Jak pracuje automatický překlad

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

ELRC Workshop, 14. duben 2016

Osnova

- Úloha strojového překladu.
- Obtížnost překladu.
- Přístupy ke strojovému překladu.
  -  Hloubkový překlad.
  -  Frázový překlad
  -  Automatické opravy chyb.
- Shrnutí.

Úloha strojového překladu

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。

Úloha strojového překladu

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.

Úloha strojového překladu

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.

Úloha strojového překladu

Úloha strojového překladu

Jakou kvalitu lze očekávat

- Srozumitelné, s hrubými chybami.
- Vhodné k úpravám (posteditaci).
- Vhodné k vydání.

Jakou kvalitu lze očekávat

- Srozumitelné, s hrubými chybami.
  - Google Translate, Microsoft Bing, ...
  - ÚFAL Moses:
    `http://lindat.mff.cuni.cz/services/moses/`
- Vhodné k úpravám (posteditaci).
- Vhodné k vydání.

Jakou kvalitu lze očekávat

- Srozumitelné, s hrubými chybami.
  - Google Translate, Microsoft Bing, ...
  - ÚFAL Moses:
    `http://lindat.mff.cuni.cz/services/moses/`
- Vhodné k úpravám (posteditaci).
  - Množství nástrojů pro podporu překladu (CAT).
  - Strojový překlad výrazně obohacuje překladovou paměť.
- Vhodné k vydání.

Jakou kvalitu lze očekávat

- Srozumitelné, s hrubými chybami.
  - Google Translate, Microsoft Bing, ...
  - ÚFAL Moses:
    `http://lindat.mff.cuni.cz/services/moses/`
- Vhodné k úpravám (posteditaci).
  - Množství nástrojů pro podporu překladu (CAT).
  - Strojový překlad výrazně obohacuje překladovou paměť.
- Vhodné k vydání.
  - I dnes jen velmi úzké oblasti (předpověď počasí).

Chyby dělají i lidští překladatelé

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

Chyby dělají i lidští překladatelé

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

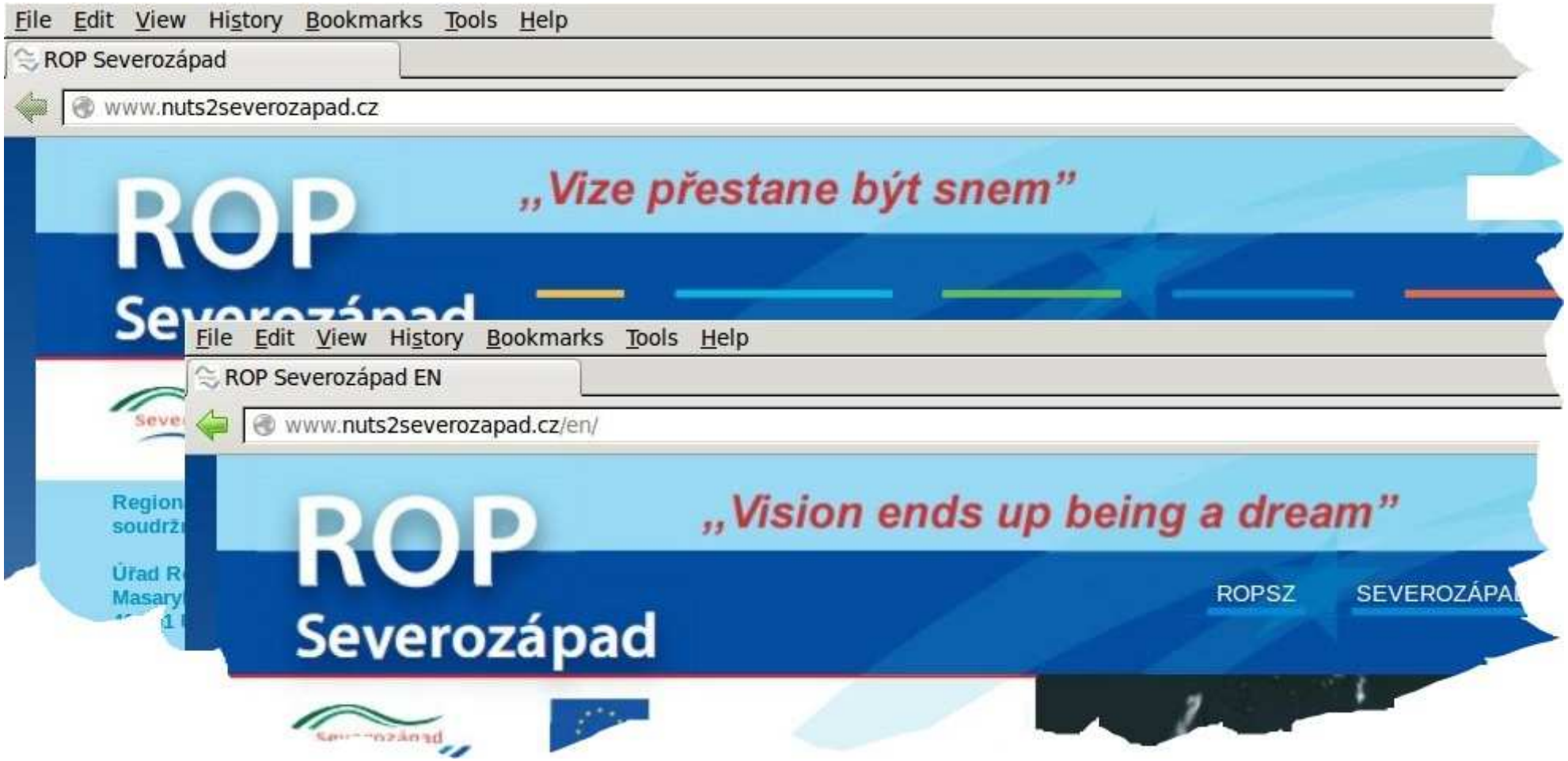
The Fentimans Logo is a based on Fearless,
Thomas Fentiman's prize Alsatian,
double winner of the Crufts Obedience Test.

Chyby dělají i lidští překladatelé

Základem tohoto loga je Nebojsa, postava Alsasana
získaná Thomasem 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.

Chyby dělají i lidští překladatelé



. . . nuts2severozapad.cz . . . Vize skončí jako sen.

