

# Jak nachytat strojový překlad na švestkách



Ondřej Bojar

[bojar@ufal.mff.cuni.cz](mailto:bojar@ufal.mff.cuni.cz)

Ústav formální a aplikované lingvistiky

Matematicko-fyzikální fakulta

Univerzita Karlova v Praze

# Optimismus na začátek



布拉格城市公共交通包括：城市火车、地铁、有轨电车、公共汽车。地铁一共有A、B、C三条线，纵横交错贯穿整个布拉格，三条地铁线在市中心都可以互相交错转换。

布拉格城市公共交通包括：城市火车、地铁、有轨电车、公共汽车。地铁一共有A、B、C三条线，纵横交错贯穿整个布拉格，三条地铁线在市中心都可以互相交错转换。

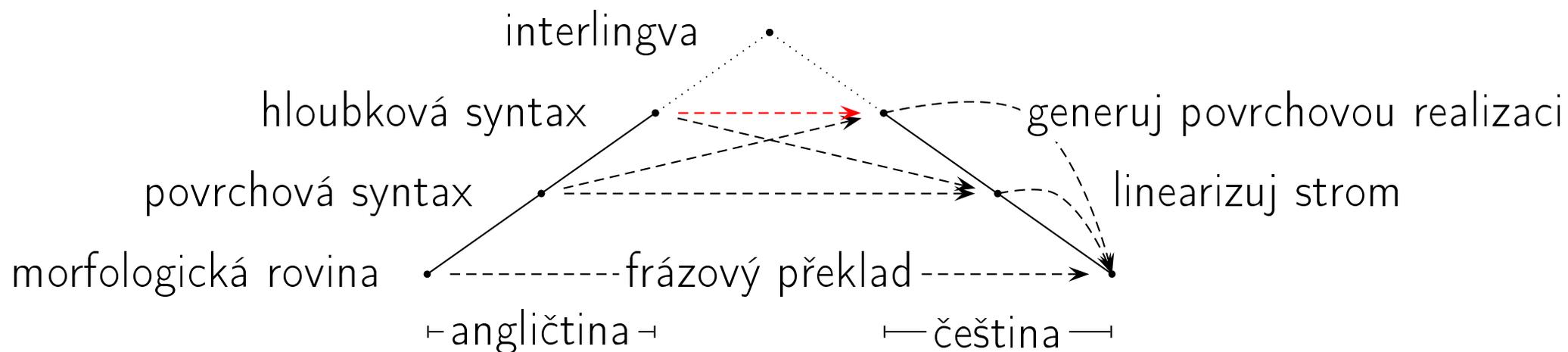
Prague city public transport, including: City train, subway, rail trams, buses. Metro, a total of A, B, C three lines, criss-cross throughout Prague, three subway lines cross each other in the city center can be converted.

布拉格城市公共交通包括：城市火车、地铁、有轨电车、公共汽车。地铁一共有A、B、C三条线，纵横交错贯穿整个布拉格，三条地铁线在市中心都可以互相交错转换。

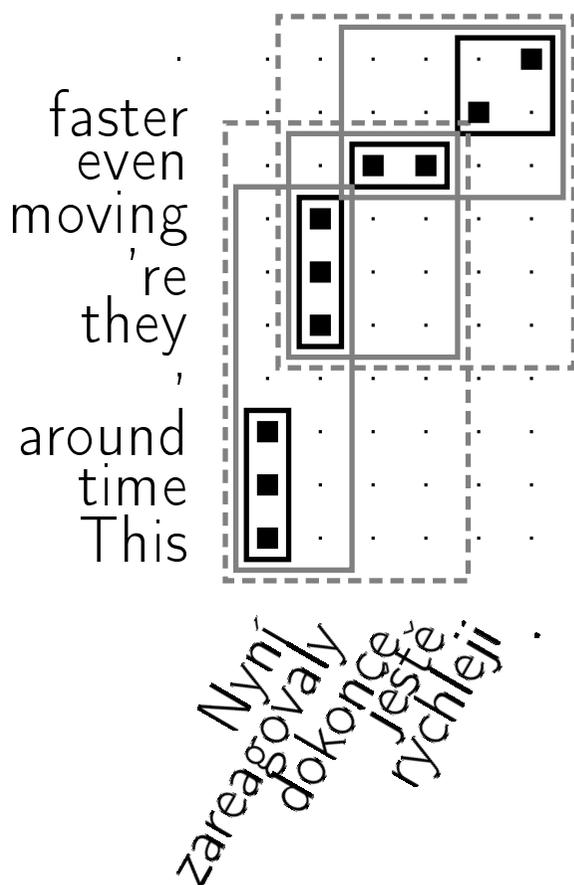
Prague city public transport, including: City train, subway, rail trams, buses. Metro, a total of A, B, C three lines, criss-cross throughout Prague, three subway lines cross each other in the city center can be converted.

