Aplikace strojového překladu

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

Obsah prezentace

- Projekt IS „Od jazyka ke znalostem a sématickému webu“
- Úvod do strojového překladu:
  - Motivace
  - Hrubé rozdělení metod.
  - Formální popis přirozeného jazyka.
  - Obtížnost překladu.
- Dva přístupy ke strojovému překladu:
  - Frázový (mj. ÚFAL, Google).
  - Stromečkový (mj. ÚFAL).
- Závěrem: Přínos projektu Informační společnost pro pracoviště.

Od jazyka ke znalostem a sématickému webu (2005-2009)

Hlavní řešitel: prof. RNDr. Jan Hajič, Dr.

Cíl: vytvořit podmínky pro integraci znalostí popsaných ve volném, nestrukturovaném textu do obecných systémů znalostí využívajících jak strukturovaná, tak nestrukturovaná data, a to podle potřeb konkrétní aplikace.

Projekt umožní implementovat systémy, které:

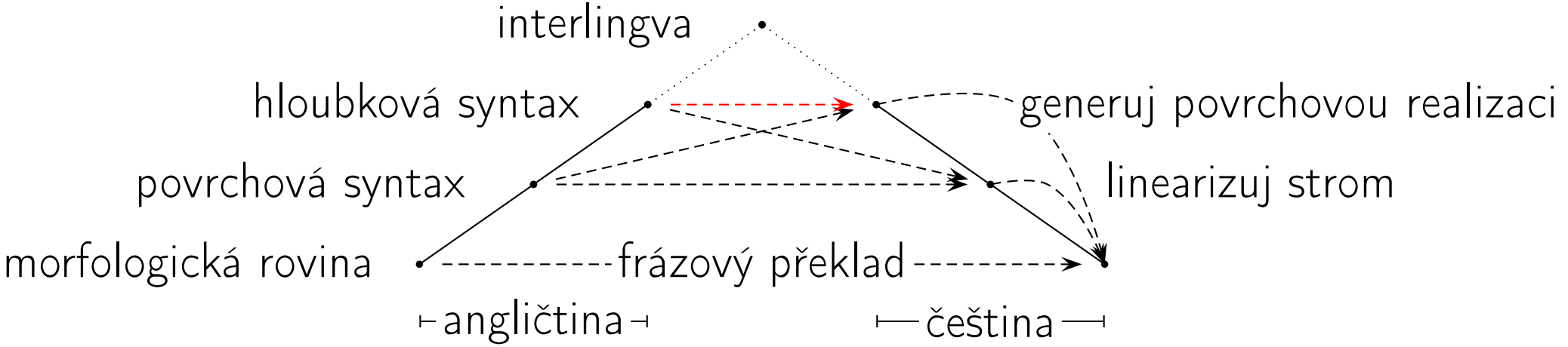
- dokážou analyzovat volný text a výsledek uložit ve strukturované podobě,
- umožní překládat z jednoho jazyka do druhého na základě obsahu.

Strojový překlad je lákavý

Strojový překlad (machine translation, MT) je zajímavý

- **akademicky**,
  - Hřiště pro testování mnoha dílčích nástrojů zpracování jazyka.
  - Test užitečnosti překladu přes hloubkový rozbor.
- **komerčně**,
  - EU utrácí ročně 1 000 000 000 eur za překlady.
  - USA investuje do překladu pro účely (kontra)rozvědky.
- i **pro uživatele**:
  - Umožňuje využít texty z webu bez ohledu na zdrojový jazyk.