Proč je překlad těžký

Na vstupu víceznačnost všeho druhu:

The plant is next to the bank.

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 **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 **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 **velvety coat rack**.
. . . sametová police na kabáty?
. . . police na sametové kabáty?

Z češtiny to není lepší:

Spal celou Petkevičovu přednášku.
Ženu holí stroj.

Proč je překlad těžký

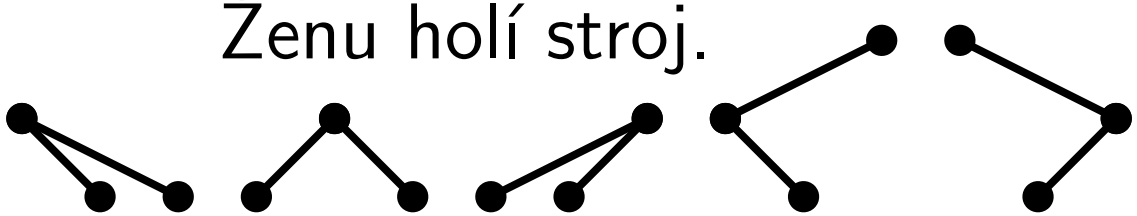
Na vstupu víceznačnost všeho druhu:

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

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

Při překladu se možnosti násobí

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			...	...		

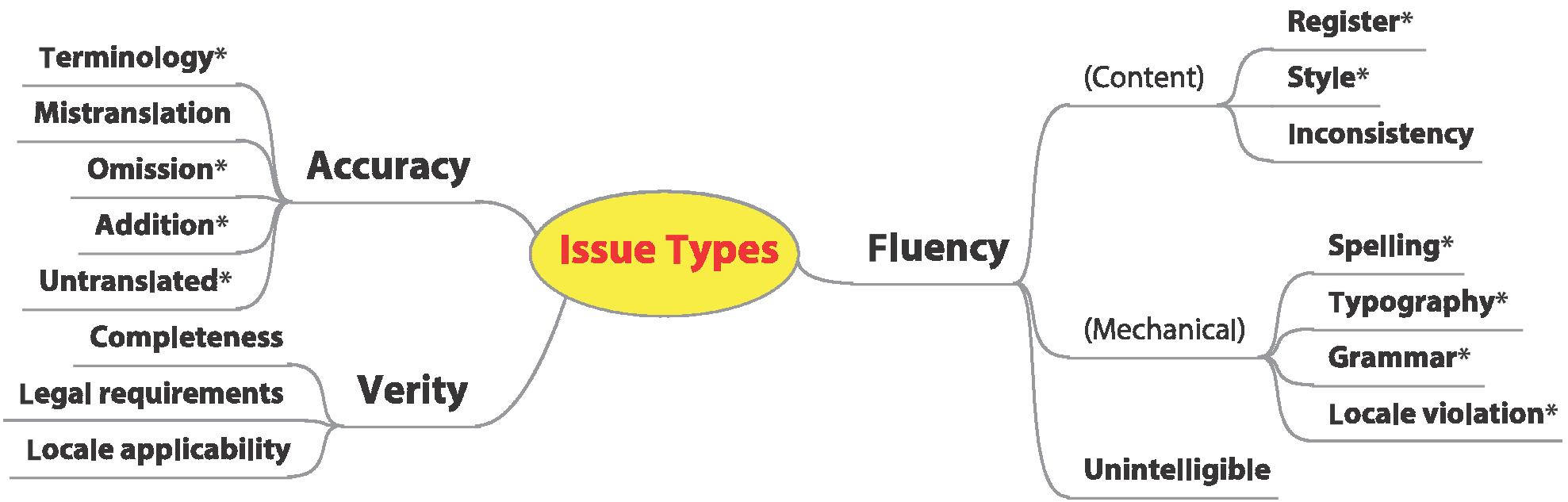
Do češtiny navíc musíme trefit tvar

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			...	...		

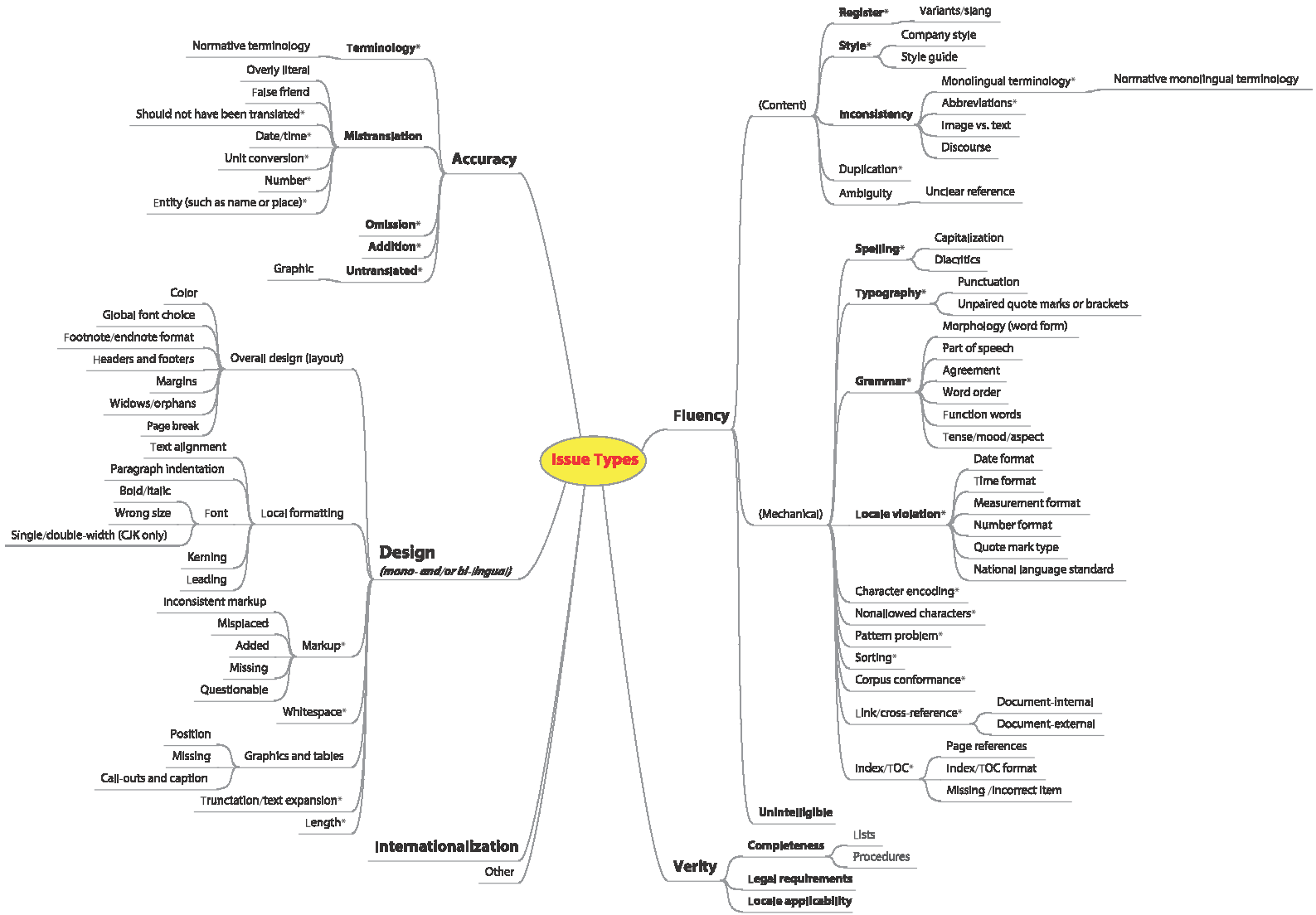
. . . ale tvar závisí na kontextu. . .

