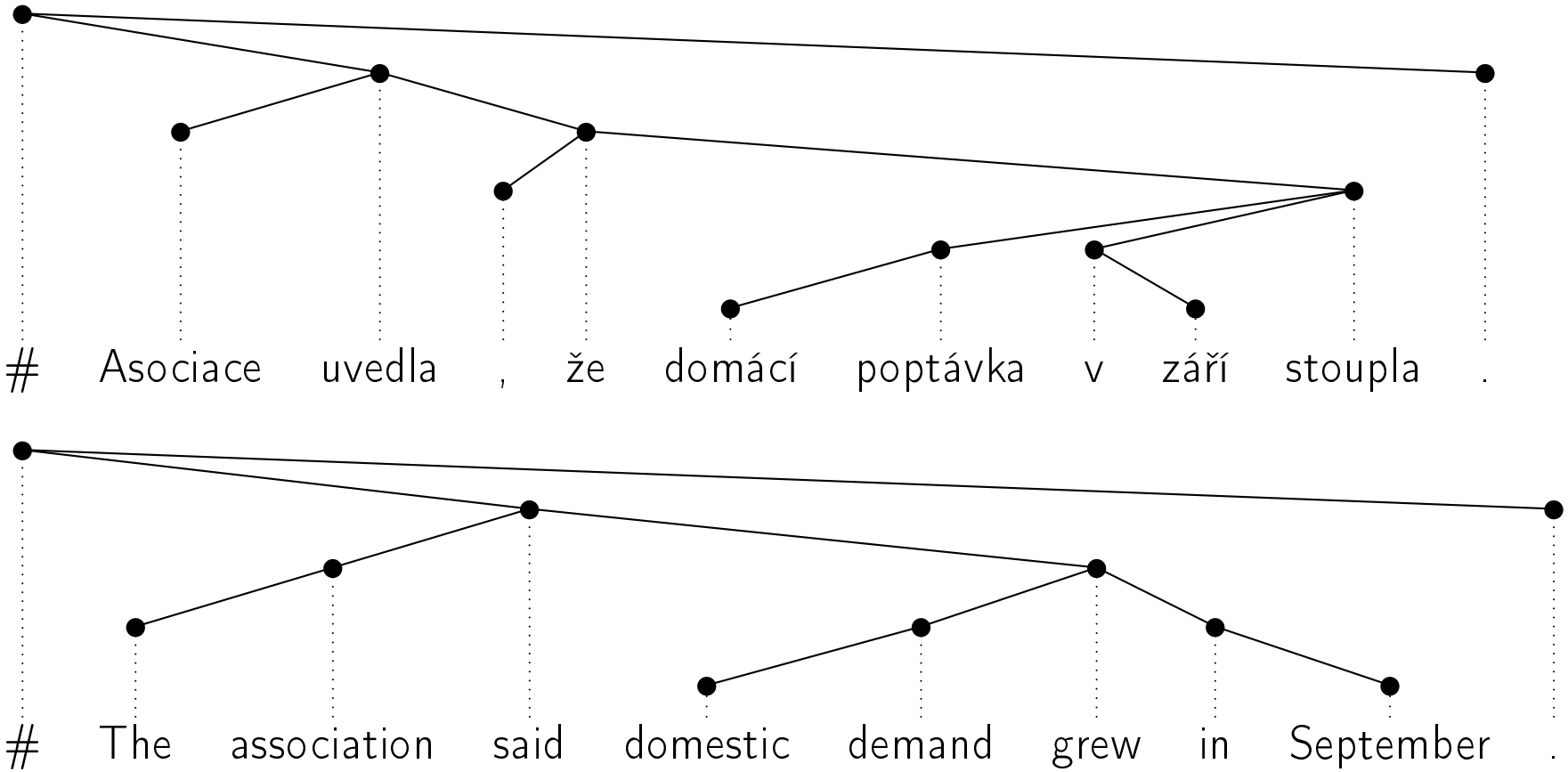
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo)

Při samotném překladu hledáme:

