Strojový překlad

. . . jako informované hádání.

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

Mixér Výzkumného záměru, 23. 11. 2011

Osnova

- Překlad jako informované hádání.
- Dva přístupy k překladu.
- Který přístup je tedy lepší?
- Dva způsoby kombinování překladu.
- Souhrn.

Zpola informované hádání

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?

Zpola informované hádání

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?

Zpola informované hádání

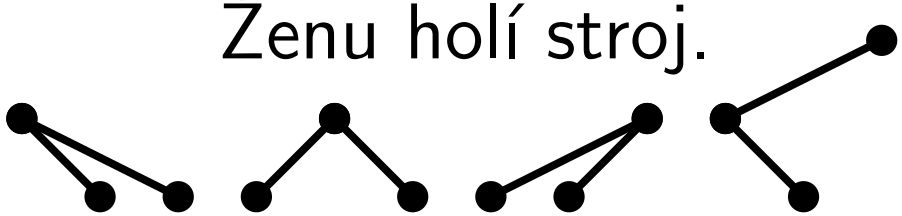
Na vstupu víceznačnost všeho druhu:

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.



Zpola informované hádání

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.

Bez hádání to asi nepůjde

Ryze pravidlové systémy:

- Často narážejí na chybějící hesla ve svých slovnících:
  - 32 % anglických a 28 % německých vět nelze zpracovat (Nicholson et al., 2008).

„Vyškrtávací“ přístupy nechávají příliš mnoho variant:

- Věta o n slovech má teoreticky $O(n^n)$ rozborů (stromů).
  - Bojar (2004): Povolíme-li jen pozorované lokální konfigurace, dostaneme 9^n struktur.
  - Kovář et al. (2008): Ruční gramatika dává pro větu průměrně 10^{14} analýz, medián 240.

Použitelné v kontrole gramatiky nebo předzpracování pro tagger: (Petkevič, 2006; Spoustová et al., 2007).

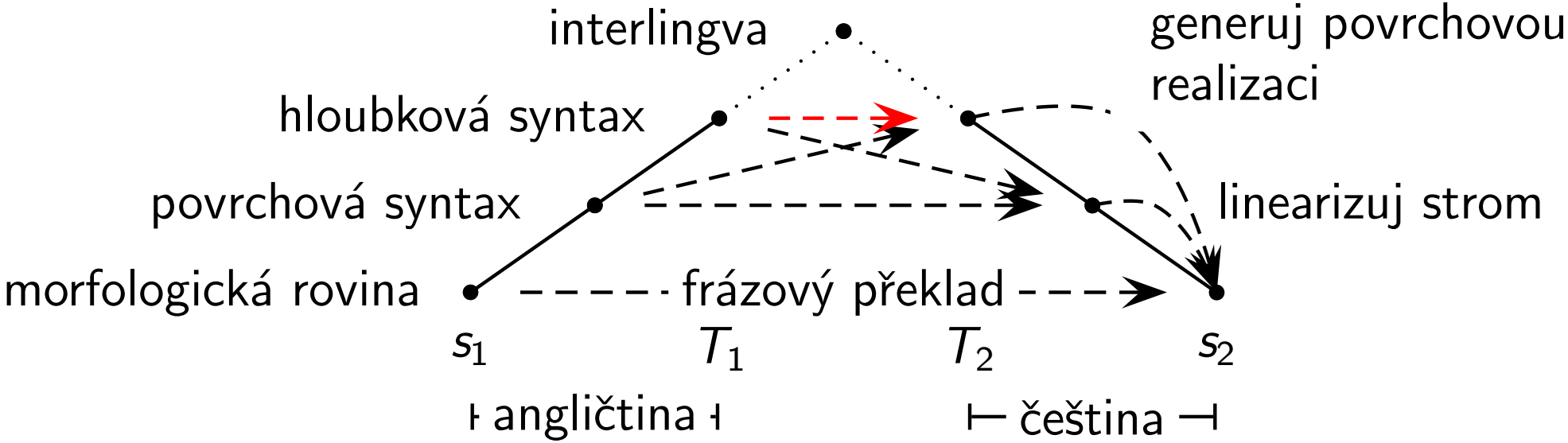
Do češtiny navíc musíme trefit tvar

Čeština má:

- pro podst./příd./... jména: 7 pádů, 4 rody, 3 čísla, ...
- pro slovesa: rod, číslo, čas, způsob, vid, ...