Přístupy ke strojovému překladu



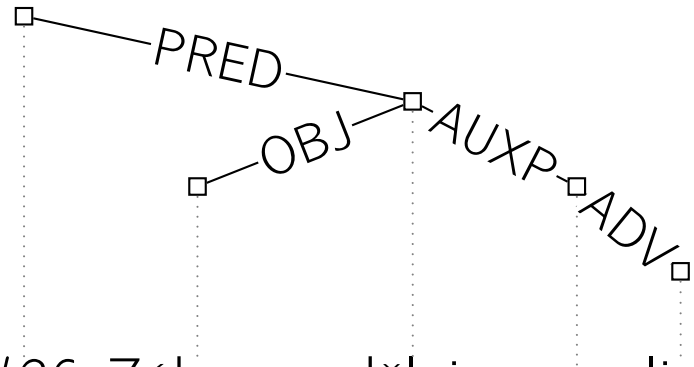
- Čím víc vstup rozeberu, tím snazší by měla být fáze transferu.
- Hypotetická interlingva zachycuje čistý význam.
- Statistické systémy se natrénují "samy" podle ukázek.
- Pravidlové systémy ručně píší lingvisté-programátoři.

Formální popis češtiny

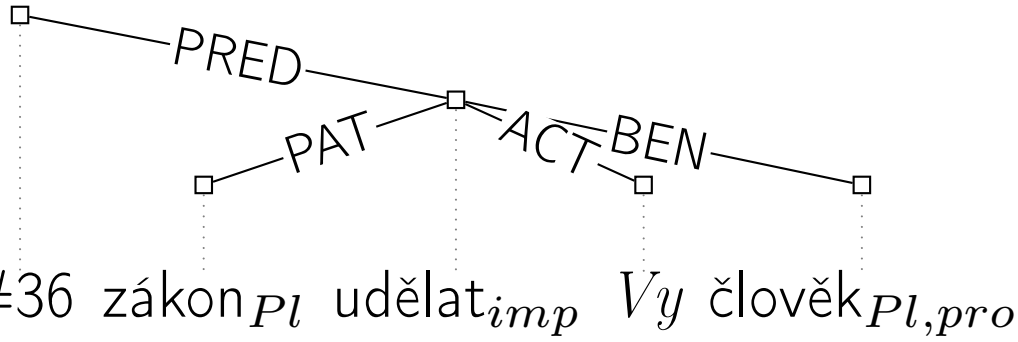
Morfologická rovina:

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):



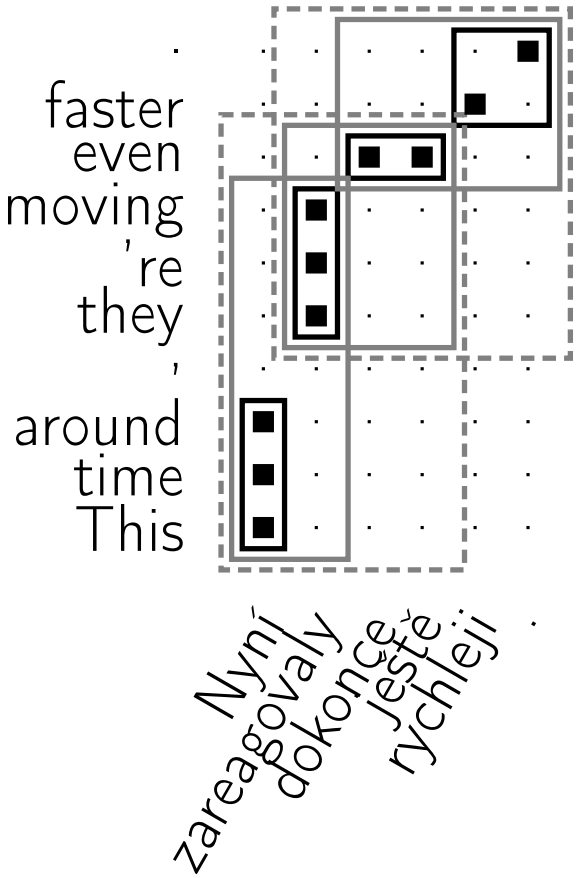
Tektogramatická rovina (hloubková syntax):



Proč je překlad těžký?