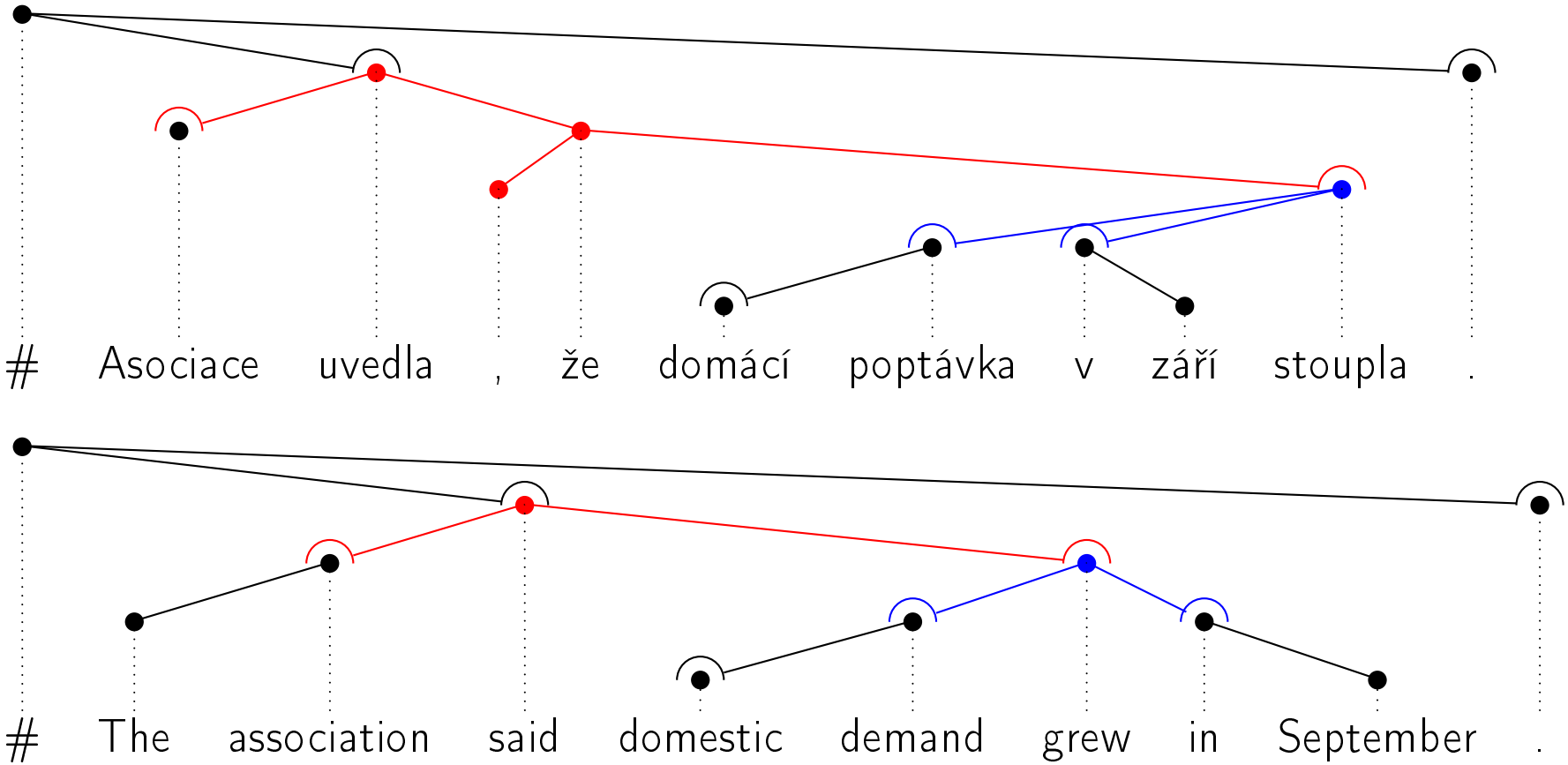
- takovou segmentaci vstupní věty na úseky („fráze“)
- a takové překlady frází

aby byl výstup co nejpravděpodobnější.