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			...	...		

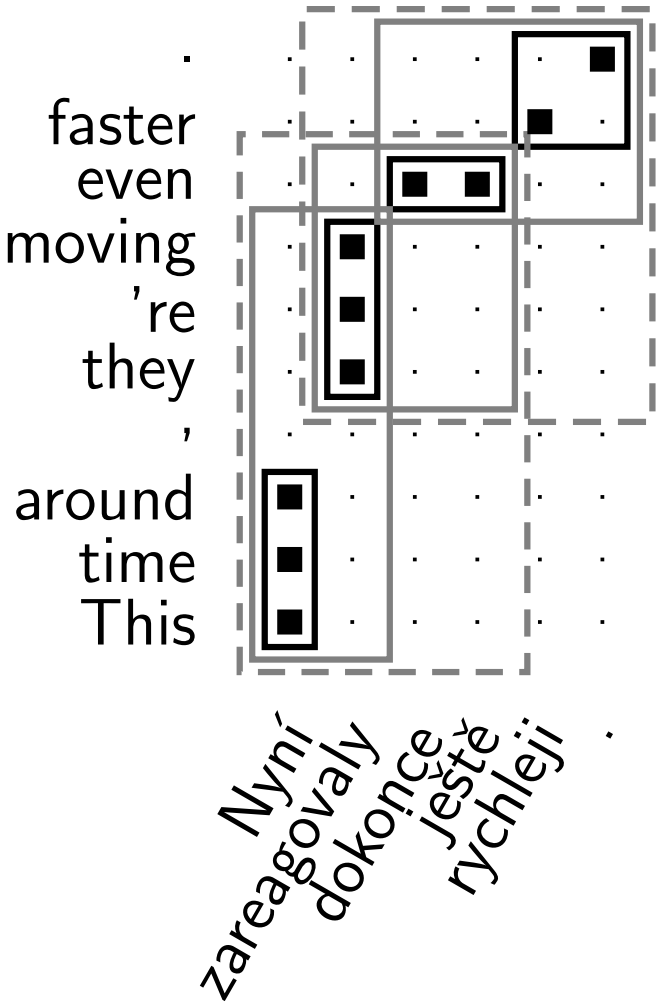
Přístupy ke strojovému překladu



$$\hat{s}_2 = \operatorname*{argmax}_{s_2} p(s_2 \mid s_1) = \left\langle \begin{array}{l} \operatorname*{argmax}_{s_2, T_1, T_2} p(T_1 \mid s_1) \cdot p(T_2 \mid T_1) \cdot p(s_2 \mid T_2) \\ \operatorname*{argmax}_{s_2} p(s_1 \mid s_2) \cdot p(s_2) \end{array} \right.$$

- Čím víc vstup rozeberu, tím snazší by měl být transfer.
  - Rozbor ovšem také není snadný.
  - Navíc čelím kumulaci chyb.