Praha městská hromadná doprava, včetně: městský vlak, metro, tramvaj, autobus. Metro, celkem A, B, C tři řádky, křížem krážem po celé Praze, tři linky metra kříží v centru města může být převeden.

- Hrubé rozdělení metod strojového překladu.
- Frázový překlad (mj. Google, ÚFAL).
  - ...a jeho nevýhody.
- Stromečkový (mj. Systran, ÚFAL).
  - Formální popis přirozeného jazyka (čj, aj, arabština, ...),
  - Obtížnost překladu.
- Srovnání obou přístupů.



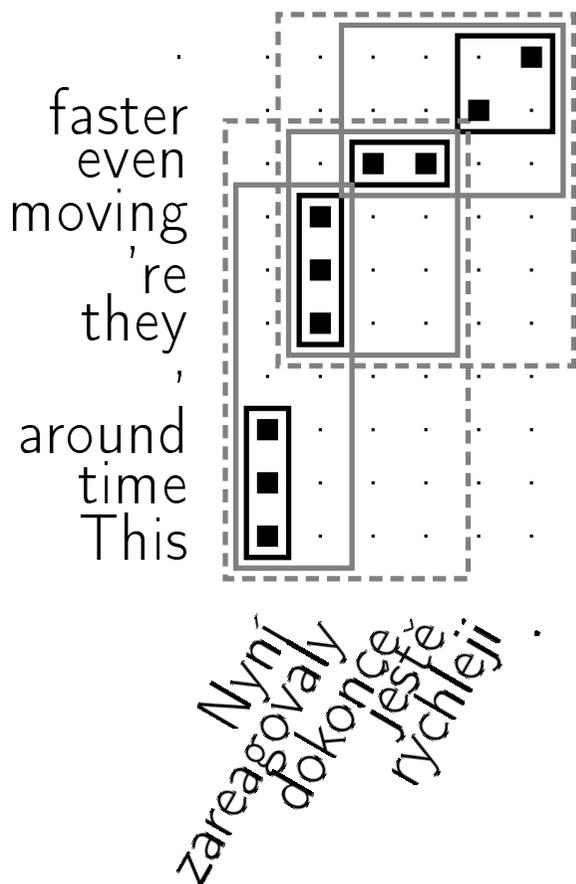
- Čím víc vstup rozeberu, tím snazší by měla být fáze transferu.
- Hypotetická interlingva zachycuje čistý význam.
- Statistické systémy se natrénují se “samy” podle ukázek.
- Pravidlové systémy ručně píší lingvisté-programátoři.



## Trénovací data:

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





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.

# (Ne)výhody frázového přístupu



- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊕ Dokáže zachytit i idiomatická spojení, jsou-li v datech.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory.

Kick the bucket.



# (Ne)výhody frázového přístupu



- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊕ Dokáže zachytit i idiomatická spojení, jsou-li v datech.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory.

Kick the bucket.



Proč musel natáhnout bačkory?

Why did he kick the bucket?



# (Ne)výhody frázového přístupu



- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊕ Dokáže zachytit i idiomatická spojení, jsou-li v datech.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory.

Kick the bucket.



Proč musel natáhnout bačkory?

Why did he kick the bucket?



Proč natáhl bačkory?

Why stretched slippers?



# (Ne)výhody frázového přístupu



- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊕ Dokáže zachytit i idiomatická spojení, jsou-li v datech.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory.

Kick the bucket.



Proč musel natáhnout bačkory?

Why did he kick the bucket?



Proč natáhl bačkory?

Why stretched slippers?



Proč musel natáhnout bačkory Karel?

Why did he kick the bucket Charles?



# (Ne)výhody frázového přístupu



- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊕ Dokáže zachytit i idiomatická spojení, jsou-li v datech.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory.

Kick the bucket.



Proč musel natáhnout bačkory?

Why did he kick the bucket?



Proč natáhl bačkory?

Why stretched slippers?



Proč musel natáhnout bačkory Karel?

Why did he kick the bucket Charles?



John se snažil natáhnout bačkory.

John tried to kick the bucket.



# Jak nachytat na švestkách...



...nemusí jít o idiomy, abyste frázový překlad odhalili:

Jan s Marií se vzali.

John and Mary were married.



# Jak nachytat na švestkách...



...nemusí jít o idiomy, abyste frázový překlad odhalili:

Jan s Marií se vzali.

