

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



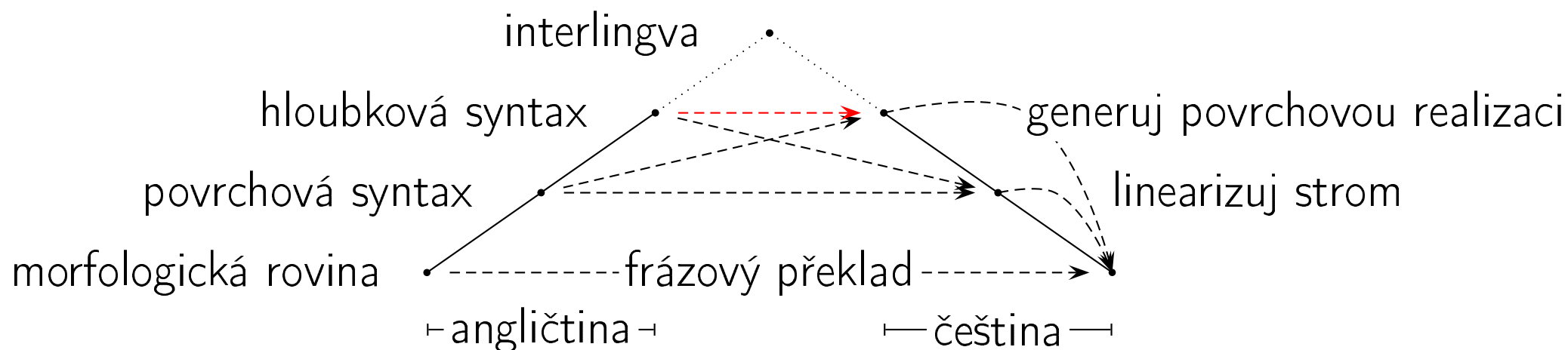
- Úvod do strojového překladu:
 - Hrubé rozdělení metod.
 - Formální popis přirozeného jazyka (čj, aj, arabština, ...),
 - Motivace k překladu.
 - Obtížnost překladu.
- Podrobněji:
 - Jazyková data (korpora, slovníky),
 - Nástroje pro automatické zpracování jazyka.
- Dva přístupy ke strojovému překladu:
 - Frázový (mj. Google).
 - Stromečkový (mj. ÚFAL).

Strojový překlad je lákavý



Strojový překlad (machine translation, MT) zajímavý akademicky, komerčně i pro uživatele:

- Hřiště pro testování užitečnosti mnoha dílčích nástrojů zpracování jazyka.
- EU utrácí ročně 1 000 000 000 eur za překlady.
- USA investuje do překladu pro účely rozvědky.
- Automatický překlad umožňuje využít texty z webu bez ohledu na zdrojový jazyk.

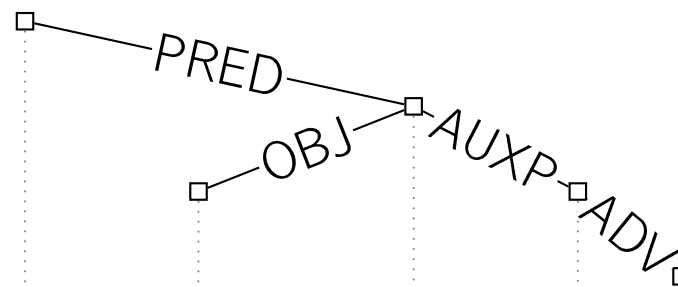


- Čím víc vstup rozeberu, tím snazší by měla být fáze transferu.
- Hypotetická interlingva zachycuje čistý význam.
- Statistické systémy se natrénují se "samy" podle ukázek.
- Pravidlové systémy ručně píší lingvisté-programátoři.

Analytická rovina (povrchová syntax):

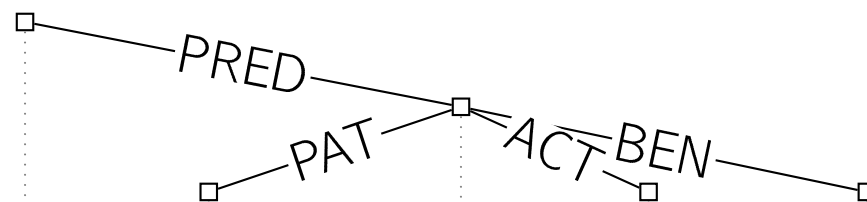
Morfologická rovina:

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----



#36 Zákony udělejte pro lidi

Tektogramatická rovina (hloubková syntax):



#36 zákon_{P1} udělat_{imp} Vy člověk_{P1,pro}

Proč je překlad těžký?



