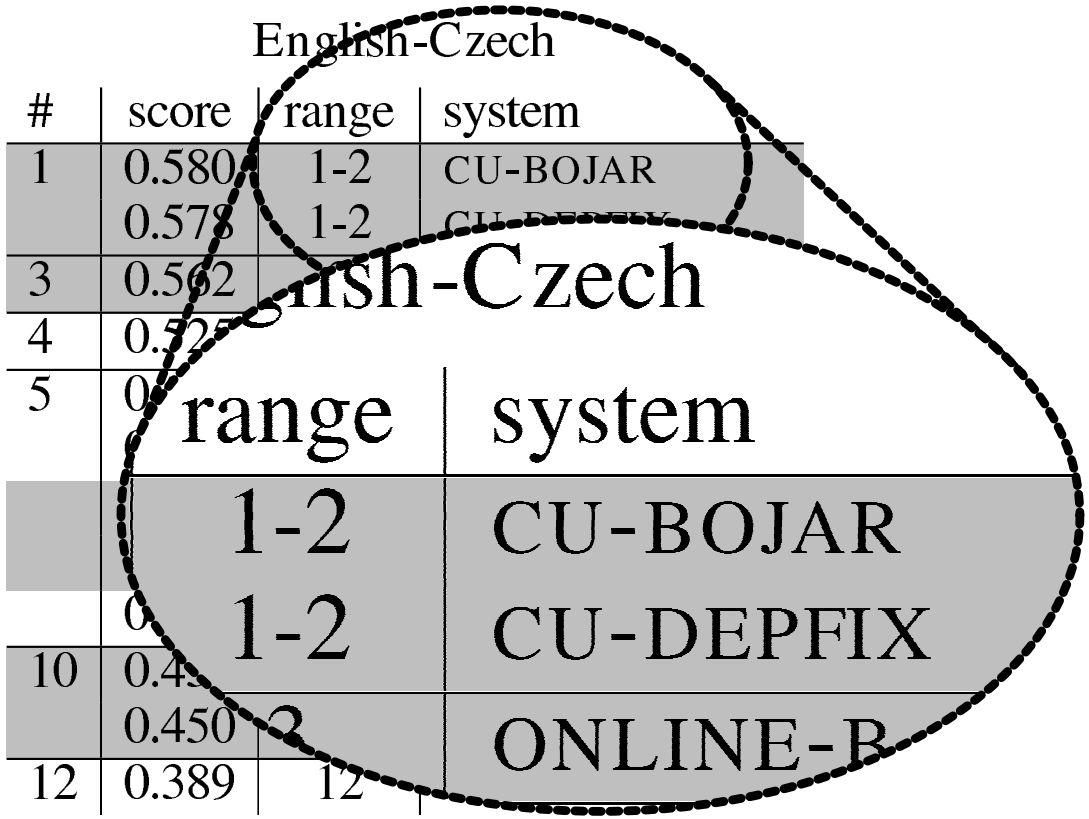
Strojový překlad lepší než Google

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
MFF UK

Jeden den s informatikou, 24. září 2013

Opravdu jsme vyhráli



English-Czech

#	score	range	system
1	0.580	1-2	CU-BOJAR
	0.578	1-2	CU-DEPFIX
3	0.562		
4	0.525		
5	0		
	0		
10	0.4		
	0.450	3	ONLINE-B
12	0.389	12	

Table 6: Official results for the WMT13 translation task.

`http://www.statmt.org/wmt13/results.html`

Výstupy systémů (ne Google):

`http://matrix.statmt.org/`

Opravdu jsme vyhráli

	The Champions League proved we could hold our own.
my	Champions League dokázala, že dokážeme obstát.
G.	Liga mistrů ukáza l bychom mohli držet naše vlastní.
Ref	V Lize mistrů jsme dokázali, že hrát umíme.

Opravdu jsme vyhráli

The Champions League proved we could hold our own.

---

my Champions League dokázala, že dokážeme obstát.

G. Liga mistrů ukázal bychom mohli držet naše vlastní.

Ref V Lize mistrů jsme dokázali, že hrát umíme.

Republican leaders justified their policy
by the need to combat electoral fraud.

---

my Republikánští vůdci ospravedlňovali svou politiku
potřebou bojovat proti volebním podvodům.

G. Republikánští vůdcové odůvodněné svou politiku
o potřebě boje proti volební podvody.

Ref Republikánští lídři odůvodnili svou politiku
nezbytností boje proti volebním podvodům.

Vyhráli jsme opravdu?

- Soustava tří složitých komponent.
- Každá ušita na míru překladu z angličtiny do češtiny.
- Potřebuje obrovská data.
- Potřebuje velmi pokročilé lingvistické nástroje.
- Žádné záruky ani pro náš výstup.

Ale pěkně popořádku

- Obtížnost překladu.
- Formální popis češtiny.
- Přístupy ke strojovému překladu.
- Tři součásti našeho systému:
  -  Hloubkový překlad.
  -  Frázový překlad
  -  Automatické opravy chyb.
- Srovnání přístupů k překladu.

- Počítačová lingvistika není jen překlad.

Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Put it on the **velvety coat rack**.
. . . sametová police na kabáty?
. . . police na sametové kabáty?

Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Put it on the **rusty coat rack**.
. . . rezavá police na kabáty?
. . . police na rezavé kabáty?

Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Put it on the **rusty coat rack**.
. . . rezavá police na kabáty?
. . . police na rezavé kabáty?

Z češtiny to není lepší:

Spal celou Petkevičovu přednášku.
Ženu holí stroj.