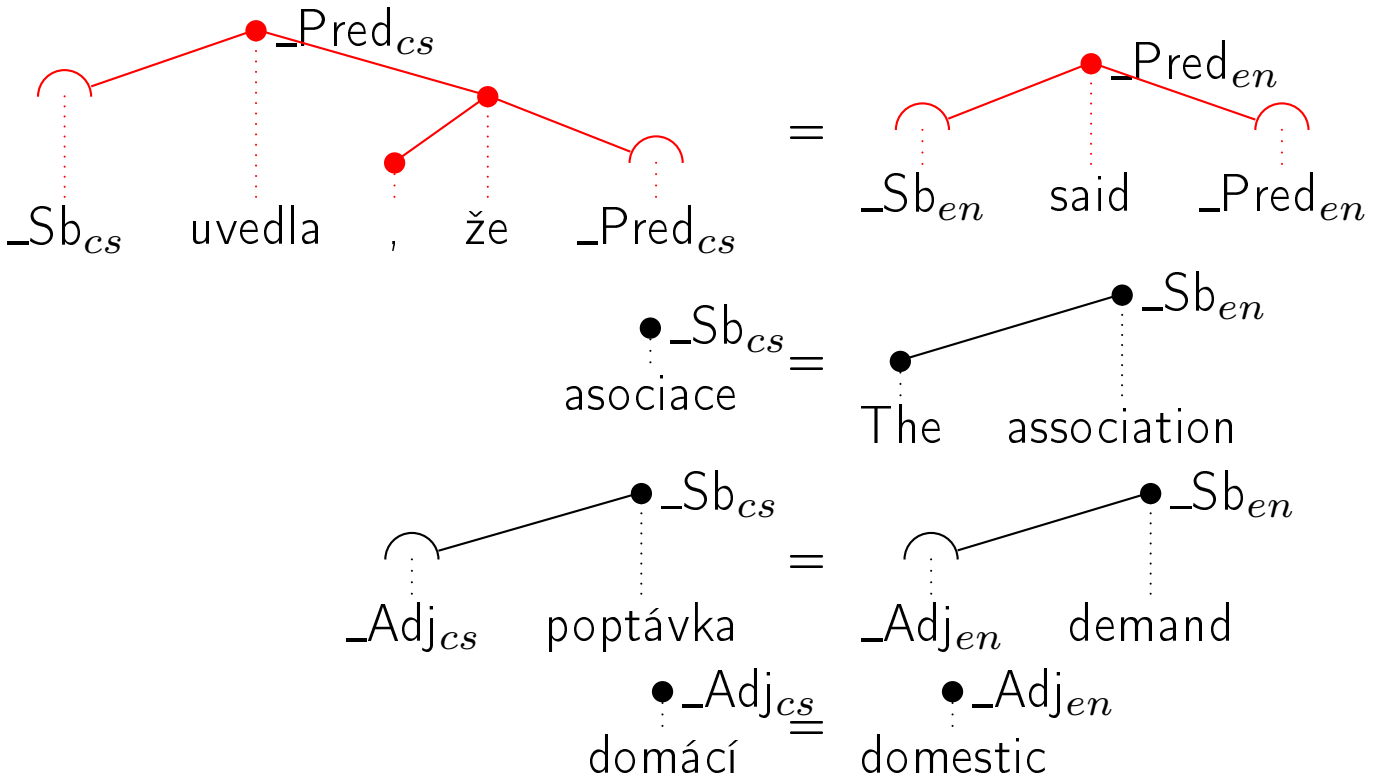
Syntaktický překlad: stromy...



...rozložíme na stromečky...



...a sebereme slovník překladů stromečků.



V čem je překlad těžký

- Významy slov (*bank* je *břeh* nebo *banka*, *plant* je *rostlina* nebo *továrna*).
- Správná volba slovního tvaru (v češtině mnoho možností).
- Pořadí slov (kombinatorická exploze variant, nestihneme projít všechny).
- Zájmena (a jejich rod – zastupuje anglické *it* něco muž/žen/střed. rodu?).
- Slovesné časy.
- Idiomy (*natáhnout bačkory* = *kick the bucket*).

Stručně: Pro danou větu:

- je těžké správně rozebrat („strojově pochopit“) vstup: *Ženu holí stroj*,
- je těžké získat překladový slovník, který by obsahoval všechno, co věta potřebuje,
- i s aktuálními slovníky je těžké v řadě možností poznat ty lepší.

Proč studovat na MFF (a ÚFALu)

Můžete se naučit mj.:

- Rozdělit složité úlohy na částečky a přispět částečkami,
- Počítat, abyste nehledali jehly v horách sena (Pravděpodobnost a statistika),
- Navrhovat datové struktury, abyste zvládli terabajty dat,
  Text na českém webu ∼1.5 TB, jeden experiment s frázovým překladem 1-2 GB ale třeba i 10 GB.
- Programovat, abyste zvládli stovky počítačů najednou,
  - Unix/Linux je naprosto nutný, Sítě a Internet velmi užitečné.
  - ÚFAL sám má 160 CPU, počítače s 16 až 32 GB RAM.
- Modelovat, jak lidé (myslí a) pracují s textem, řečí, gesty, ...
- Soutěžit na mezinárodní úrovni v překládání, analýzách, generování, ...

Více: `http://ufal.mff.cuni.cz/`
Ukázky: `http://ufal.mff.cuni.cz/` → Tools (→ překladový systém Moses)