- Víceznačnost a význam slov.
- Cílový slovní tvar.
- Pořádek slov.
- Negace.
- Zájmena.
- Idiomatická spojení.

Víceznačnost a význam slov



Kolik rozborů mají věty:

Time flies like an arrow.

The bank is next to the plant.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Slovníková hesla na tom nejsou lépe:

kniha účetní

napětí dovolené

plán prací

tři prdele

Cílový slovní tvar



Časy:

- Angličtina má předpřítomný čas pro nedávnou minulost.
- Španělština má dvě varianty minulého času: pro určitý čas v minulosti a pro neznámý čas v minulosti.

Pády, rody,:

- Čeština má 7 pádů, 3 čísla a 4 rody:

The *cat* is on the mat. → kočka

He saw a *cat*. → kočku

He saw a dog with a *cat*. → kočkou

He talked about a *cat*. → kočce

⇒ Při překladu nutno vybrat správný tvar.

Pořádek slov



- Anglicky: subject-verb-object (SVO)
- Japonsky: subject-object-verb (SOV)

IBM bought Lotus.

IBM Lotus bought.

Reporters said IBM bought Lotus.

Reporters IBM Lotus bought said.

- Německy: Satzklammer (SV_1OV_2 , OV_1SV_2)

Die Satzklammer oder Klammerform **stellt** den typischen Satzbau der deutschen Sprache **dar**.

- Kombinatorická exploze možností, nestihneme probrat všechny.

- Francouzská negace je *okolo* slovesa:
Je ne parle pas français.
- Česká negace bývá zdvojená:
Nemám žádné námitky.
- Umístění negace mění význam:
Nemohl jsem přijít, ...
...ráno se mi udělalo špatně.
...ráno se mi neudělalo dobře.
- V severní a jižní Itálii se prý jízdenka v MHD procvaknutím:
zneplatňuje nebo *učiní platnou* (in/validare).

- V angličtině musí být podmět vyjádřen \Rightarrow nutno doplnit podle slovesa:

Četl knihu. = He read a book.

Spal jsem. = I slept.

- Rod českého zájmena musí odpovídat odkazovanému slovu:

He saw a book. *It was red.*

Viděl knihu. Byla černá.

He saw a pen. *It was red.*

Viděl pero. Bylo černé.

Idiomatická spojení



Kromě známého:

kick the bucket = natáhnout bačkory
a bone of contention = jablko sváru

jde i „obyčejná“ frázová slovesa:

run into = potkat
show up = přijít, ukázat se, stavit se
make up = vymyslet si
talk sb. into sth. = přemluvit někoho, aby ...

- **Korpusy** jsou (velké) sbírky textů:
 - Texty typicky označované nebo včetně větných rozborů.
Pražský závislostní korpus (PDT): 1 mil. vět ručně.
 - Některé vícejazyčné: CzEng (7 mil. vět, 50 mil. slov, odpovídá ~1-2 tis. knih, ale většinu tvoří právní texty).
 - **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
- ⇒ Lze využít v programech (pravidlových i statistických).

- Identifikace kódování dokumentu a jazyka.

- Rozpoznávání mluvené řeči.

- Rozpoznání hranic vět a slov:

Švejk 12. prosince dorazil na král. Vinohrady s dopisem.
ajskrím → I scream / icecream.

- Morfologická analýza.