Proč je překlad těžký

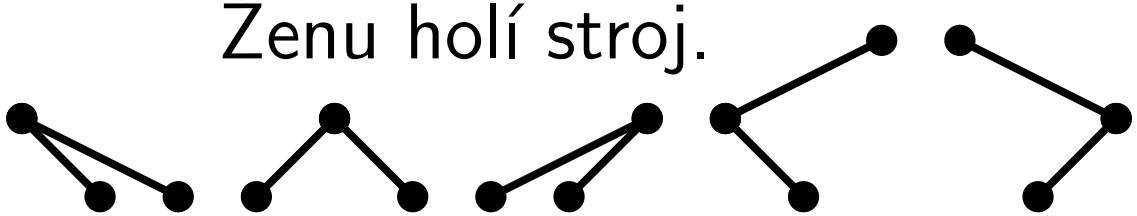
Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Put it on the **rusty coat rack**.
. . . rezavá police na kabáty?
. . . police na rezavé kabáty?

Z češtiny to není lepší:

Spal celou Petkevičovu přednášku.
Ženu holí stroj.



Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The **plant** is next to the **bank**.
rostlina? továrna? *banka? břeh?*

Put it on the **rusty coat rack**.
. . . rezavá police na kabáty?
. . . police na rezavé kabáty?

Reálné věty jsou stejně těžké:

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 složenk**y** s číslem.
Moses 2	Jeden **úder** a stroj vydá složenk**y** s číslem.
Google	Jedním klepnutím a stroj **problémy skluzu** s číslem.

Do češtiny navíc musíme trefit tvar

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	...	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	...		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	...		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			...	...		

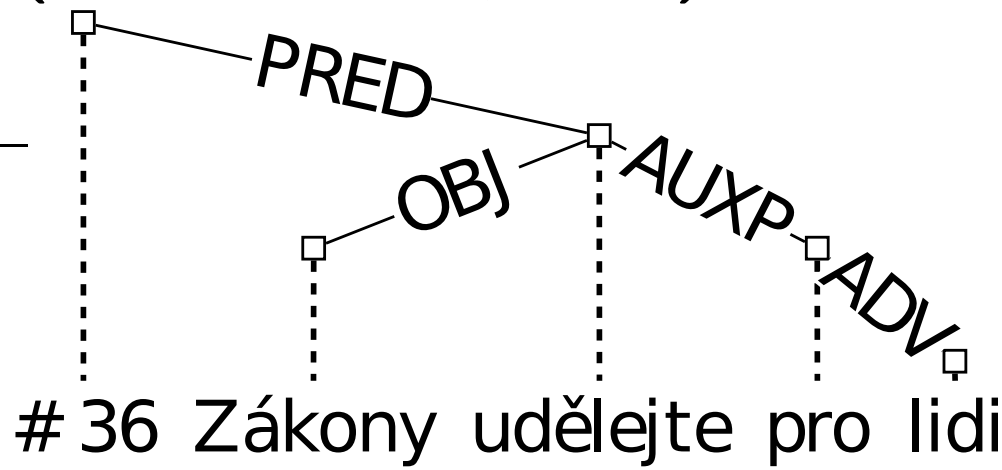
A tohle mám naprogramovat?!

Formální popis češtiny

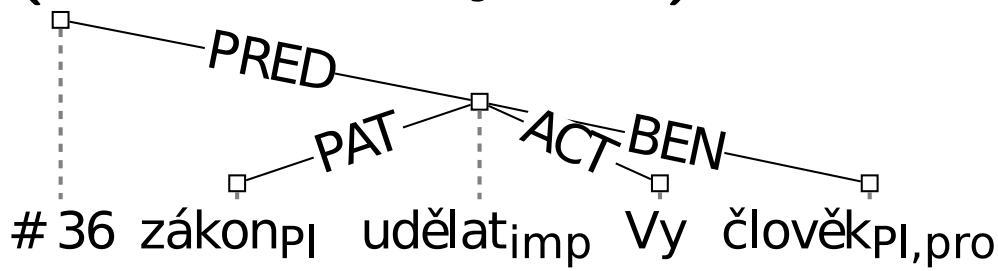
Morfologická rovina:

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----

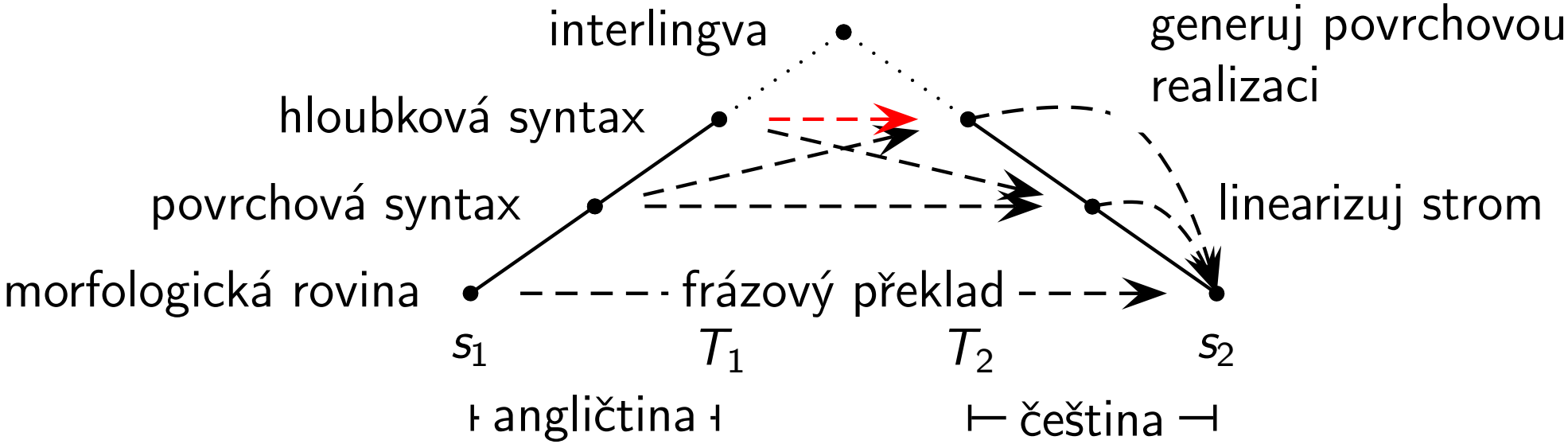
Analytická rovina (povrchová syntax):