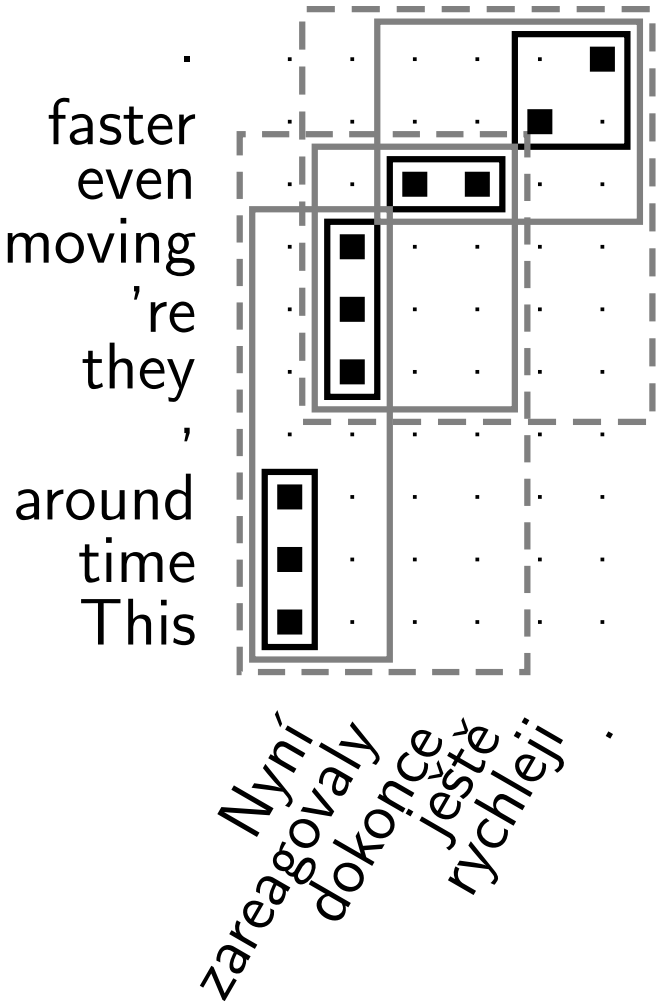
Frázový překlad...



Trénovací data:

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

Frázový překlad...

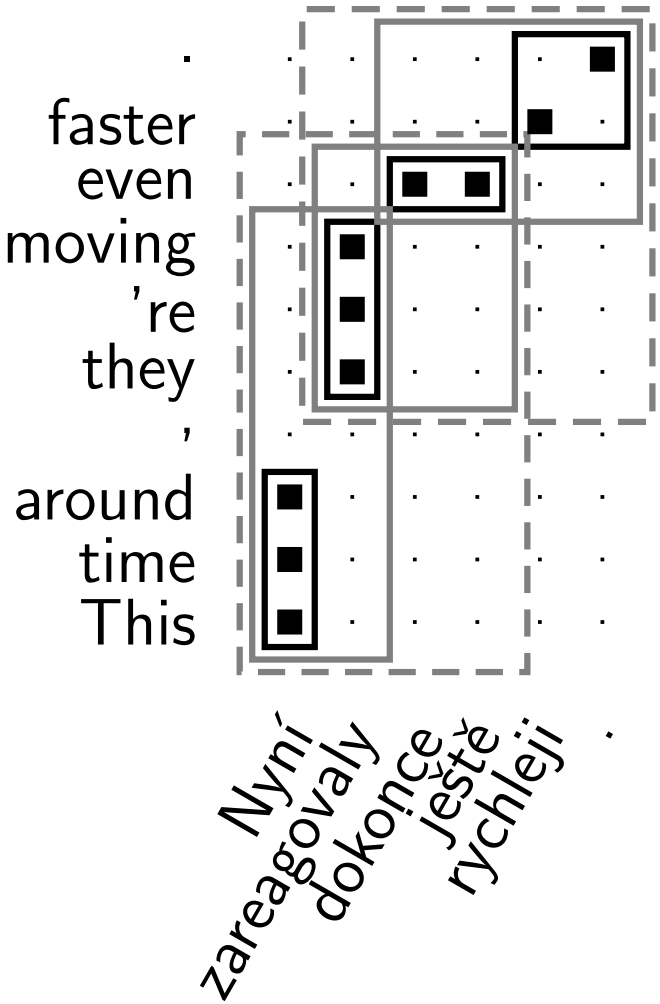


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
even faster	=	dokonce ještě rychleji
...	=	...

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:

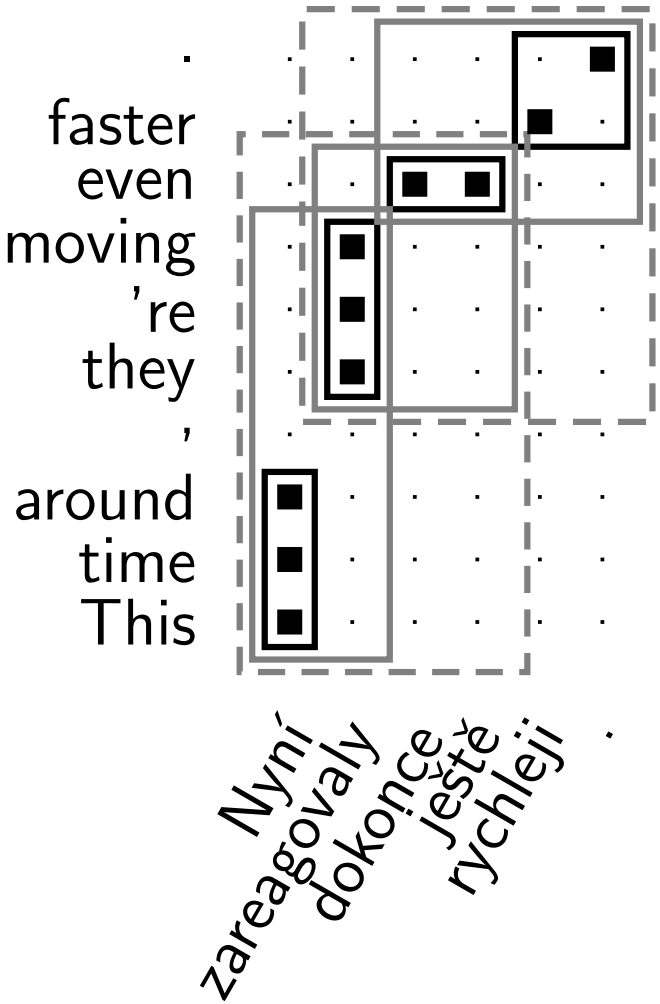
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo $\sim$ anglické slovo)

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) ... 9 mil. párů vět
- automatické zarovnání slov (české slovo ∼ anglické slovo) ∼ 2×90 M

Při samotném překladu hledáme:

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

aby byl výstup co nejpravděpodobnější.

...nachytat na švestkách?

Natáhnout bačkory.

Kick the bucket. ✓

...nachytat na švestkách?

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓

...nachytat na švestkách

Natáhnout bačkory.	Kick the bucket.	✓
Proč musel natáhnout bačkory?	Why did he kick the bucket?	✓
Proč natáhl bačkory?	Why stretched slippers?	×

...nachytat na švestkách

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. ✓

...nachytat na švestkách

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. ✓

...nachytat na švestkách

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. ∼

...nachytat na švestkách

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. ×

...nachytat na švestkách

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.

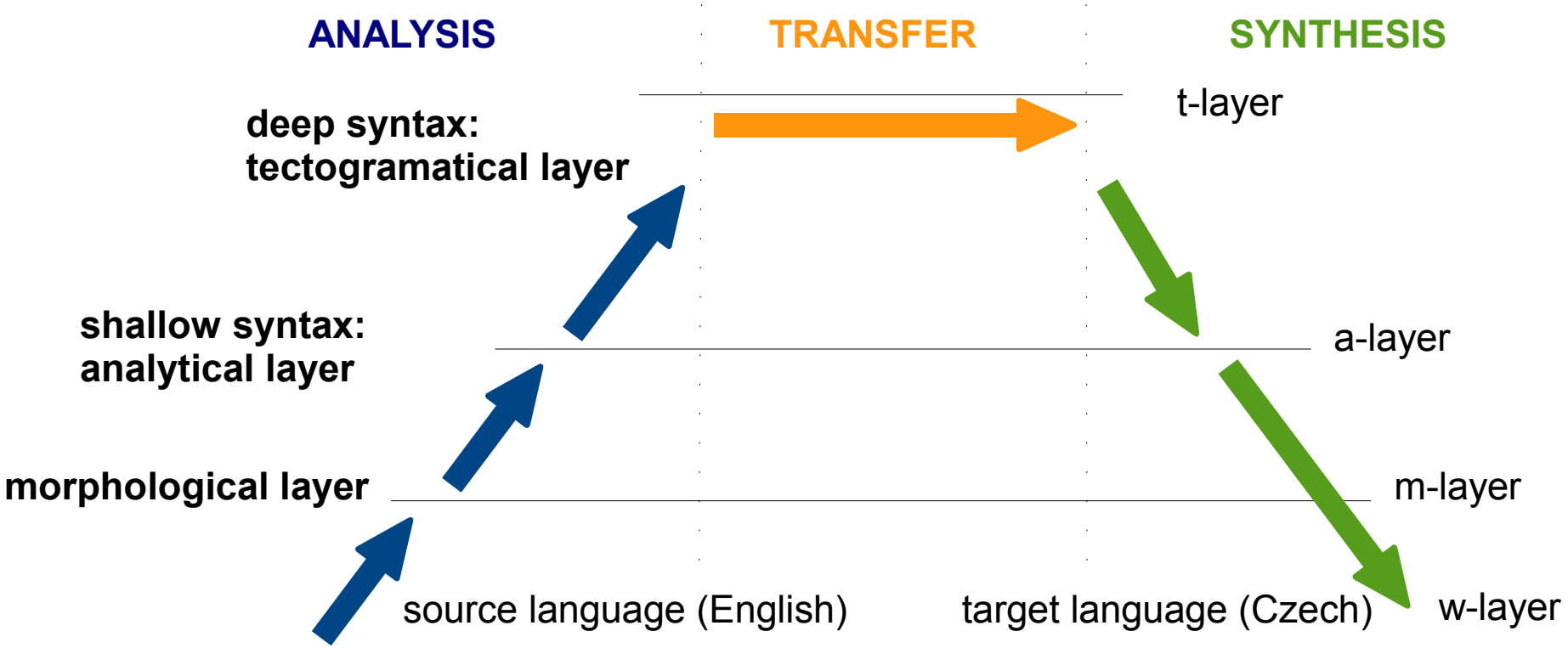
John and Mary are married in church yesterday. ∼