- Víceznačnost a význam slov.
  - *Spal celou Petkevičovu přednášku.*
- Cílový slovní tvar.
  - 7 pádů, 3 čísla a 4 rody ⇒ kombinatorická exploze variant výstupu.
- Pořádek slov.
  - Pro aj→čj malý problém, opačně nutno „normalizovat“.
- Negace.
  - *Nemám žádné námitky.*
  - *Udělalo se mi špatně.* ≠ *Neudělalo se mi dobře.*
- Zájmena: *Give me ... The red one.* → *Tu červenou./Ten červený.*
- Idiomatická spojení: *kick the bucket* = *natáhnout bačkory*

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:

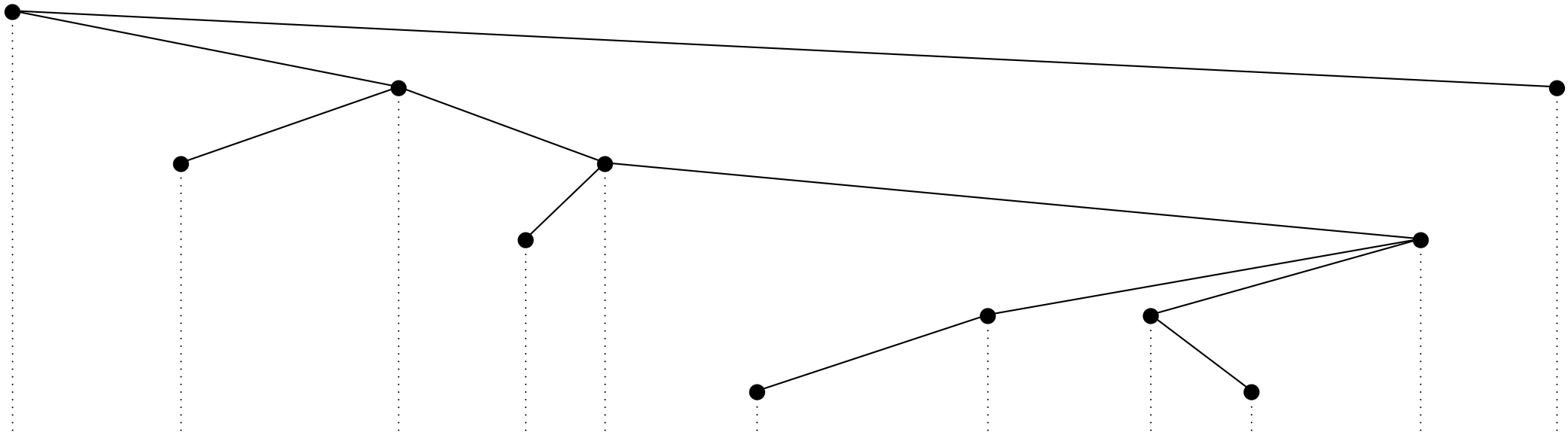
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo)

Při samotném překladu hledáme:

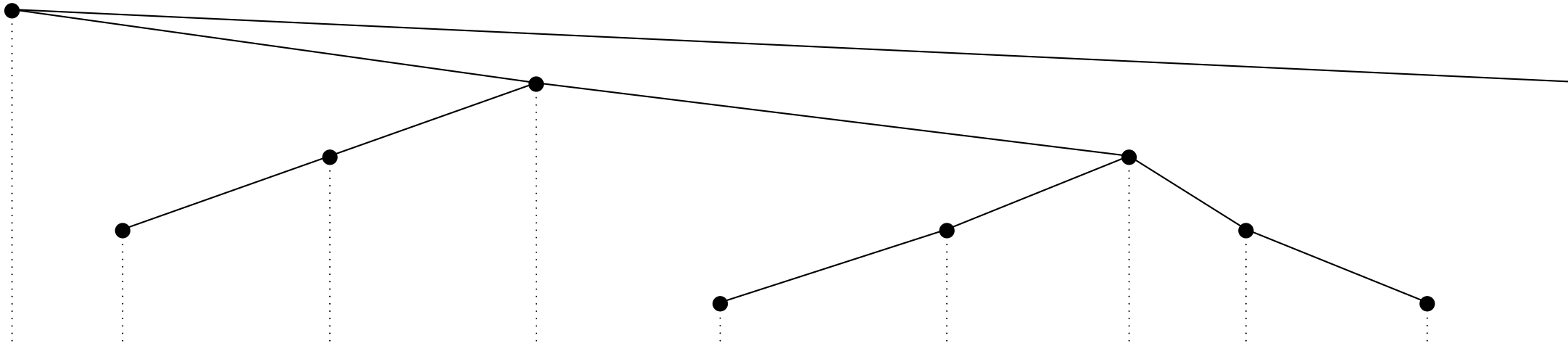
- takovou segmentaci vstupní věty na úseky („fráze“)
- a takové překlady frází