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**		
			zelenému	pruhovanému		
zrak mi utkvěl na			zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

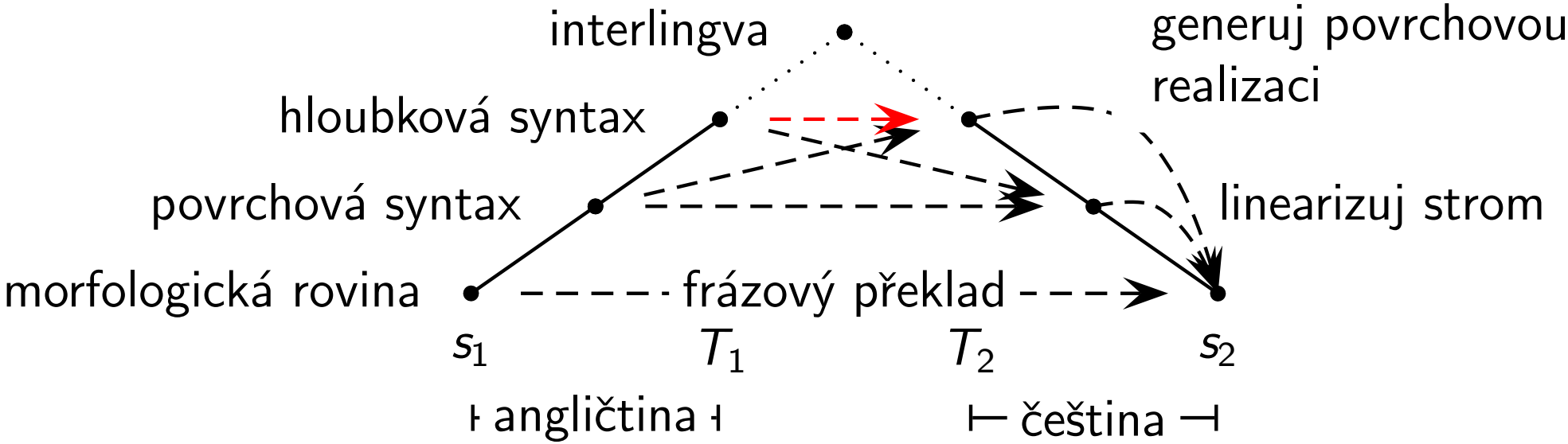
Co všechno může být špatně



Co všechno může být špatně



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.

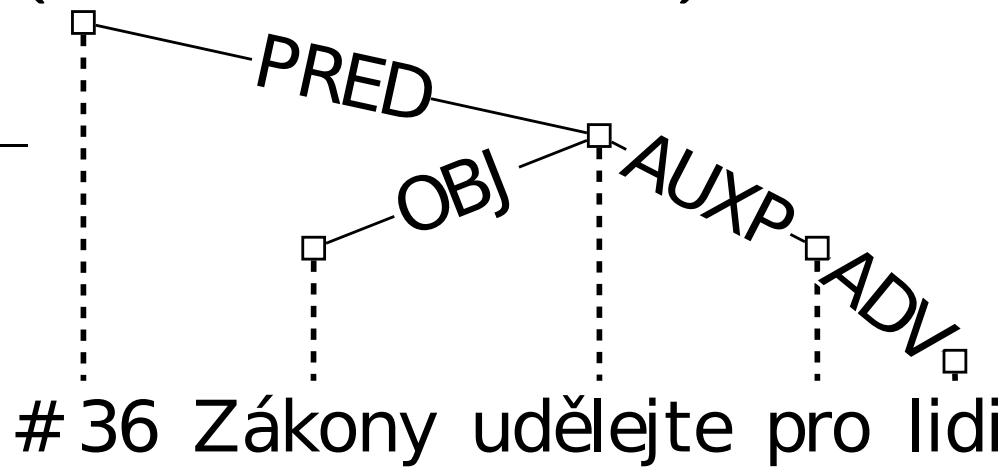
Překlad přes hloubkovou rovinu TectoMT

Formální popis češtiny

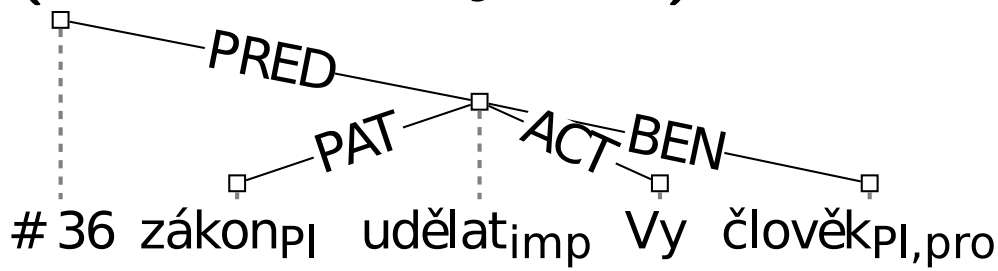
Morfologická rovina:

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----

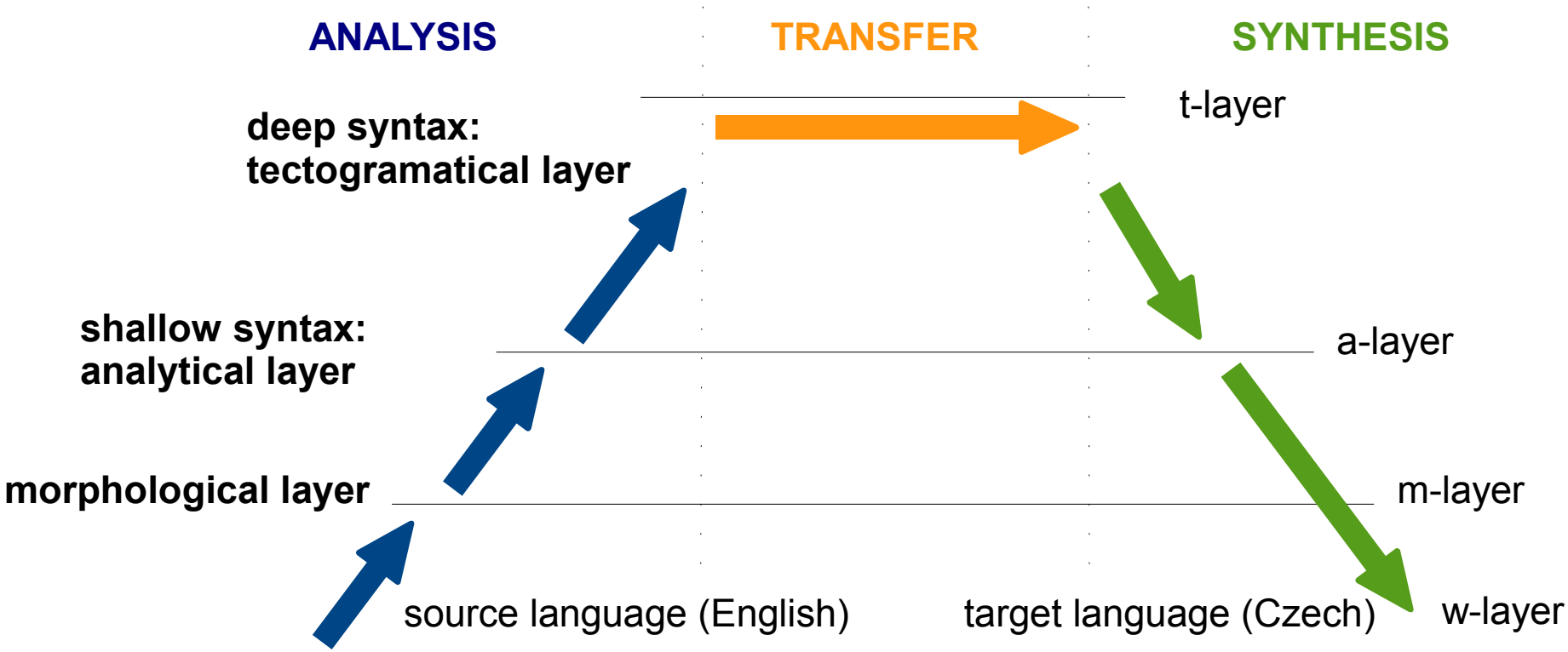
Analytická rovina (povrchová syntax):



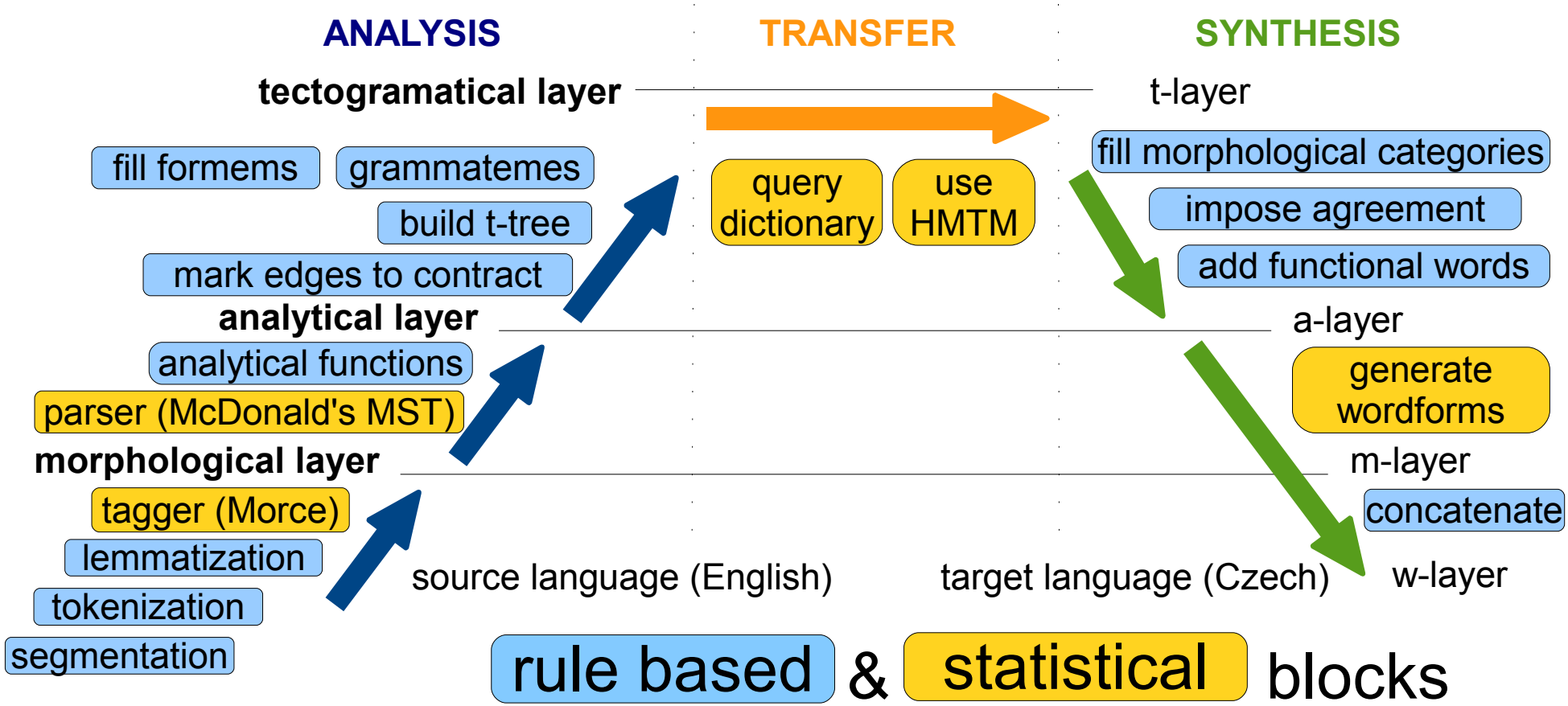
Tektogramatická rovina (hloubková syntax):