Jan s Marií se včera v kostele svatého Ducha vzali.

John and Mary yesterday in the Church of the Holy Spirit took. ×

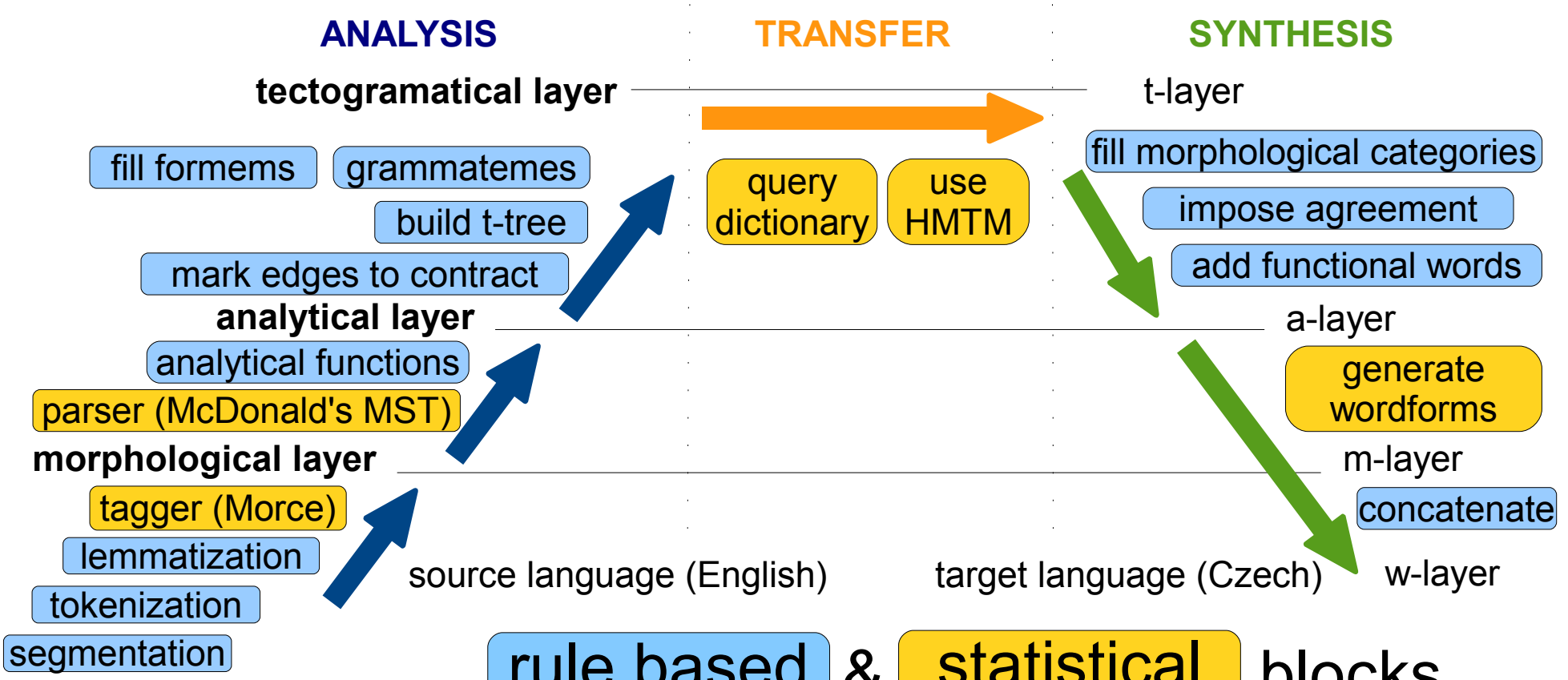
...zkusme tedy překlad dělat pořádně.