John and Mary were married.



Jan s Marií se včera vzali.

John and Mary married yesterday.



# Jak nachytat na švestkách...



...nemusí jít o idiomy, abyste frázový překlad odhalili:

Jan s Marií se vzali.      John and Mary were married.      ✓

Jan s Marií se včera vzali.      John and Mary married yesterday.      ✓

Jan s Marií se včera v kostele vzali.

John and Mary are married in church yesterday.      ~

# Jak nachytat na švestkách...



...nemusí jít o idiomy, abyste frázový překlad odhalili:

Jan s Marií se vzali. John and Mary were married. ✓

Jan s Marií se včera vzali. John and Mary married yesterday. ✓

Jan s Marií se včera v kostele vzali.  
John and Mary are married in church yesterday. ~

Jan s Marií se včera v kostele svatého Ducha vzali.  
John and Mary yesterday in the Church of the Holy Spirit took. ✗

# Jak nachytat na švestkách...



...nemusí jít o idiomy, abyste frázový překlad odhalili:

Jan s Marií se vzali. John and Mary were married. ✓

Jan s Marií se včera vzali. John and Mary married yesterday. ✓

Jan s Marií se včera v kostele vzali.  
John and Mary are married in church yesterday. ~

Jan s Marií se včera v kostele svatého Ducha vzali.  
John and Mary yesterday in the Church of the Holy Spirit took. ✗

...zkusme tedy překlad dělat pořádně.

zákony

udělejte

pro

lidi

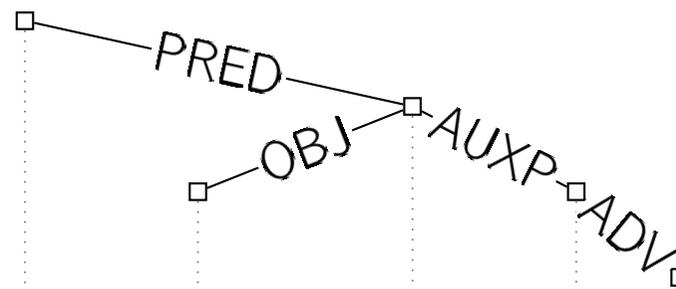
## Morfologická rovina:

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):

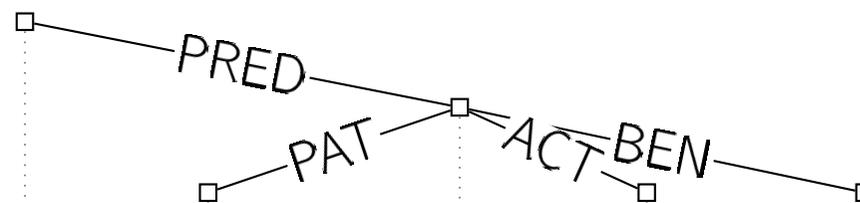
## Morfologická rovina:

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----



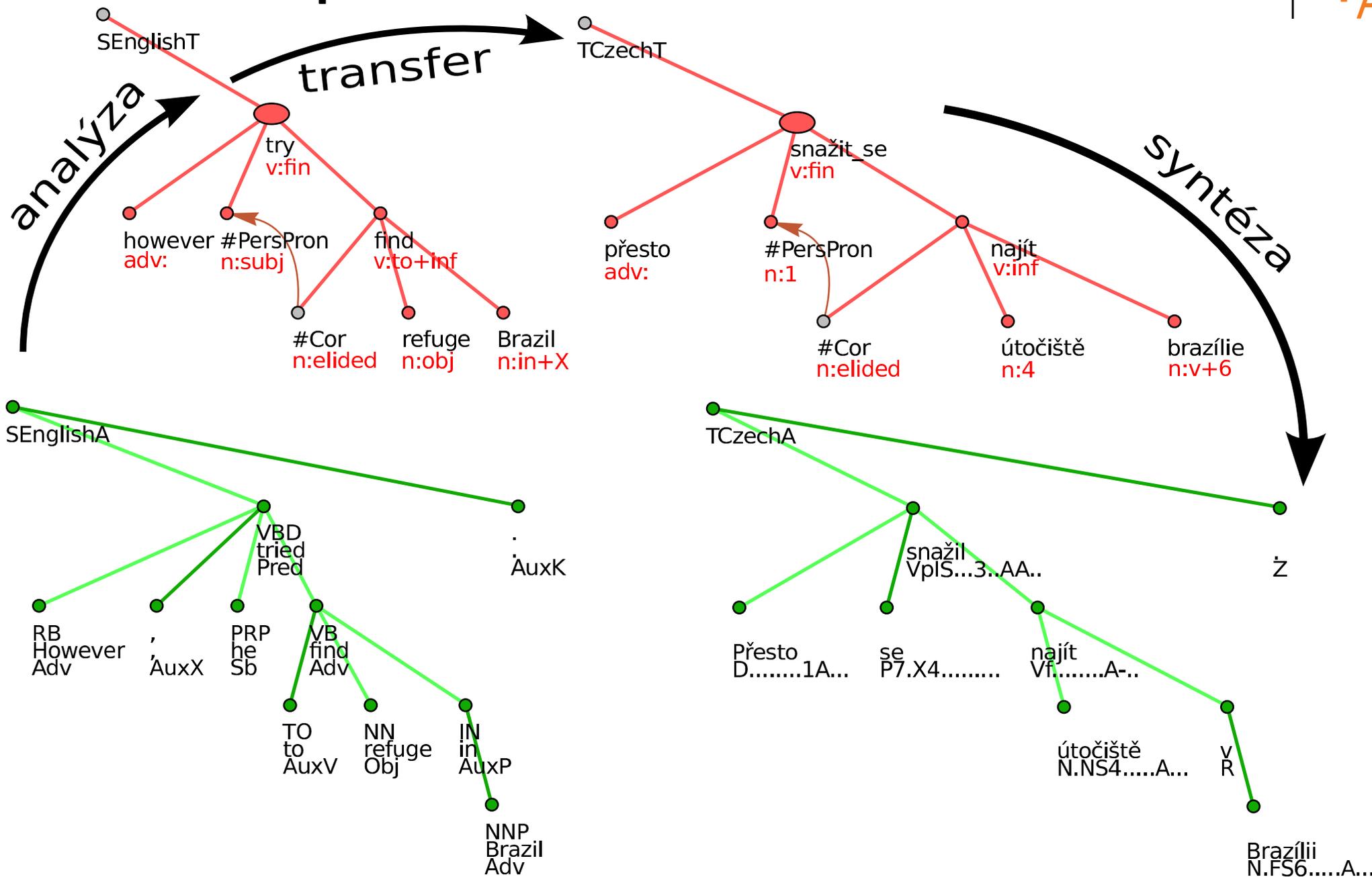
#36 Zákony udělejte pro lidi

Tektogramatická rovina  
(hloubková syntax):



#36 zákon<sub>Pl</sub> udělat<sub>imp</sub> Vy člověk<sub>Pl,pro</sub>

# Překlad přes hloubkovou rovinu



# Proč je překlad těžký?



- Víceznačnost a význam slov.
- Cílový slovní tvar.
- Negace.
- Zájmena.

A už zmiňované:

- Pořádek slov (tj. i vzdálenost mezi slovy).
- Idiomatická spojení.

# Víceznačnost a význam slov



Time flies like an arrow.

# Víceznačnost a význam slov



Time flies like an arrow.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

# Víceznačnost a význam slov



Time flies like an arrow.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Slovníková hesla na tom nejsou lépe:

kniha účetní, napětí dovolené, plán prací, tři prdele

# Víceznačnost a význam slov



Time flies like an arrow.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Slovníková hesla na tom nejsou lépe:

kniha účetní, napětí dovolené, plán prací, tři prdele

Reálný příklad:

SRC One tap and the machine issues a slip with a number.

REF Jedno ťuknutí a ze stroje vyjede papírek s číslem.

---

Moses 1 Z jednoho kohoutku a stroj vydá složenky s číslem.

Moses 2 Jeden úder a stroj vydá složenky s číslem.

Google Jedním klepnutím a stroj problémy skluzu s číslem.

# 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.

- 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.

# Frázový vs. syntaktický v praxi



Stell dir das vor.

Google Imagine that.

Systran Imagine.



# Frázový vs. syntaktický v praxi



Stell dir das vor.

Google Imagine that.



Systran Imagine.



Stell dir ein Haus vor.

Google Imagine a house before.



Systran Imagine a house.



# Frázový vs. syntaktický v praxi



Stell dir das vor.

Google Imagine that.



Systran Imagine.



Stell dir ein Haus vor.

Google Imagine a house before.



Systran Imagine a house.



Stell dir ein kleines Haus vor.

Google Imagine a small house in front.



Systran Imagine a small house.



# Frázový vs. syntaktický v praxi



Stell dir das vor.

Google Imagine that.



Systran Imagine.



Stell dir ein Haus vor.

Google Imagine a house before.



Systran Imagine a house.



Stell dir ein kleines Haus vor.

Google Imagine a small house in front.



Systran Imagine a small house.