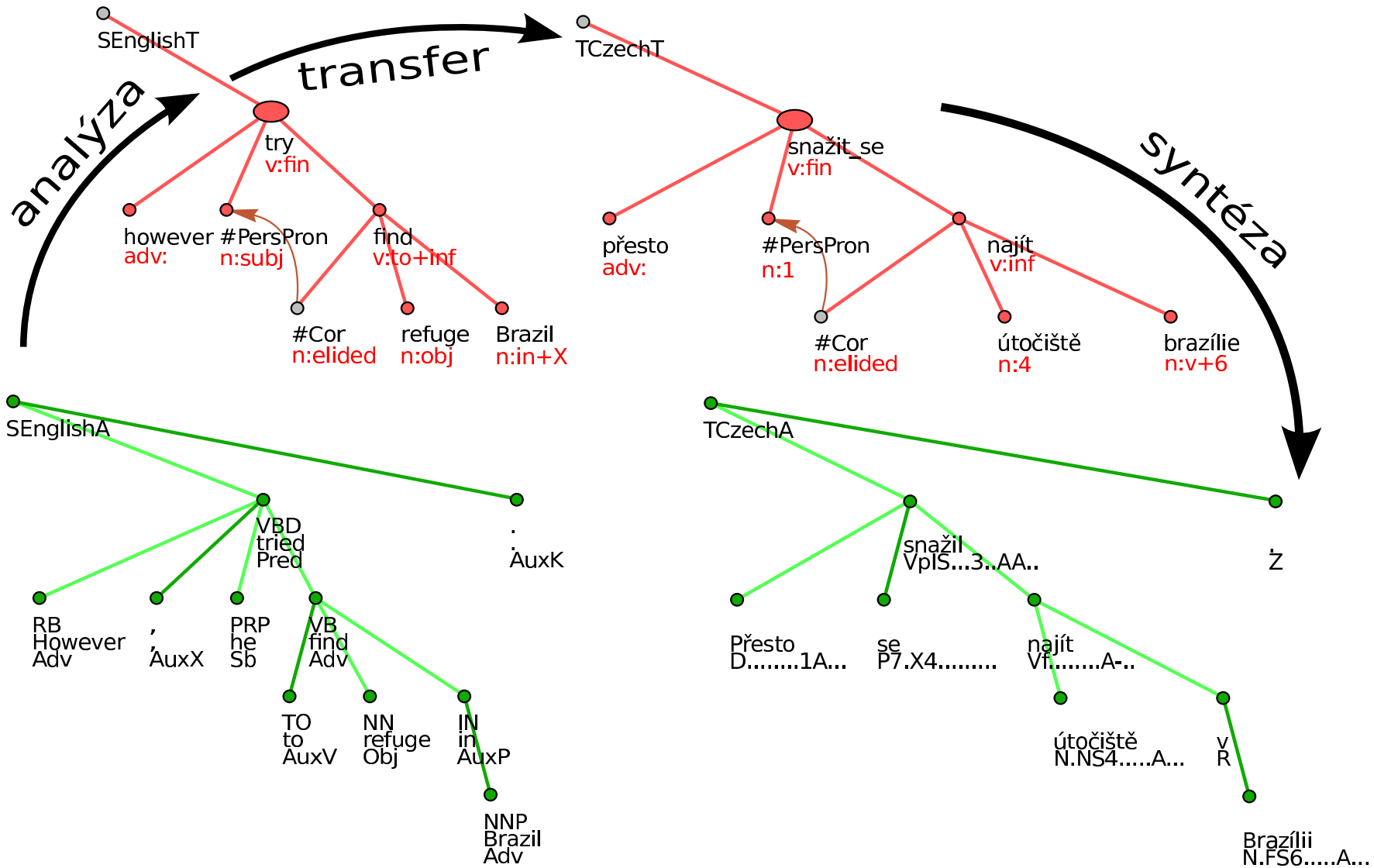
TectoMT: Hloubkový překlad



TectoMT: Hloubkový překlad



Jádro: Překlad stromu na strom



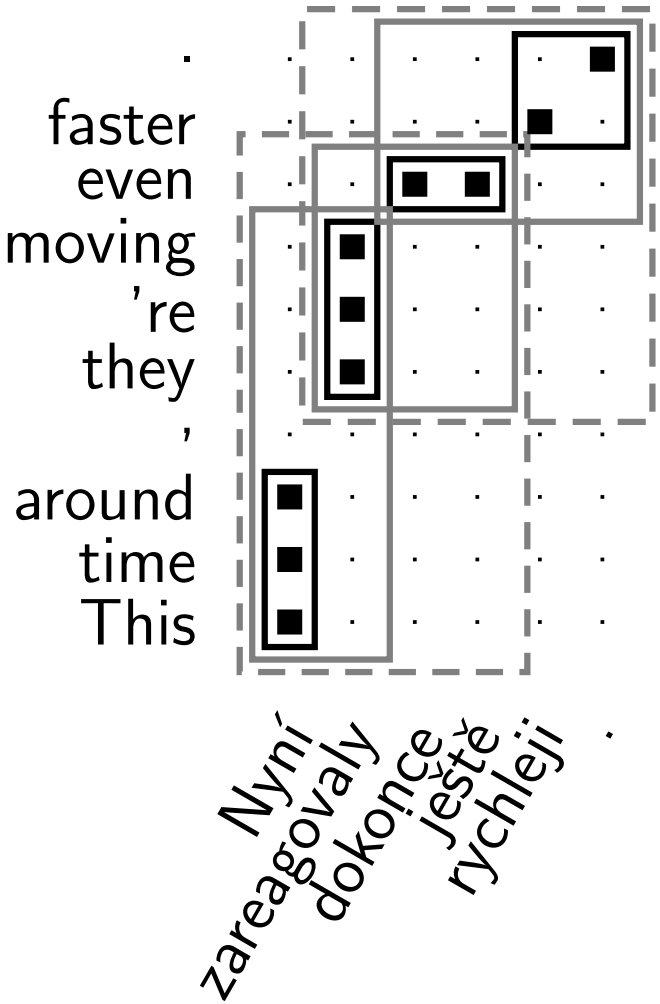
- Díky t-rovině lze tvar stromu přenést beze změn.