Hloubkový překlad: TectoMT



- TectoMT = MT z bloků platformy Treex (Perl; CPAN).
- Transfer na tektogramatické rovině (∼PDT).

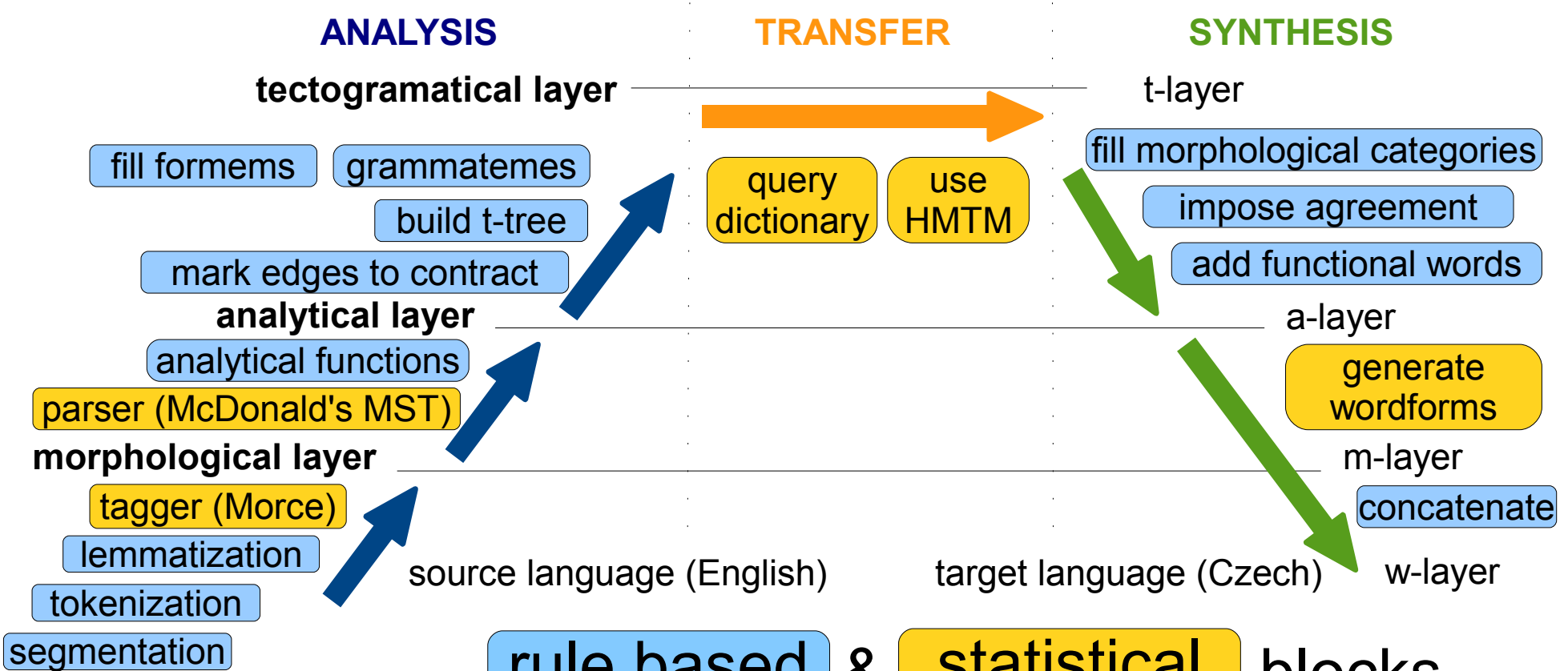
Hloubkový překlad: TectoMT



rule based & statistical blocks

- TectoMT = MT z bloků platformy Treex (Perl; CPAN).
- Transfer na tektogramatické rovině (∼PDT).

Hloubkový překlad: TectoMT



rule based & statistical blocks

- TectoMT = MT z bloků platformy Treex (Perl; CPAN).
- Transfer na tektogramatické rovině (∼PDT).
- Statistické bloky uvnitř uvažují mnoho možností, mezi bloky je předáván výstup jen jeden.

Jak nachytat syntaktický překlad

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

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. ×