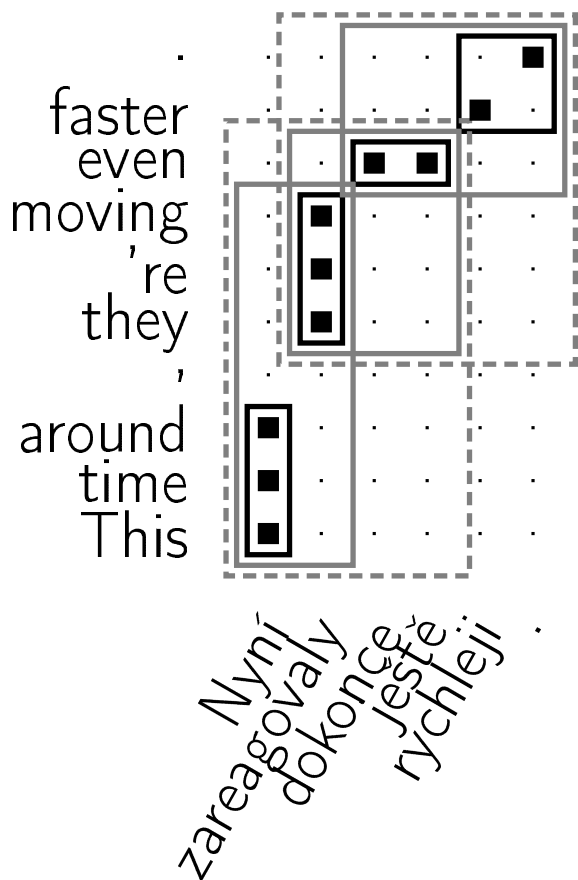
- Povrchový a hloubkový větný rozbor.

- Identifikace pojmenovaných entit:

Bílý dům se nechal slyšet.

- Koreference (mj. identifikace, co zastupují zájmena).

- Vyhledávání dokumentů (na webu).
- Kontrola překlepů.
- Kontrola pravopisu.