Tektogramatická rovina (hloubková syntax):



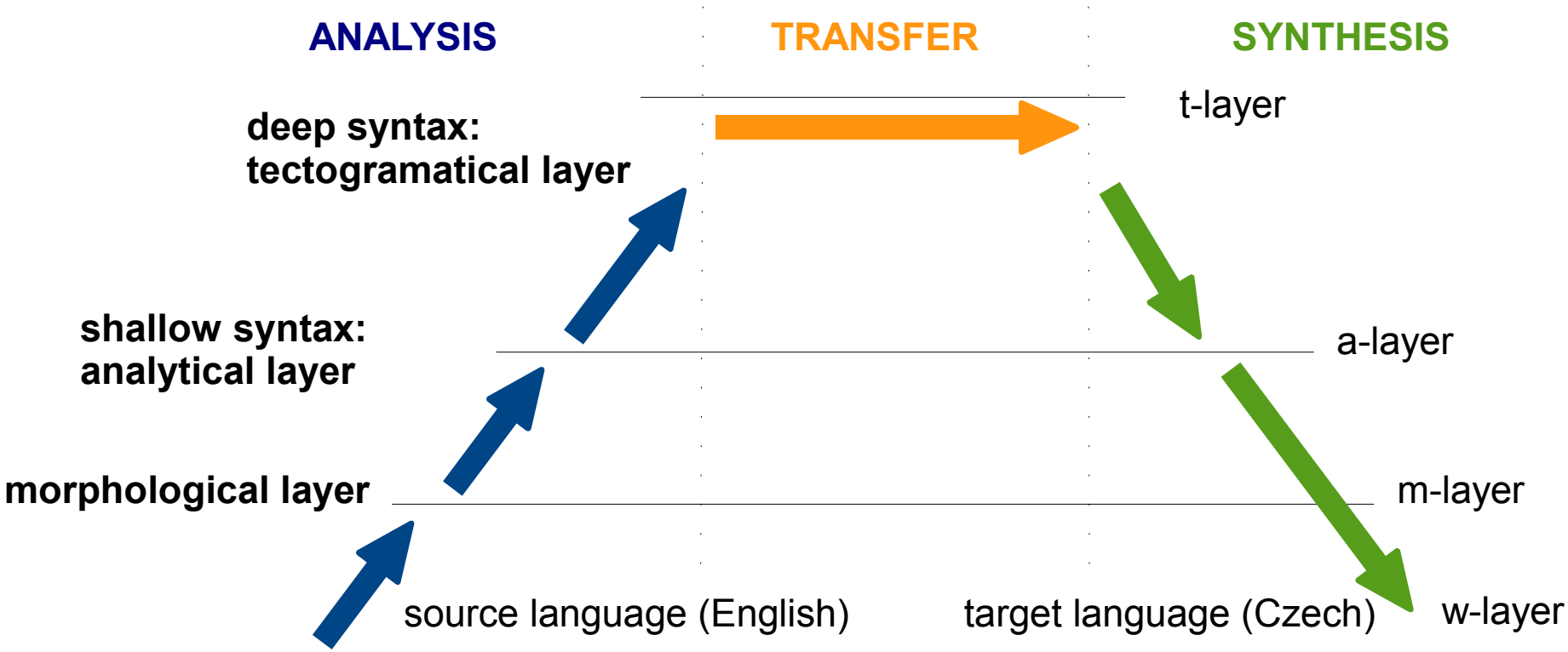
Přístupy ke strojovému překladu



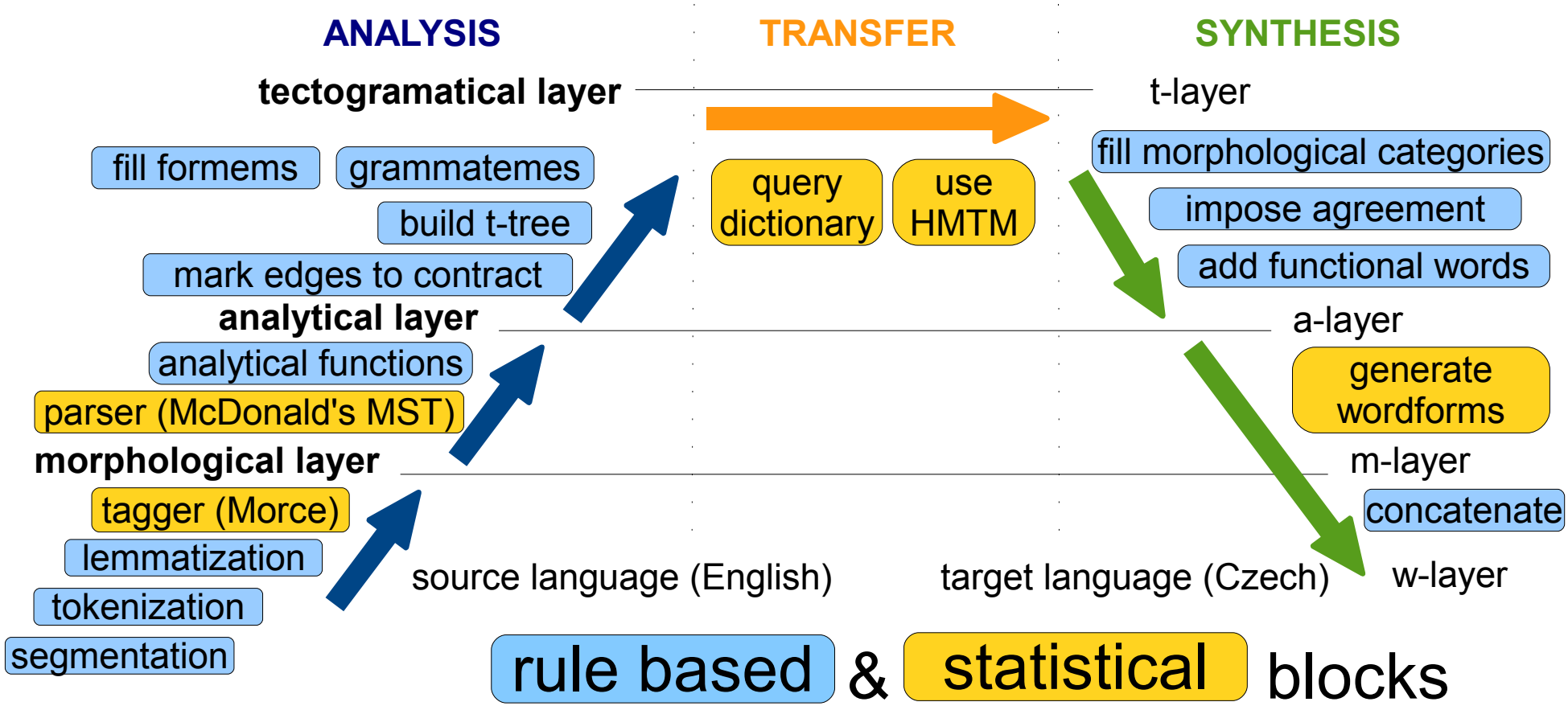
- Čím víc vstup rozeberu, tím snazší by měl být transfer.
  - Rozbor ovšem také není snadný.
  - Navíc čelím kumulaci chyb.
- Pravidlový vs. statistický přístup:
  - Pravidlové systémy píší lingvisté-programátoři.
  - Statistické systémy se naučí samy podle dat.