- A u pravidlových systémů 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 Frázový	ÚFAL Hloubkový	Komerční Google	Komerční PC Translator
Seřaďte hypotézy od nejlepší po nejhorší. Shody povoleny.				
WMT10 > ostatní	**45**	44	49	**49**
WMT10 >= ostatní	**66**	60	**70**	62
WMT11 > ostatní	39	**40**	**44**	34
WMT11 >= ostatní	**64**	58	**65**	51
WMT10: Člověk zkusil výstup MT opravit bez znalosti originálu.				
Je to dobrý překlad? (%)	**40**	34	**55**	43
MT přeložil krátký text. Dokážete zodpovědět kontrolní otázky?				
% správných odpovědí	73.6	**80.6**	78.7	80.2

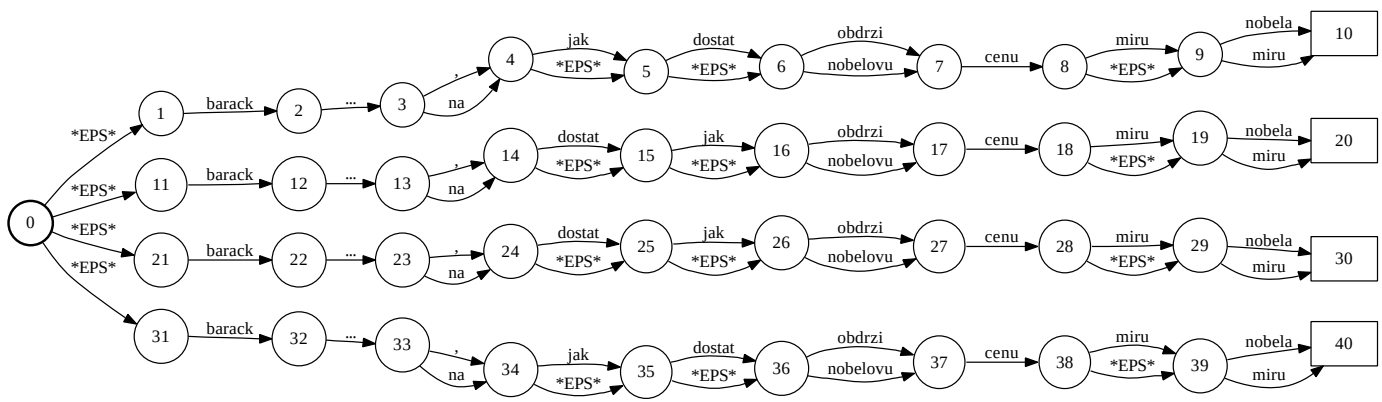
- Cílené porovnání systémů ÚFALu (63 vět):
  - 3 z 6 lidí radši hloubkový, 3 radši frázový.
- Pravidelné soutěže: `http://www.statmt.org/wmt11/`.