Frázový překlad
Moses (a také Google)

MOSES CORE

Frázový překlad
Moses (a také MT@EC)

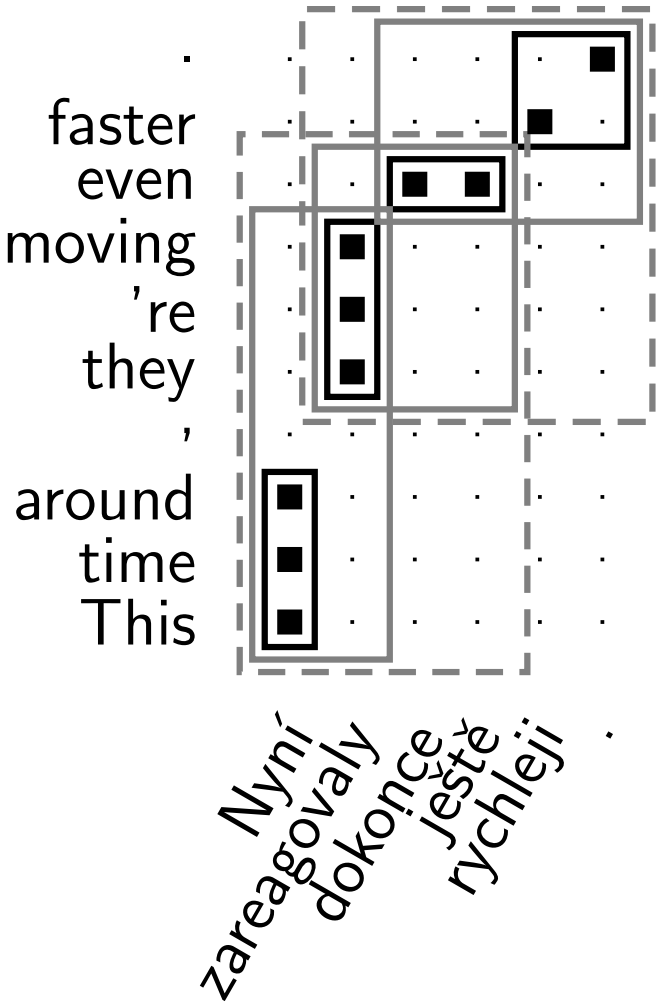
Frázový překlad



Trénovací data:

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

Frázový překlad

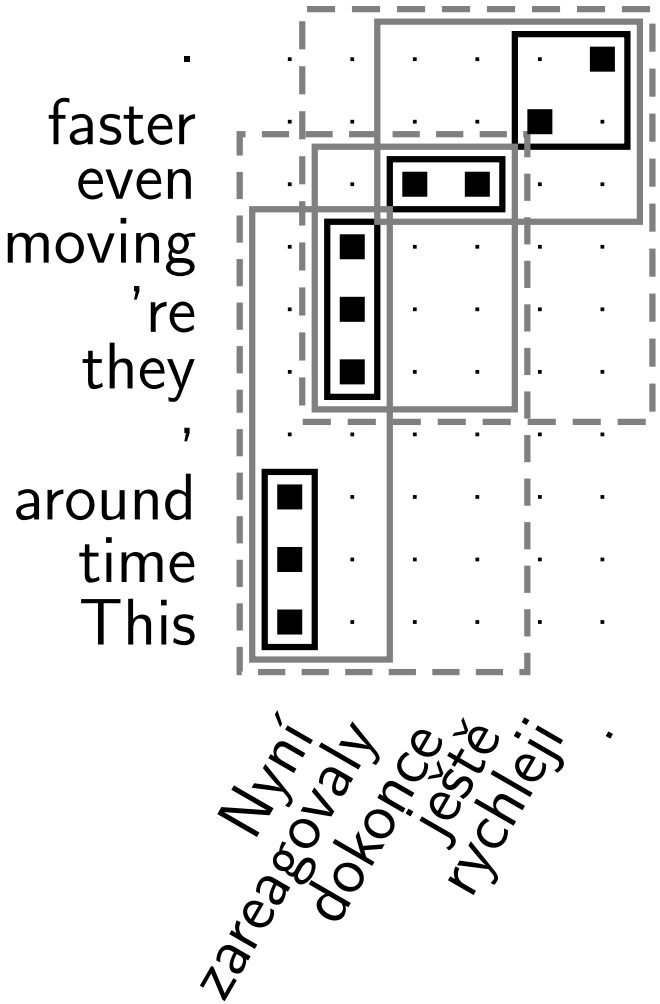


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
even faster	=	dokonce ještě rychleji
...	=	...