Překlad přes hloubkovou rovinu TectoMT

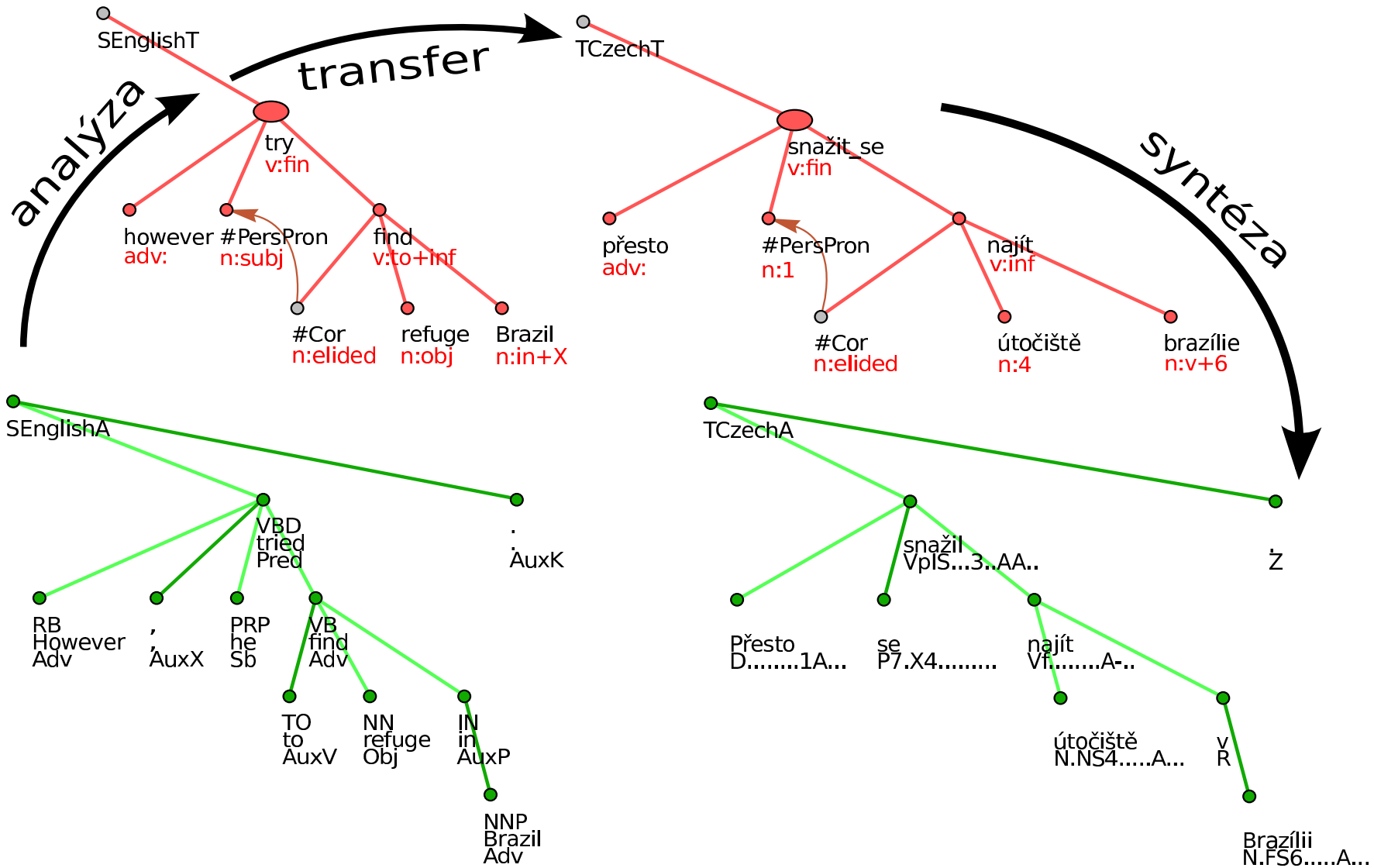
TectoMT: Hloubkový překlad



TectoMT: Hloubkový překlad



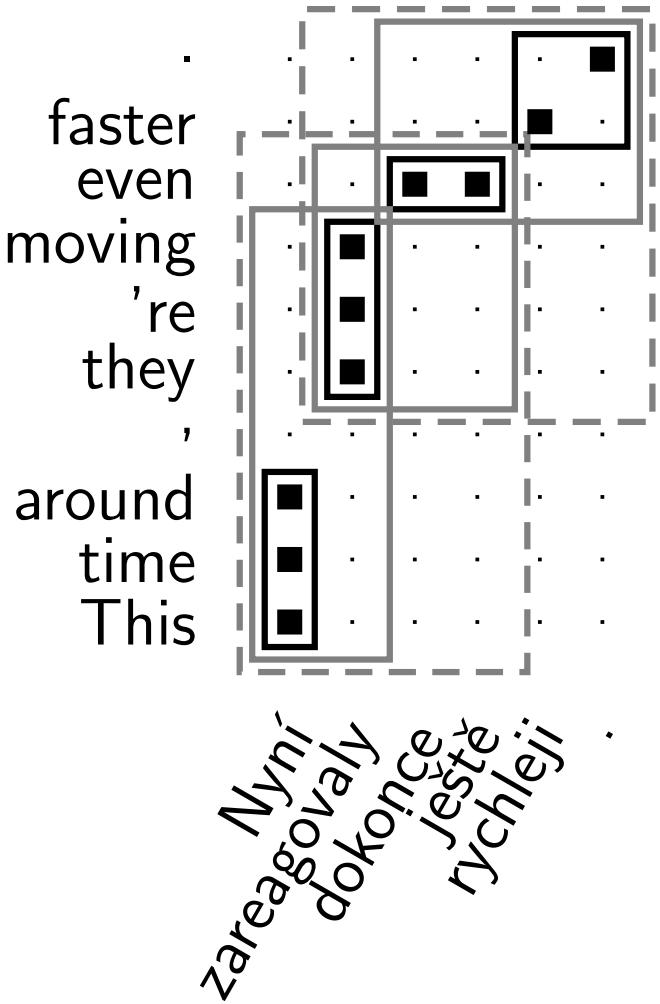
Jádro: Překlad stromu na strom