- Syntéza a rozpoznávání řeči.
- Automatická sumarizace textů.
- Strojový překlad.
- Strojový překlad mluvené řeči.



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

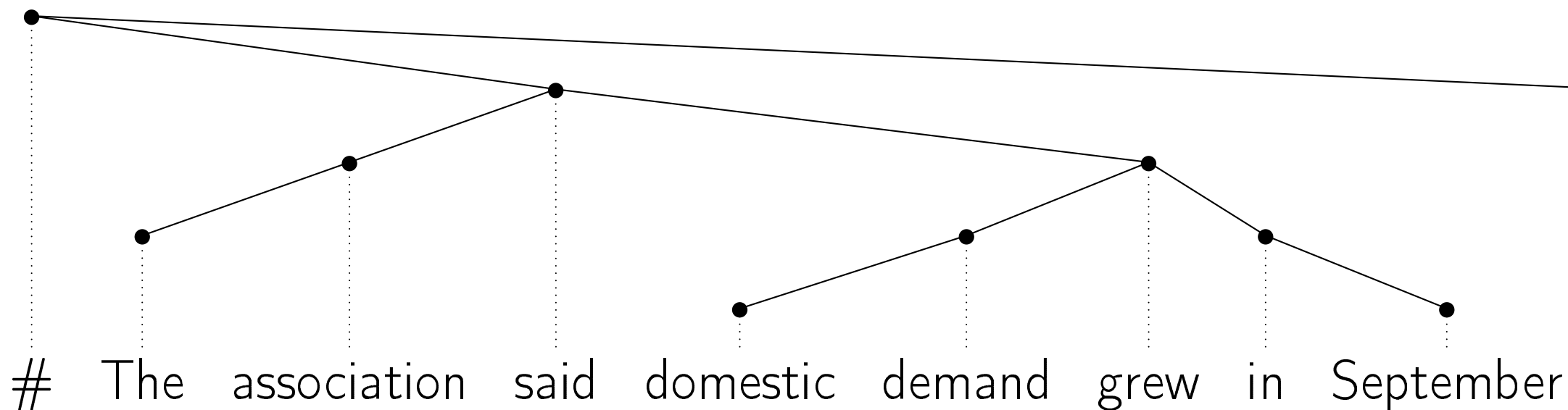
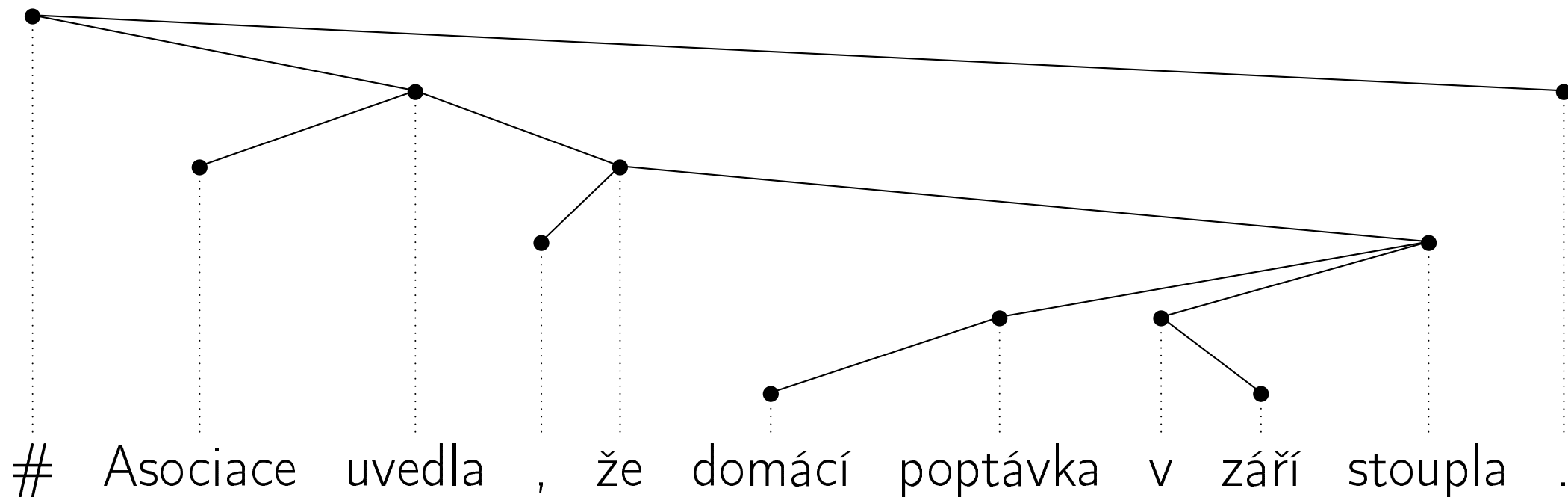
Trénovací data:

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

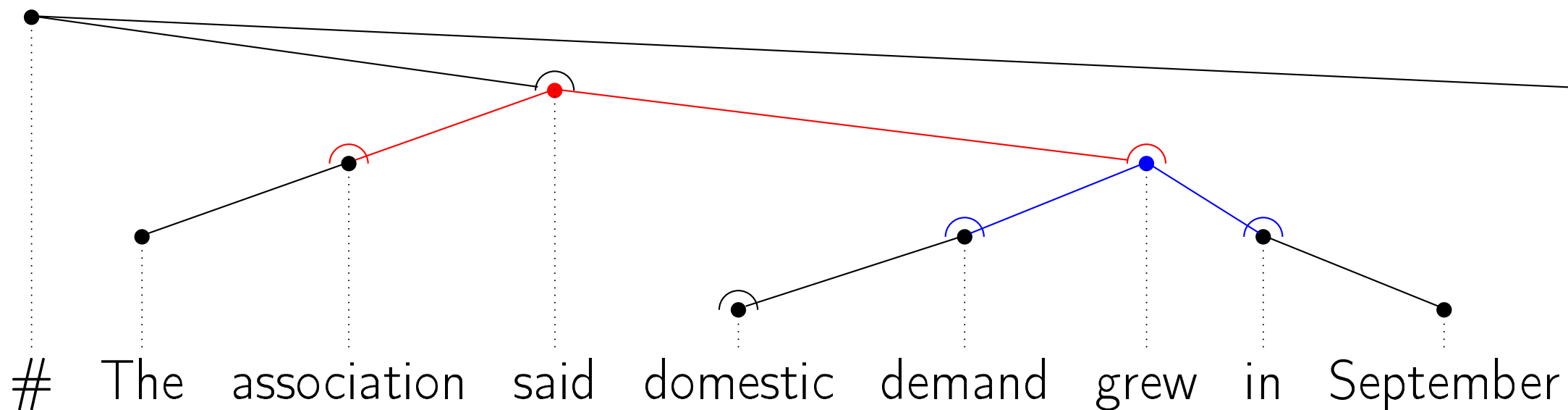
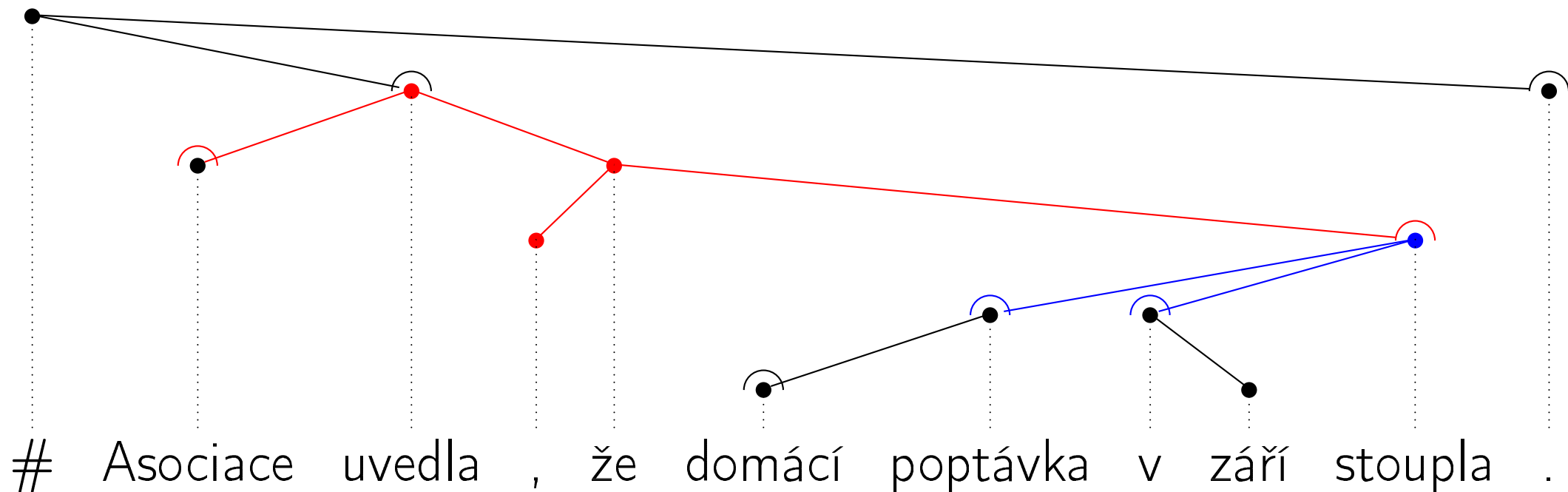
Při samotném překladu hledáme:

- takovou segmentaci vstupní věty na úseky („fráze“)
- a takové překlady frází
aby byl výstup co nejpravděpodobnější.