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:

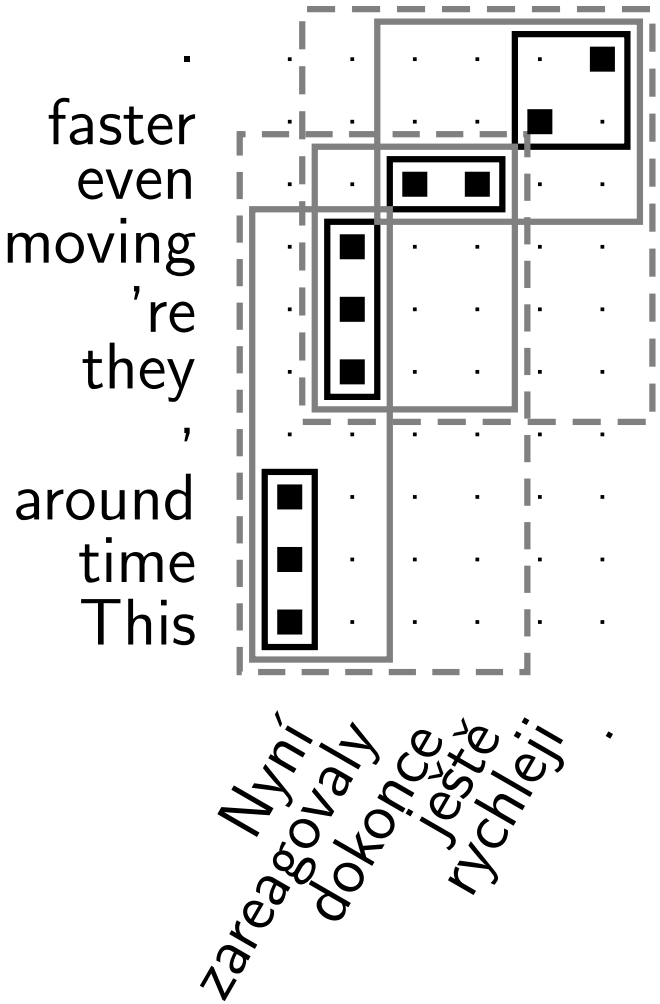
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo)

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.
- ⊕ 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. ✓

Výhody a nevýhody frázového překladu

- ⊕ 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?	√

Výhody a nevýhody frázového překladu

- ⊕ 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?	✗

Výhody a nevýhody frázového překladu

- ⊕ 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?	×

Výhody a nevýhody frázového překladu

- ⊕ 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.	×

Nachytat překlad na švestkách...

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

Jan s Marií se vzali.

John and Mary were married. ✓

Nachytat překlad 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. ✓

Nachytat překlad 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. ∼

Nachytat překlad 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. ×

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

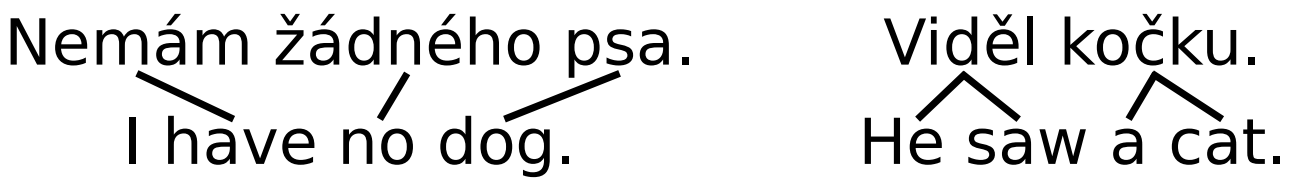
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:



Problém negace

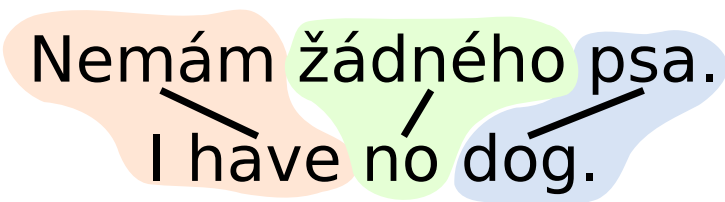
- Francouzská negace je okolo slovesa:

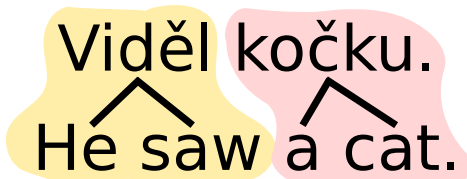
  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:





Problém negace

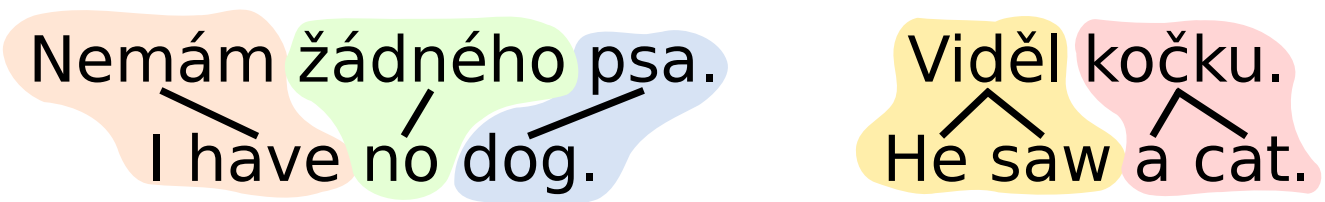
- Francouzská negace je okolo slovesa:

  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

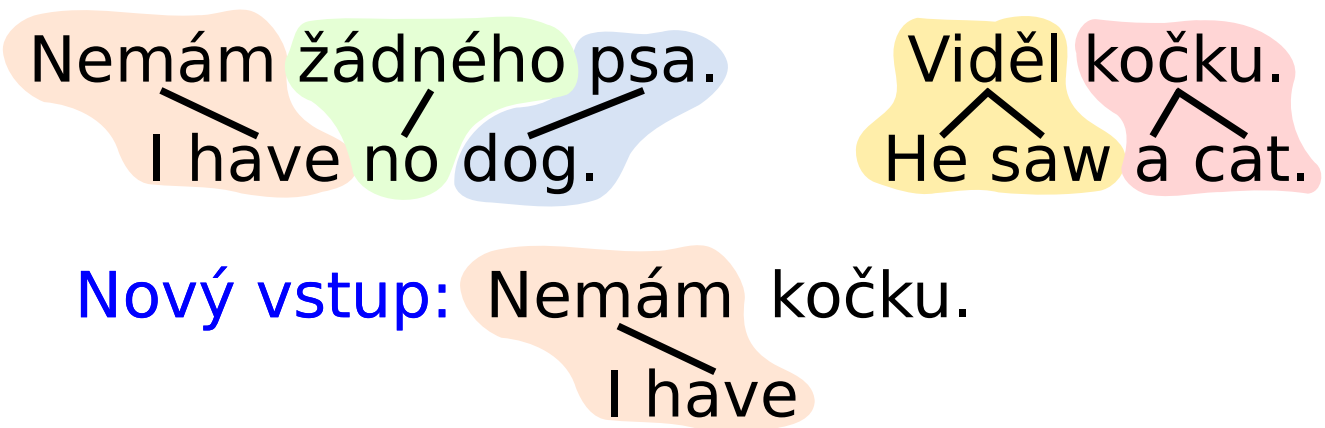
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace vede ke ztrátě negace při překladu:



Problém negace

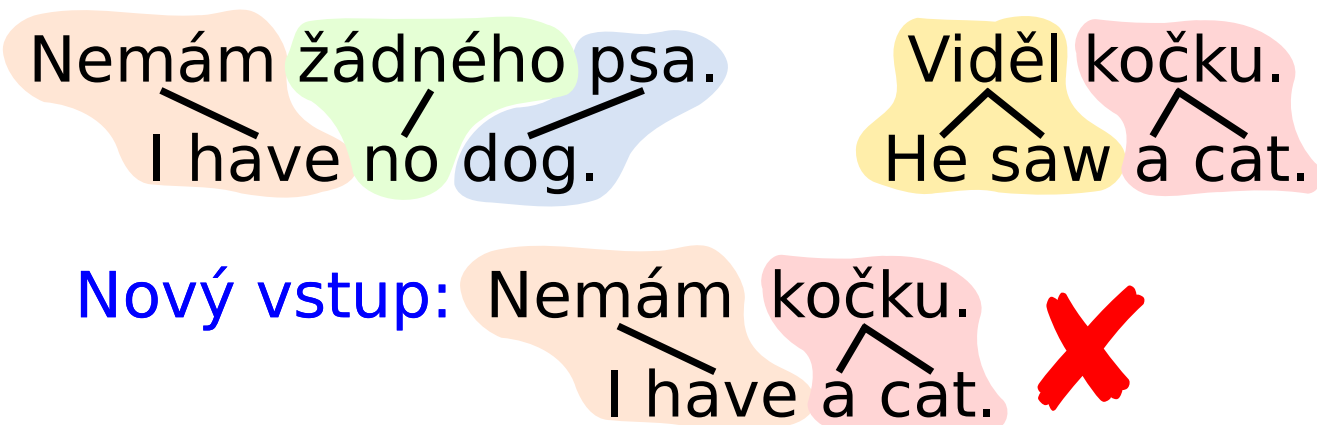
- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

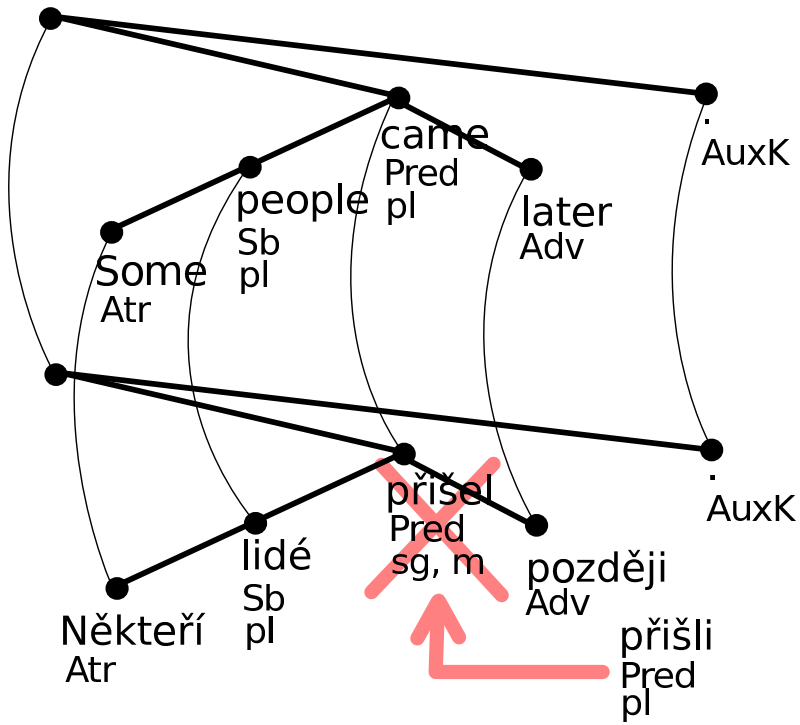
Zdvojená negace vede ke ztrátě negace při překladu:



Oprava negace a gramatiky Depfix

Oprava gramatiky (DEPFIX)

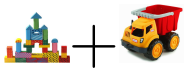
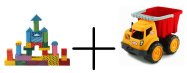
1. Zarovnání vstupu a hypotézy.
2. Větný rozbor vstupu a hypotézy.
3. Pravidla opravující časté chyby:
   - Korekce rozboru hypotézy.
   - Negace, gramatické shody, pády po předložce...



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

... přesnost vrácení ztracené negace: 90 %

Nejlepší je kombinace

	System	BLEU	TER	Manual
WMT13		20.0	**0.693**	**0.664**
		**20.1**	0.696	0.637
		19.5	0.713	–
	Google Translate	18.9	0.720	0.618
		14.7	0.741	0.455
WMT14		21.1	0.670	**0.373**
	uedin-unconstr.	**21.6**	**0.667**	0.357
		20.9	0.674	0.333
	Google Translate	20.2	0.687	0.168
		15.2	0.716	-0.177
WMT15		**18.8**	**0.715**	**0.686**
		18.7	0.717	–
		17.6	0.730	–
	Google Translate	16.4	0.750	0.515
		13.4	0.763	0.209

Náš vítězný systém: Chiméra

- Vstup:
  - Famous cases also relate to graphic elements.
-  Hloubkový překlad přeloží pomocí větného rozboru:
  - Slavné případy se být týkají grafické prvky.
-  Frázový překlad přidá 200M en-cs a 3,6G cs slov:
  - Slavné případy se týkají také grafické prvky.
-  Automatické opravy chyb jako shoda či negace:
  - Slavné případy se týkají také grafických prvků.

(Google: Slavné případy týkat i grafické prvky.)

Shrnutí

- Nejlepší výsledek: kombinace přístupů.
  -  Hloubkový +  frázový +  korektura.
- Hloubkový překlad náročný na lingvistické zdroje.
- Frázový překlad potřebuje jen paralelní data.
- Chyby v současných modelech zatím nevyhnutelné (negace, . . .)
- Při dostatku dat výstup vhodný k posteditaci.