- Díky t-rovině lze tvar stromu přenést beze změn.

Frázový překlad
Moses (a také Google)

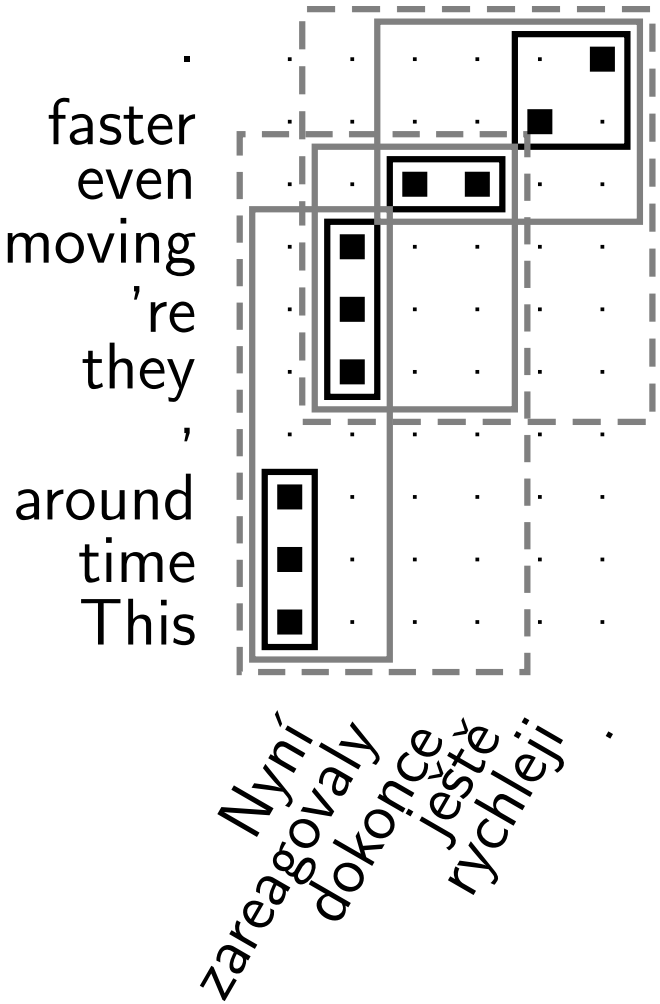
Frázový překlad



Trénovací data:

- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo $\sim$ anglické slovo)