Kombinace výstupů zvaná „Rover“

- Systémy hlasují, která jednotlivá slova do výstupu dát.
- Pořadí slov dle jednoho ze syst. (volí se v každé větě).

barack	...	,	ϵ	ϵ	obdrží	cenu	míru	nobela
barack	...	na	ϵ	ϵ	nobelovu	cenu	míru	ϵ
barack	...	,	dostat	jak	nobelovu	cenu	míru	ϵ
barack	...	na	ϵ	ϵ	nobelovu	cenu	míru	ϵ

- Formalizováno jako nejlevnější cesta grafem:
  - Váhy hran vyjadřují počet hlasů, jazykový model, ...



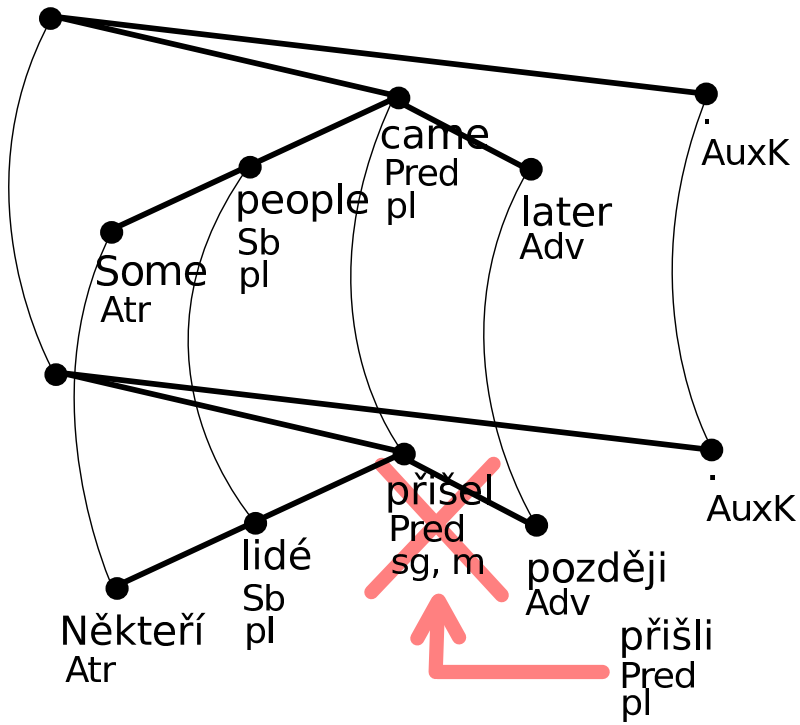
- Kombinace vždy lepší než nejlepší ze systémů.

Oprava gramatiky (DEPFIX)

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

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

... opět platforma Treex.



Shrnutí

- Překlad je zpola informované hádání.
  - Zábavné hřiště pro informatiky i statistiky.
- Frázový vs. (hloubkově) syntaktický přístup k překladu.
  - Žádný není obecně lepší.
- I kombinování hypotéz lze dělat různě.

- Čeština je zajímavá a lingvisticky dobře popsaná.
  - Naše výsledky a data jsou inspirací i jinde.

References

Ondřej Bojar. 2004. Czech Syntactic Analysis Constraint-Based, XDG: One Possible Start. Prague Bulletin of Mathematical Linguistics, 81:43–54.

Vojtěch Kovář, Aleš Horák, and Vladimír Kadlec. 2008. New methods for pruning and ordering ofÂ syntaxÂ parsingÂ trees. In Petr Sojka, Aleš Horák, Ivan Kopecek, and Karel Pala, editors, Text, Speech and Dialogue, volume 5246 of Lecture Notes in Computer Science, pages 125–131. Springer Berlin / Heidelberg.

Jeremy Nicholson, Valia Kordoni, Yi Zhang, Timothy Baldwin, and Rebecca Dridan. 2008. Evaluating and Extending the Coverage of HPSG Grammars: A Case Study for German. In European Language Resources Association (ELRA), editor, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may.

Vladimír Petkevič. 2006. Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In Mária Šimková, editor, Insight into Slovak and Czech Corpus Linguistics, pages 26–44, Bratislava, Slovakia. Veda, vydavateľstvo SAV.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007, pages 67–74, Praha.