aby byl výstup co nejpravděpodobnější.

Syntaktický překlad: stromy...

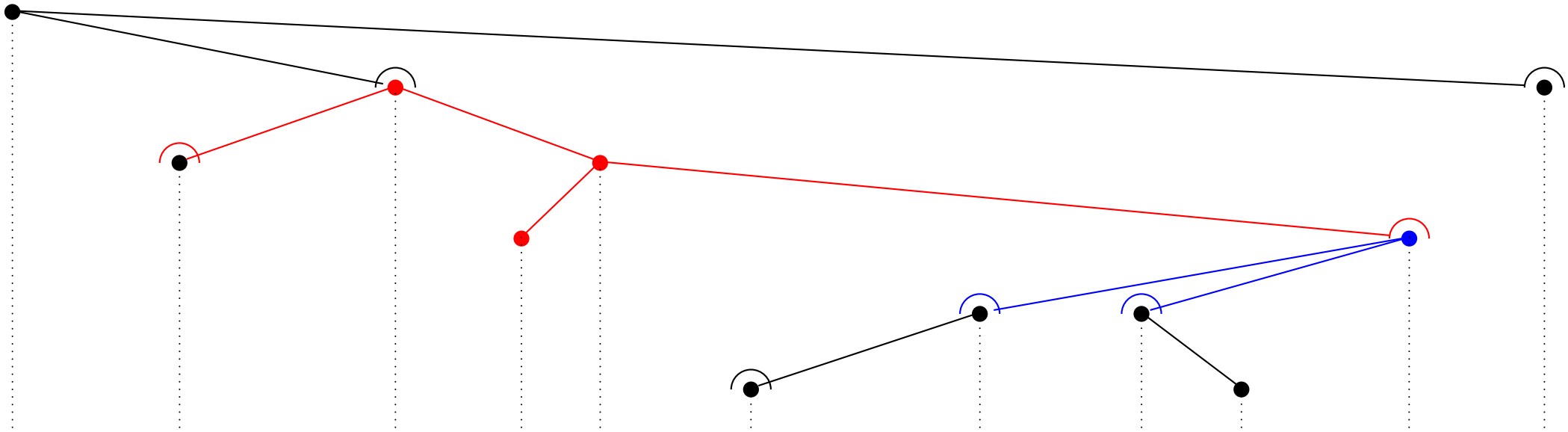


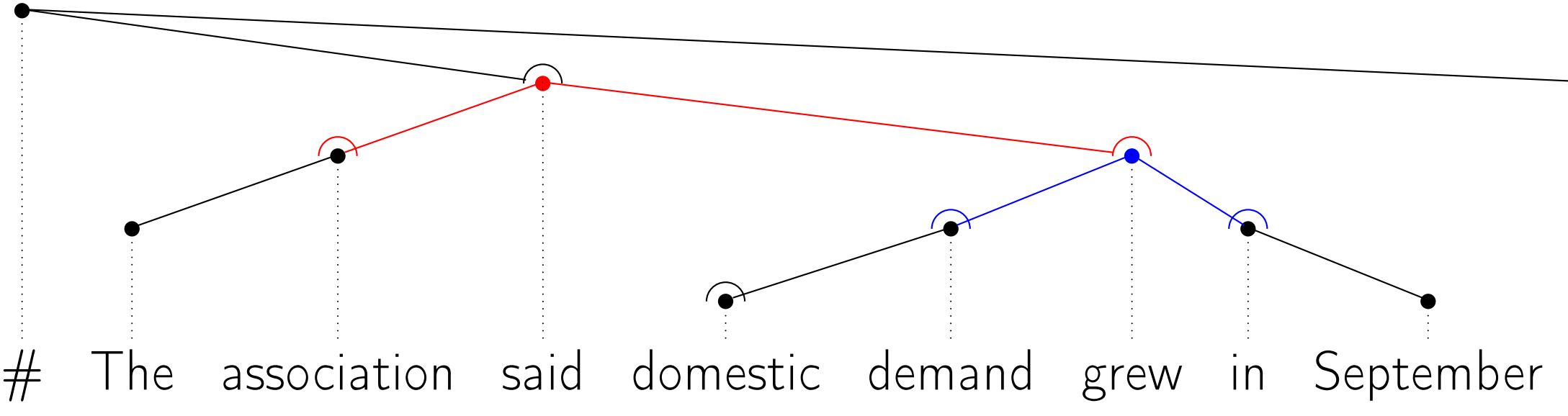