Frázový překlad

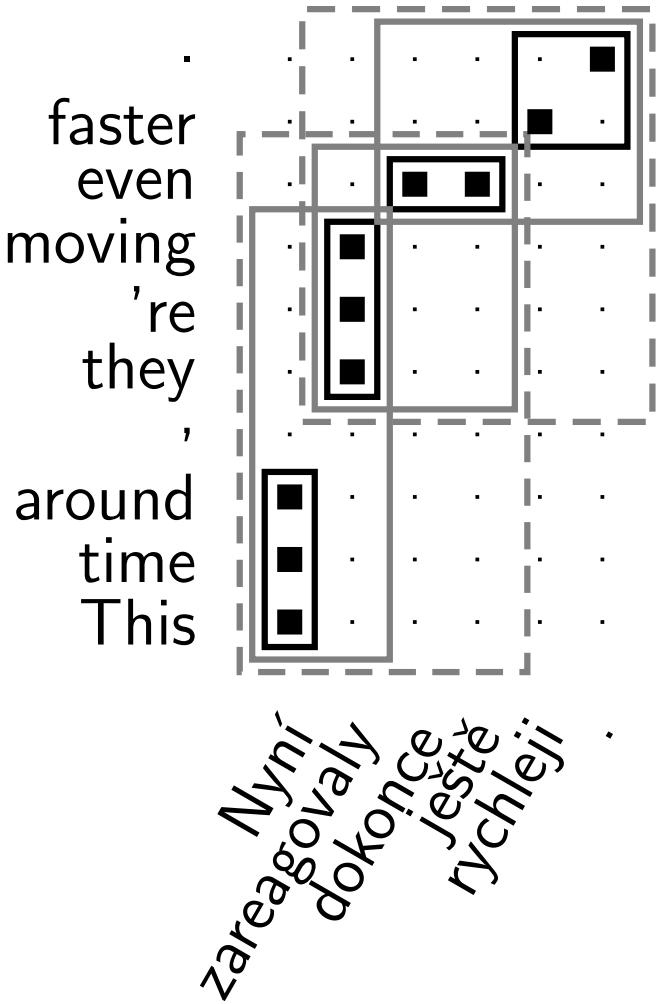


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data:

- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo $\sim$ anglické slovo)

Frázový překlad



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

Trénovací data:

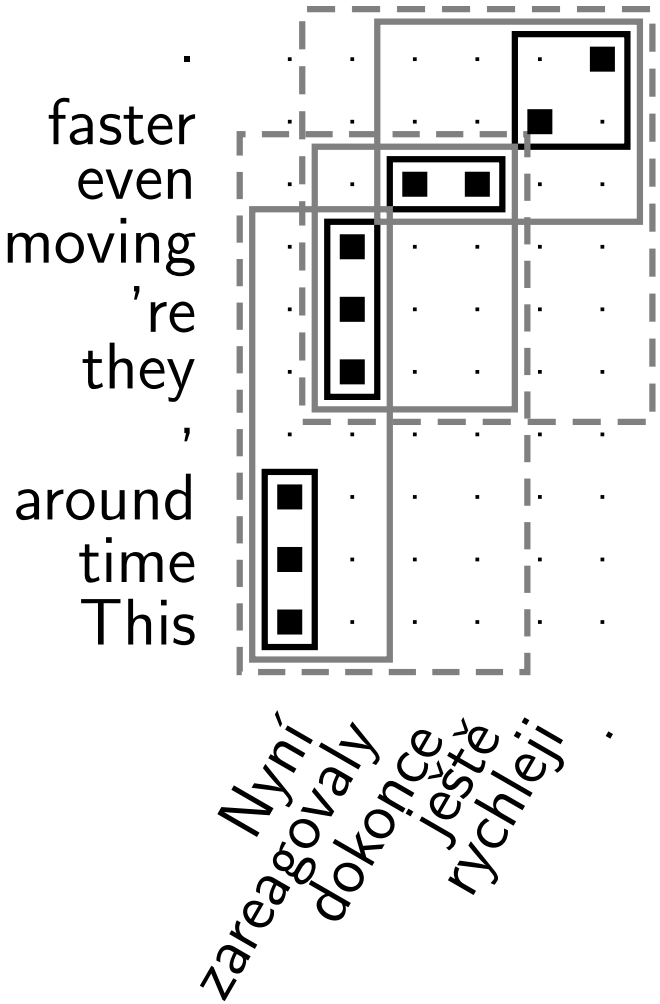
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo)

Při samotném překladu hledáme:

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

aby byl výstup co nejpravděpodobnější.

Frázový překlad



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

Trénovací data:

- paralelní korpus (česká věta = anglická věta) ... 15 mil. párů vět
- automatické zarovnání slov (české slovo ∼ anglické slovo) ∼ 2×200 M

Při samotném překladu hledáme:

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

aby byl výstup co nejpravděpodobnější.

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ Nectí gramatiku, snadno sestaví slovní salát.
- ⊖ Neumí nové tvary slov, a to ani známých.

Natáhnout bačkory. Kick the bucket. ✓

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	√

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	×

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	×

Jan s Marií se vzali.

John and Mary were married. √

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	×

Jan s Marií se vzali.

John and Mary were married. √

Jan s Marií se včera vzali.

John and Mary married yesterday. √

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	×

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. ∼

Výhody a nevýhody frázového překladu

- ⊕ Není třeba žádná znalost jazyka, stačí paralelní data.
- ⊖ 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?	×

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. ×

Problém negace

- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

Problém negace

- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Problém negace

- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:

Nemám žádného psa.	Viděl kočku.
I have no dog.	He saw a cat.

Problém negace

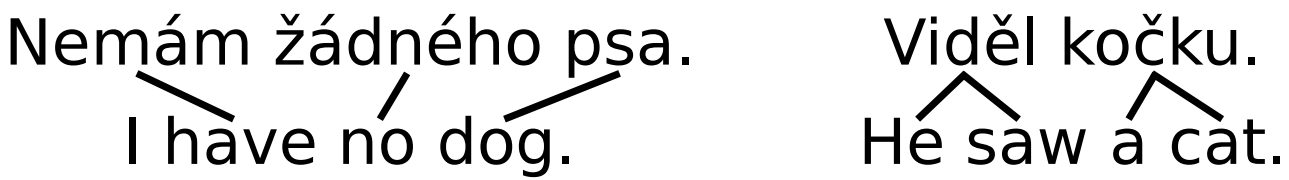
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:



Problém negace

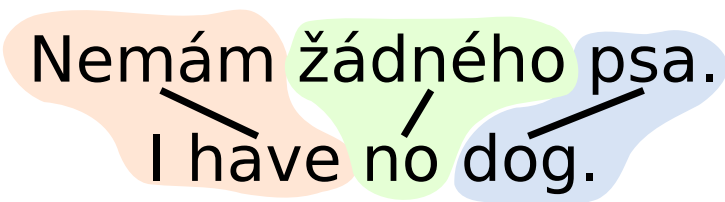
- Francouzská negace je okolo slovesa:

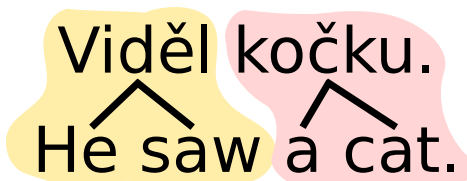
  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:





Problém negace

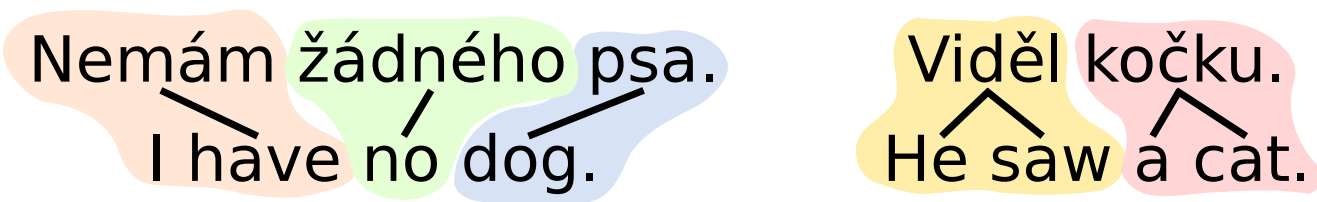
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:



Nový vstup: Nemám kočku.

Problém negace

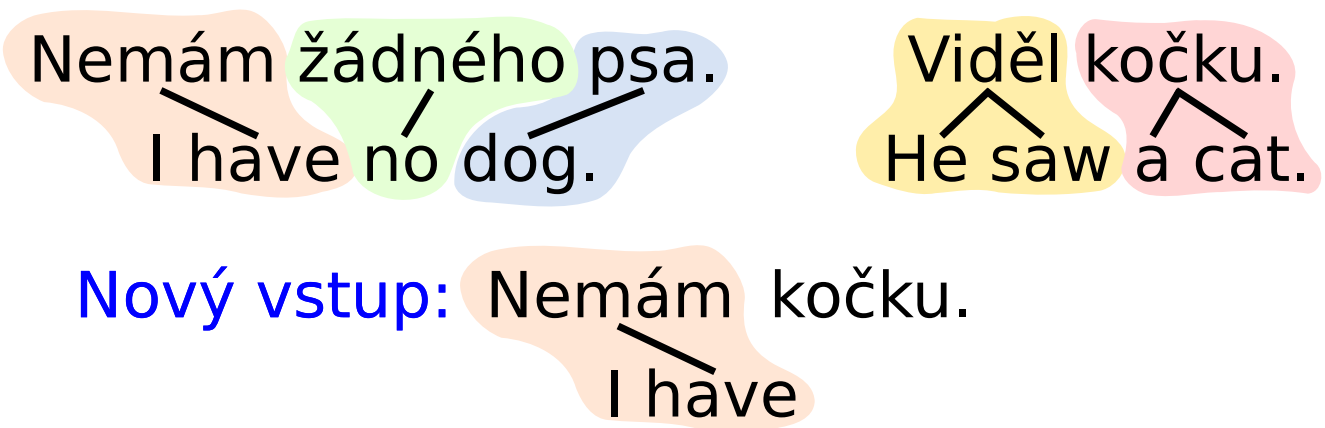
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:



Problém negace

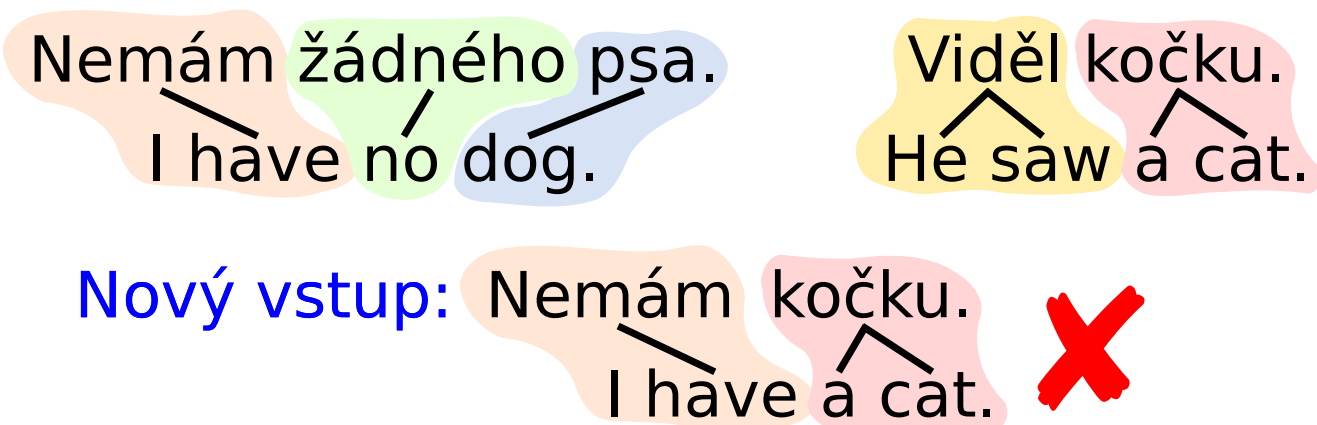
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