Stell dir ein kleines Haus mit vierzehn Fenster vor.

Google Imagine a small house with fourteen windows in front.



Systran Imagine a small house with fourteen windows.



# Jak nachytat syntaktický překlad |

- Stačí „pumpovat“ gramatické jevy, ne jen slova.

Stell dir ein Haus vor.

⇒ Imagine a house.



# Jak nachytat syntaktický překlad |

- Stačí „pumpovat“ gramatické jevy, ne jen slova.

Stell dir ein Haus vor.

⇒ Imagine a house.



Stell dir ein Haus, das einen Garten hat, vor.

⇒ Imagine a house, which has a garden.



# Jak nachytat syntaktický překlad |

- Stačí „pumpovat“ gramatické jevy, ne jen slova.

Stell dir ein Haus vor.

⇒ Imagine a house.



Stell dir ein Haus, das einen Garten hat, vor.

⇒ Imagine a house, which has a garden.



Stell dir ein Haus, das einen Garten, der berühmt ist, hat, vor.

⇒ Place to you a house, which a garden, which has is famous, forwards.



# Jak nachytat syntaktický překlad |

- Stačí „pumpovat“ gramatické jevy, ne jen slova.

Stell dir ein Haus vor.

⇒ Imagine a house. ✓

Stell dir ein Haus, das einen Garten hat, vor.

⇒ Imagine a house, which has a garden. ✓

Stell dir ein Haus, das einen Garten, der berühmt ist, hat, vor.

⇒ Place to you a house, which a garden, which has is famous, forwards. ✗

- Ale také stačí negramatický vstup.

Stell dir ein Haus, das  Garten hat, vor.

⇒ Place to you a house, the garden intends. ✗

# Který přístup vítězí? Nevíme.



Angličtina → čeština

ÚFAL

Komerční

FRÁZOVÝ	HLOUBKOVÝ	GOOGLE	PC TRANS.
---------	-----------	--------	-----------

Oficiální WMT10: Seřadte hypotézy od nejlepší po nejhorší. Shody povoleny.

> ostatní	<b>45.0</b>	44.1	49.1	<b>49.4</b>
>= ostatní	<b>65.6</b>	60.1	<b>70.4</b>	62.1

Neoficiální WMT10: Člověk zkusil výstup MT opravit bez znalosti originálu.

Je to dobrý překlad? (%)	<b>40</b>	34	<b>55</b>	43
--------------------------	-----------	----	-----------	----

Neoficiální: MT přeložil krátký text. Dokážete správně zodpovědět kontrolní otázky?

% správných odpovědí	73.6	<b>80.6</b>	78.7	80.2
----------------------	------	-------------	------	------

- Pravidelné soutěže (<http://www.statmt.org/wmt11/>).

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 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.

# Takže nezapomeňte...

...že překladu není radno důvěřovat moc:



# Takže nezapomeňte...



...i když vám může často dost pomoci:

天府路南洋冠盛酒店经理称：“出现蟑螂避无可避”

# Takže nezapomeňte...



...i když vám může často dost pomoci:

天府路南洋冠盛酒店经理称：“出现蟑螂避无可避”

Tianfu Road, Nanyang Royal Hotel manager,  
said: "cockroaches appear inevitable."

Tianfu Road, Nanyang Royal Hotel manažer,  
řekl: "šváby zdají nevyhnutelné."

# Takže nezapomeňte...



...a lidští překladatelé nemusí být nutně lepší:

Základem tohoto loga je Nebojsa, postava Alsasana  
získaná Thomase Fentimanem dvakrát  
při profesionálních zkouškách Crufts Obedience Test.

The Fentimans Logo is based on Fearless, Thomas Fentiman's  
prize Alsatian, double winner of the Crufts Obedience Test.

# Takže nezapomeňte...

...a lidští překladatelé nemusí být nutně lepší:

Základem tohoto loga je Nebojsa, postava Alsasana  
získaná Thomassem Fentimanem dvakrát  
při profesionálních zkouškách Crufts Obedience Test.

The Fentimans Logo is a based on Fearless, Thomas Fentiman's  
prize Alsatian, double winner of the Crufts Obedience Test.



# Lingvistika není jen překlad



Počítačová lingvistika má řadu aplikací:

Kontrola překlepů. Kontrola pravopisu.

Vyhledávání dokumentů (na webu). Syntéza a rozpoznávání řeči.

Automatická sumarizace textů. Strojový překlad (mluvené řeči).

Počítačová lingvistika na Matfyzu:

<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>

→ Tools → překladový systém Moses

Viz též <http://tool.statmt.org/> a <http://demo.statmt.org/>.

- **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).