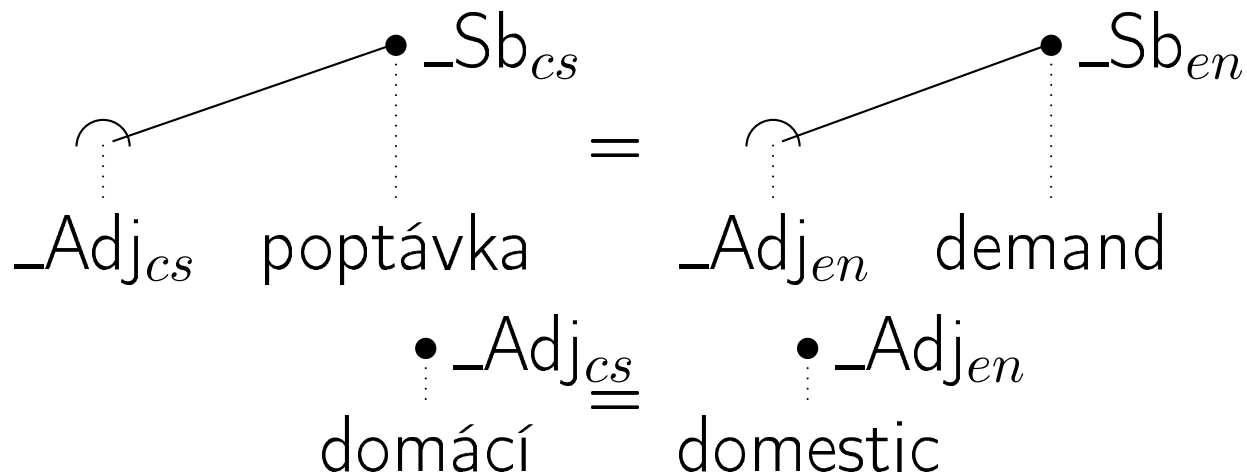
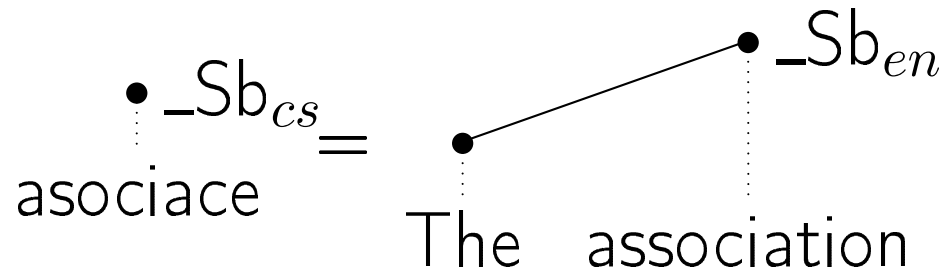
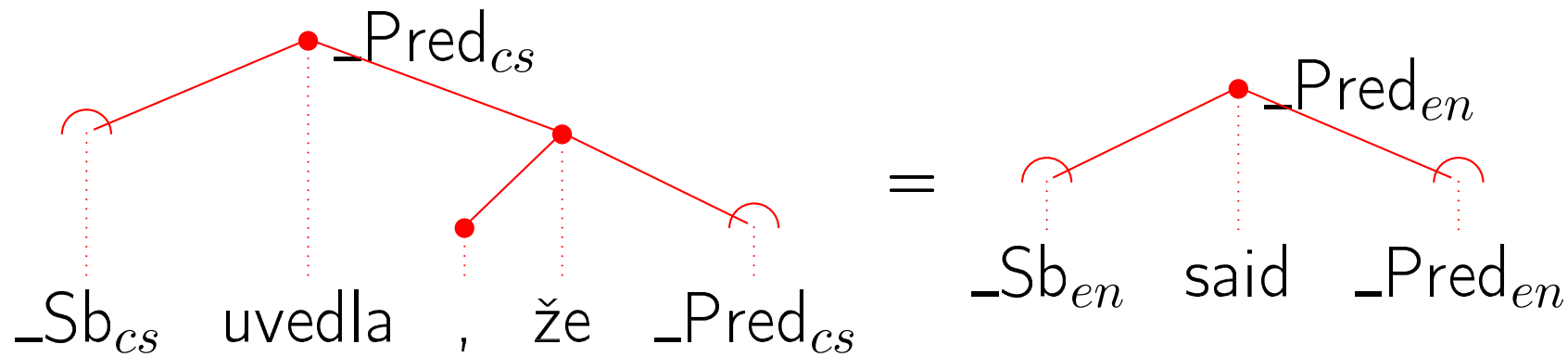
Syntaktický překlad: stromy...



...rozložíme na stromečky...



...a sebereme slovník stromečků.



Ke zdrojovému stromu hledáme rozklad a překlady stromečká, aby byl cílový závislostní strom co nejpravděpodobnější:

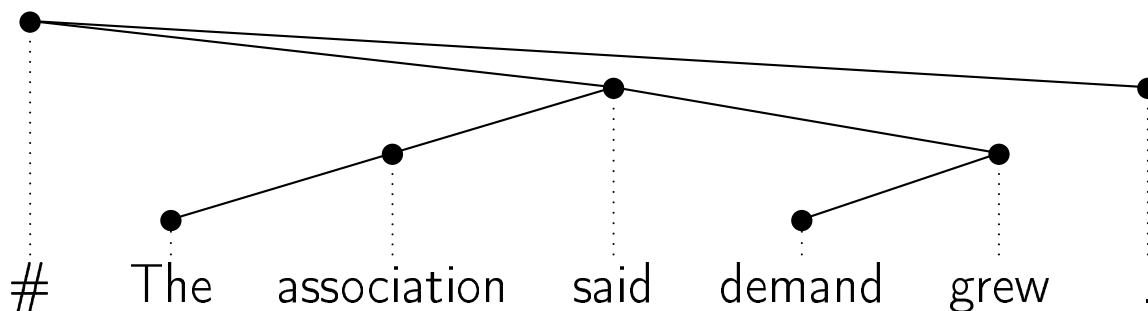
1. Příprava tabulky **možností překladu**:

- Pro každý vstupní uzel studuji všechny stromečky, které zde mohou začínat.
- Pokud ke zvolenému stromečku existuje cílový, našli jsme možnost překladu.
- Uchováváme jen τ nejlepších možností překladu pro každý uzel.

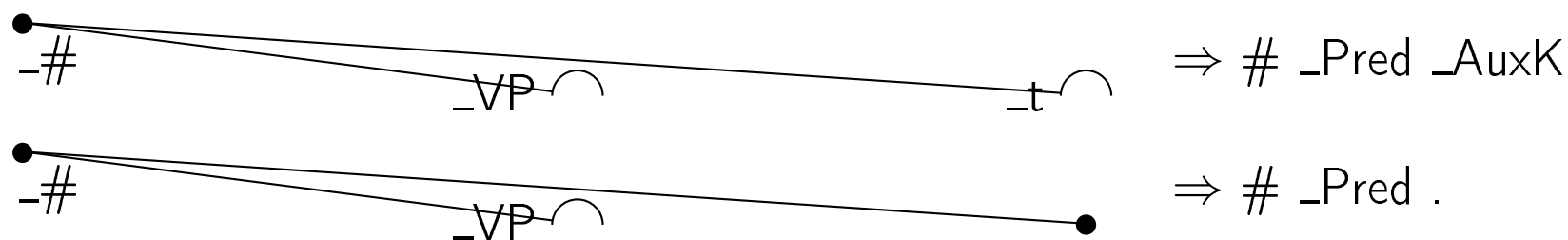
2. Postupné **budování částečných hypotéz**:

- Od kořene dolů zdrojový strom pokrýváme překladovými možnostmi.
- Uchováváme jen σ nejlepších částečných hypotéz dané velikosti (počet vstupních uzlů pokrytých vnitřními uzly)

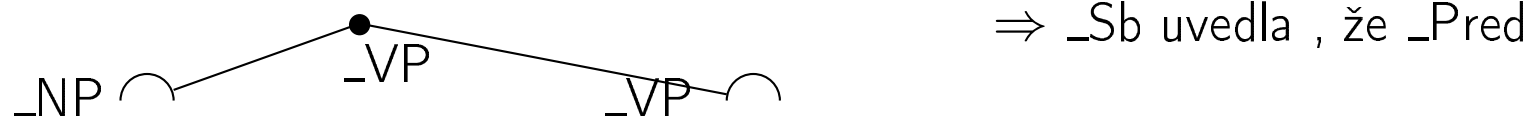
Ukázka možností překladu



Možnosti překladu v kořeni:



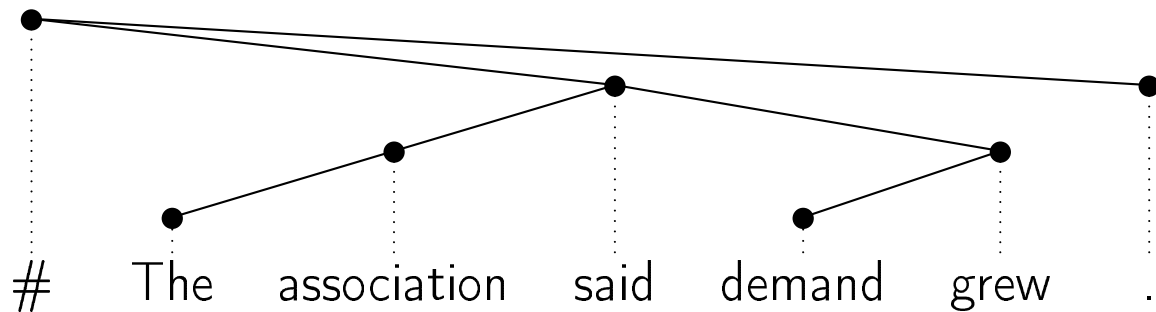
Možnosti překladu v uzlu „said“:



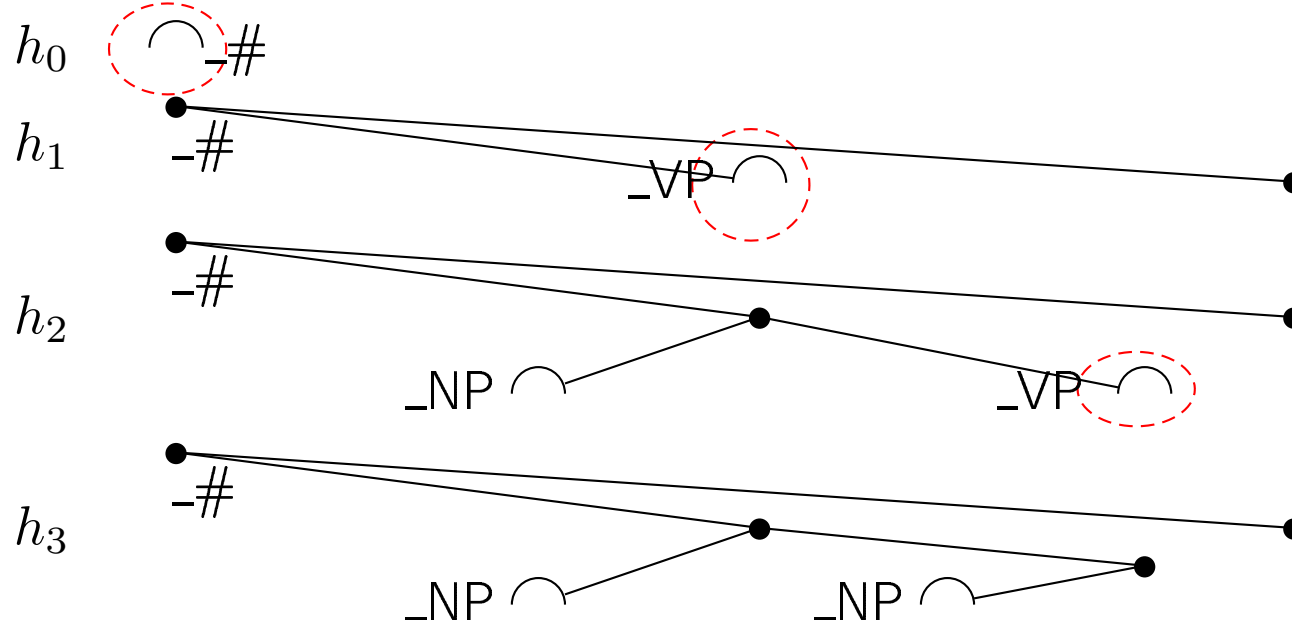
Možnosti překladu v uzlu „.“:



Postupné budování hypotéz



Ukázková derivace:



Linearizovaný výstup:

- ⇒ #
- ⇒ # _Pred
- ⇒ # _Sb uvedla , že _Pred
- ⇒ # _Sb uvedla , že _Sb stoupla .

Pro danou větu:

- Je těžké správně rozebrat („strojově pochopit“) vstup.
- Je těžké získat překladový slovník, který by obsahoval všechno, co věta potřebuje.
- Možností je příliš mnoho (varianty slov, slovních tvarů, pořadí slov).
⇒ Nutno studovat jen ty nadějně.
- Je těžké poznat lepší možnosti.
(I lidé se neshodnou v tom, jak něco přeložit.)

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.
- Naráží na chyby v kaskádě nástrojů (morf.+synt. analýza).
- Naráží na „negramatický“ vstup (cokoli, co v trénovacích stromech nebylo).

\Rightarrow Zatím funguje lépe frázový překlad.

\Rightarrow Syntaktický překlad má ale potenciál řešit těžší problémy.

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

→ Research → Prague Dependency Treebank 2.0

Ukázková data: <http://ufal.mff.cuni.cz/pdt2.0/visual-data/sample/index.htm>

→ Video Recordings

→ Tools (→ překladový systém Moses)

Další ukázky frázového překladu:

<http://demo.statmt.org/>

<http://tool.statmt.org/>