Asociace uvedla , že domácí poptávka v září stoupla .



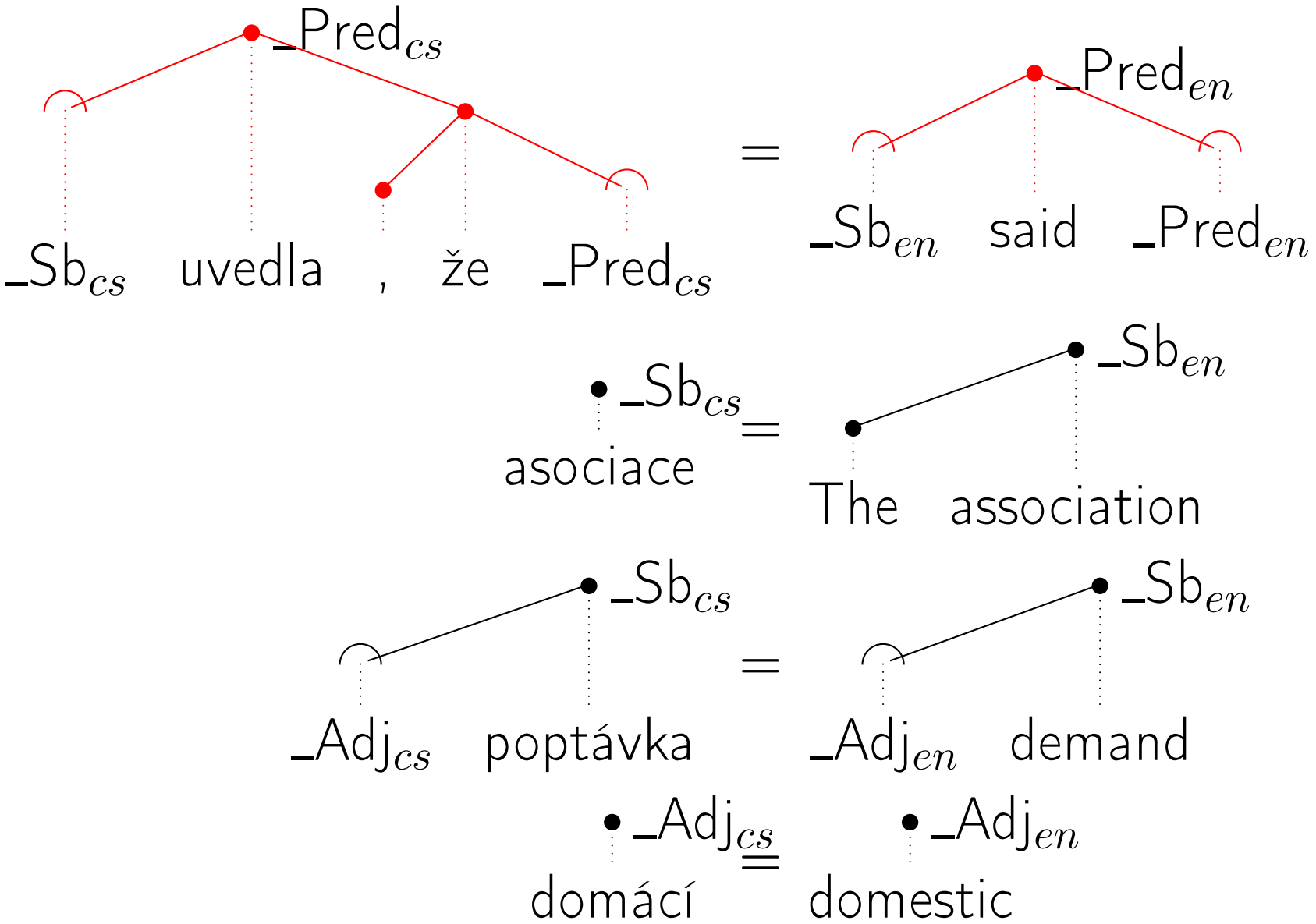
The association said domestic demand grew in September