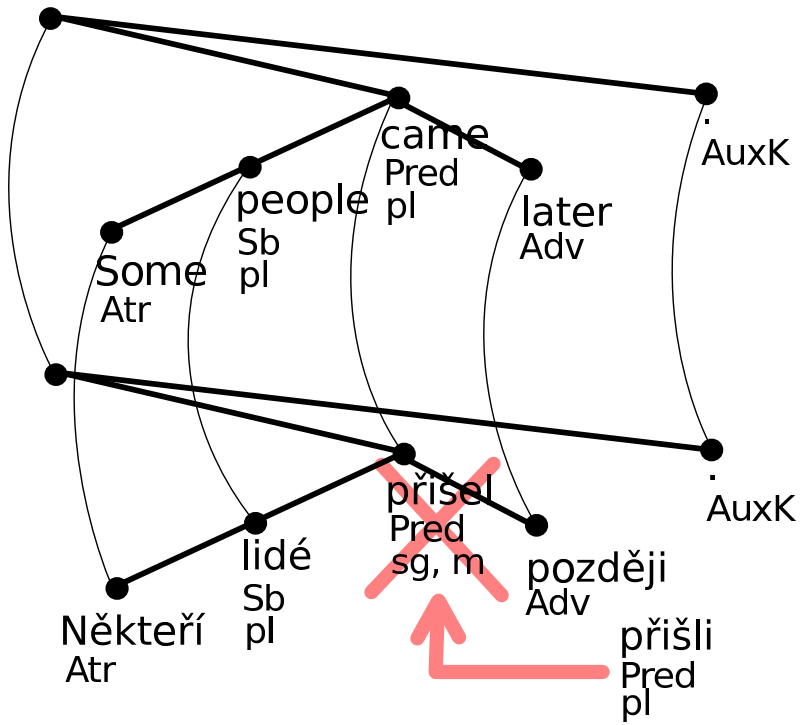
Zdvojená negace vede ke ztrátě negace při překladu:



Oprava negace a gramatiky Depfix

Oprava gramatiky (DEPFIX)

1. Zarovnání vstupu a hypotézy.
2. Větný rozbor vstupu a hypotézy.
3. Pravidla opravující časté chyby:
   - Korekce rozboru hypotézy.
   - Negace, gramatické shody, pády po předložce...



... 50–60 % změněných vět změněno k lepšímu.

... přesnost vrácení ztracené negace: 90 %

Frázový vs. syntaktický překlad

	Stell dir das vor.	
Google	Imagine that.	√
Systran	Imagine.	√

Frázový vs. syntaktický překlad

	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 překlad

	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 překlad

	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.	✓

I syntaktický překlad má své limity

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

I syntaktický překlad má své limity

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

Stell dir ein Haus vor.
⇒ Imagine a house. ✓

I syntaktický překlad má své limity

- 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. ✓

I syntaktický překlad má své limity

- 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. ✗

I syntaktický překlad má své limity

- 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. ×

- A u pravidlových systémů stačí negramatický vstup:

Stell dir ein Haus, das ⊘ Garten hat, vor.
⇒ Place to you a house, the garden intends. ×

Frázový vs. syntaktický překlad: souhrn

Frázový překlad volí primitivní řešení:

- Větu nerozebírá, jen opisuje známé posloupnosti slov.
- Spoléhá na dostatek dat.
- Často produkuje negramatické věty, rád zahodí negaci.

Syntaktický překlad:

- Výstup zahrnuje větný rozbor $\Rightarrow$ naděje gramatičnosti.
- Naráží na chyby v kaskádě nástrojů (zejm. analýza).
- Naráží na „negramatický“ vstup.

$\Rightarrow$ V průměru zatím funguje lépe frázový překlad.
$\Rightarrow$ Syntaktický překlad má potenciál řešit těžší problémy.
$\Rightarrow$ Nejlepší je přístupy kombinovat.

Počítačová lingvistika: Na hranici oborů

Kontrola překlepů. Kontrola pravopisu.
Vyhledávání dokumentů (na webu). Sumarizace textů.
Syntéza a rozpoznávání řeči. Dialogové systémy.
Strojový překlad (mluvené řeči).

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.

⇒ Lze využít v programech (pravidlových i statistických).

Proč studovat na MFF a ÚFALu

Můžete se naučit mj.:

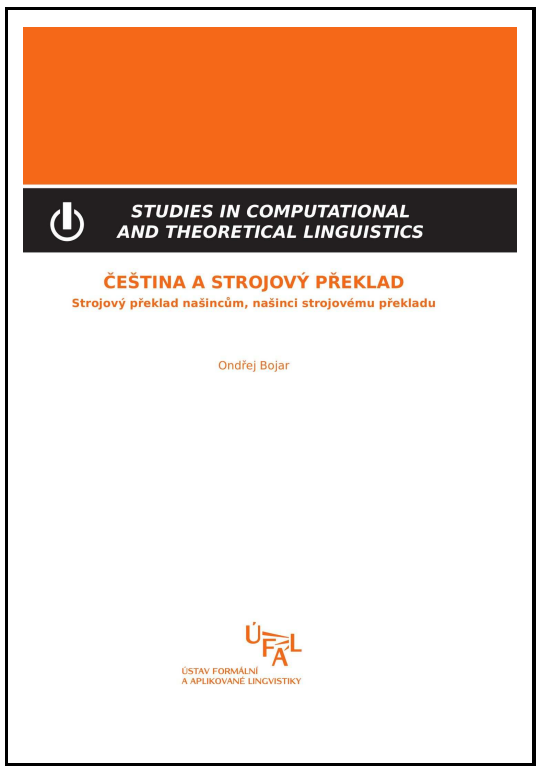
- Modelovat, jak lidé (myslí a) pracují s textem, řečí, ...
- Rozdělit složité úlohy na částečky a přispět částečkami,
- Počítat, abyste našli jehly v kupkách i horách sena (Pravděpodobnost a statistika),
- Navrhovat datové struktury, abyste zvládli terabajty dat,

  Text na českém webu $\sim$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 $>$600 CPU, počítače s 32 GB až 0.5 TB RAM.
- Soutěžit na mezinárodní úrovni v překladu, sumarizaci...

Chcete-li vědět víc

Navštivte stánek Lingvistika v 1. patře.

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



Knížka *Čeština a strojový překlad: Strojový překlad našincům, našinci strojovému překladu*. 168 pp. Knihkupectví Karolinum (Celetná 18; eshop)