...rozložíme na stromečky...





...a posbíráme slovník stromečků.



Frázový vs. syntaktický překlad

Frázový překlad volí primitivní řešení:

- Větu nerozebírá, jen opisuje známé podposloupnosti slov.
- Spoléhá na dostatek dat. V základní variantě neumí ani skloňovat, pokud tvar neviděl.
- Často produkuje negramatické věty, rád zahodí negaci.

Syntaktický překlad:

- Garantuje existenci větného rozboru výstupu $\Rightarrow$ naděje gramatičnosti.
- Explicitně zpřístupňuje závislosti mezi významovými jednotkami věty.
- Naráží na chyby v kaskádě nástrojů (morf.+synt. analýza).
- Naráží na „negramatický“ vstup (cokoli, co v trénovacích stromech nebylo).

$\Rightarrow$ Syntaktický překlad je těžší, má však potenciál řešit těžší problémy.

Který přístup vítězí? Nevíme.

	Frázový	Hloubkový	Google	PC Trans.
Oficiální WMT10: Seřaďte hypotézy od nejlepší po nejhorší. Shody povoleny.				
> ostatní	**45.0**	44.1	49.1	49.4
>= ostatní	**65.6**	60.1	70.4	62.1
Neoficiální WMT10: Člověk zkusil výstup MT opravit bez znalosti originálu.				
Je to dobrý překlad? (%)	**40**	34	55	43
Neoficiální: MT přeložil krátký text. Dokážete správně zodpovědět kontrolní otázky?				
% správných odpovědí	73.6	80.6	78.7	80.2

- Pravidelné soutěže (`http://www.statmt.org/wmt10/`).

Přínos projektu pro pracoviště

cca 90 publikací, **7 obhájených disertací** studentů z projektu.

Navazující projekty:

- **EuroMatrix(Plus)** (2007-2012) `http://www.euromatrixplus.net/`
  - Strojový překlad mezi všemi jazyky EU.
- **Faust** (2010-2013) `http://www.faust-fp7.eu/`
  - Strojový překlad zapojující korektury od uživatelů.
- **Khresmoi** (2010-2014) `http://www.khresmoi.eu/`
  - Vícejazyčná extrakce informací z lékařských textů a vyšetření.
- **META-NET** (2010-2013) `http://www.meta-net.eu/`
  - Platforma projektů pro vybudování technologického základu evropské mnohojazyčné společnosti.

Děkuji za pozornost

Další podrobnosti a odkazy:

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