Naučíme počítač (cizí) řeči?

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Science Café Kladno

Osnova

- 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.

Na hranici oborů...

čeština,
angličtina, němčina ...

matematika

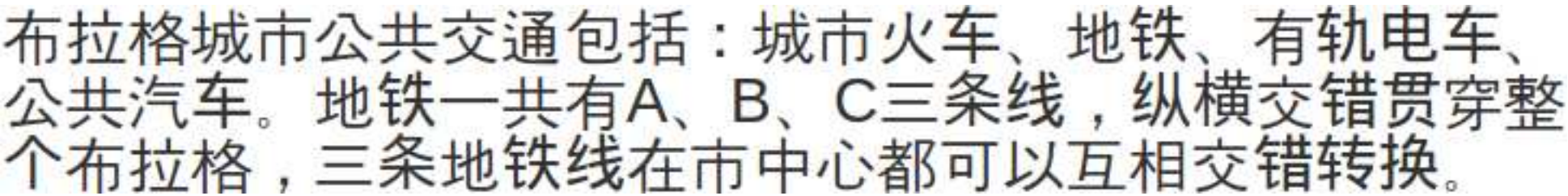
počítače

Počítačová lingvistika

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。

Optimismus na začátek



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.

Optimismus na začátek

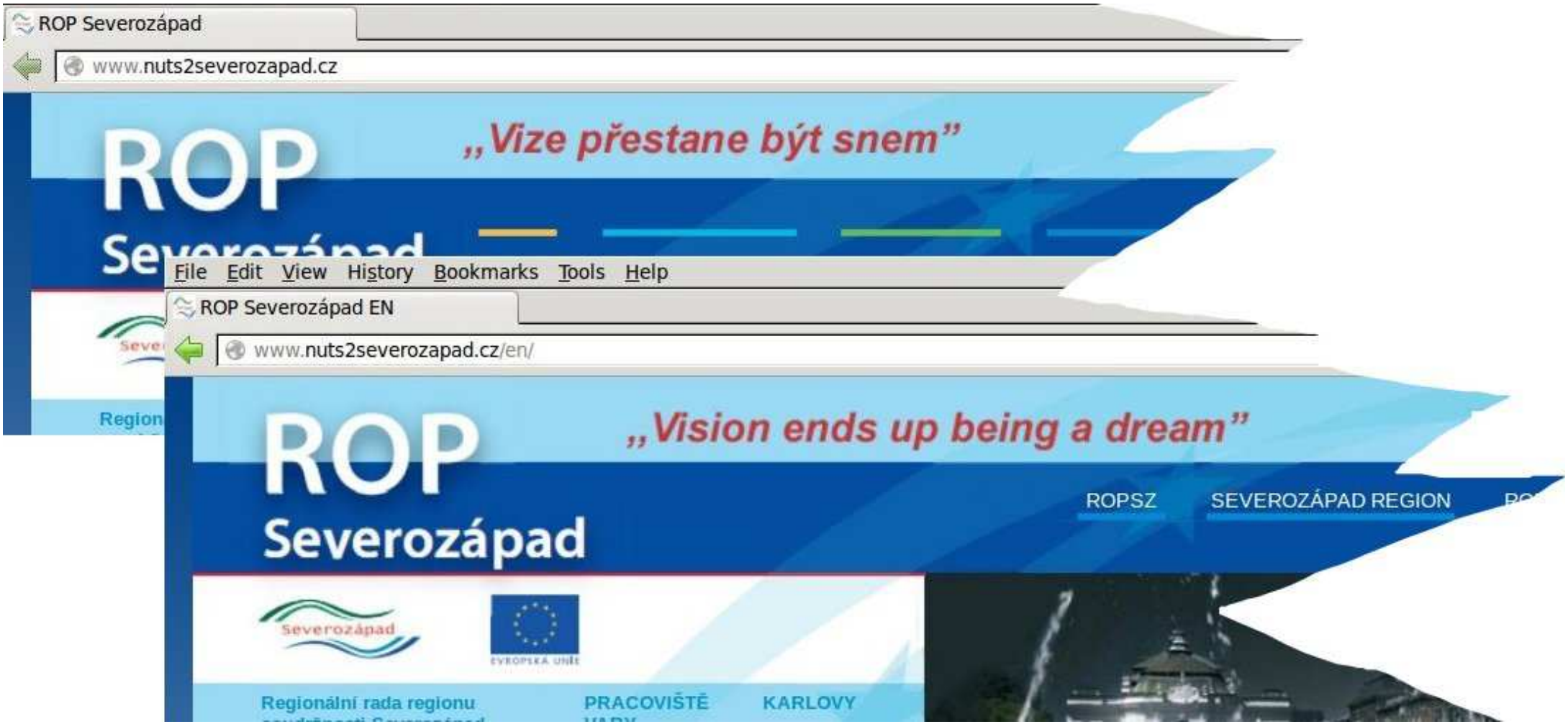
布拉格城市公共交通包括：城市火车、地铁、有轨电车、公共汽车。地铁一共有A、B、C三条线，纵横交错贯穿整个布拉格，三条地铁线在市中心都可以互相交错转换。

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.

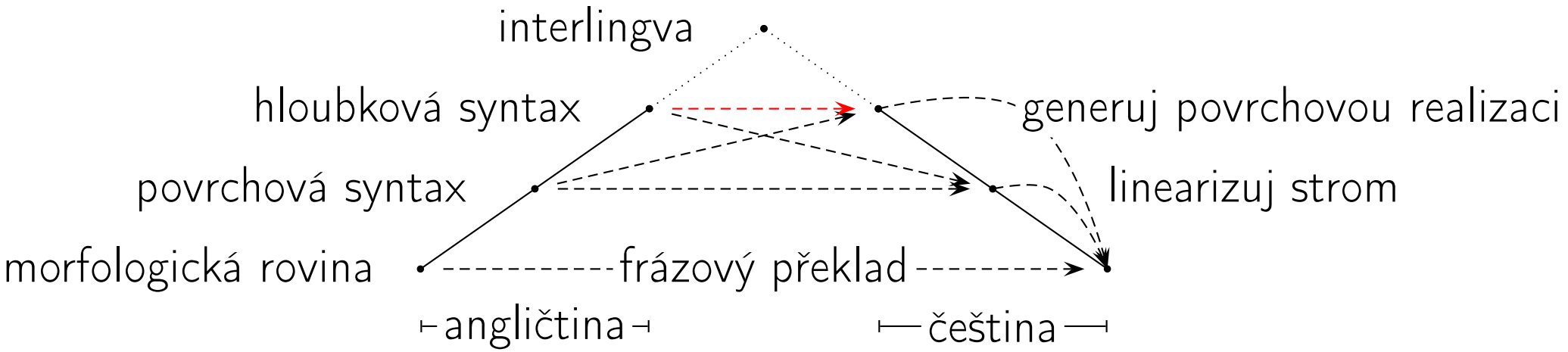
Vystřízlivění...

Ovšem chybují i lidé



Velka sbírka podobných: `http://www.engrish.com/`

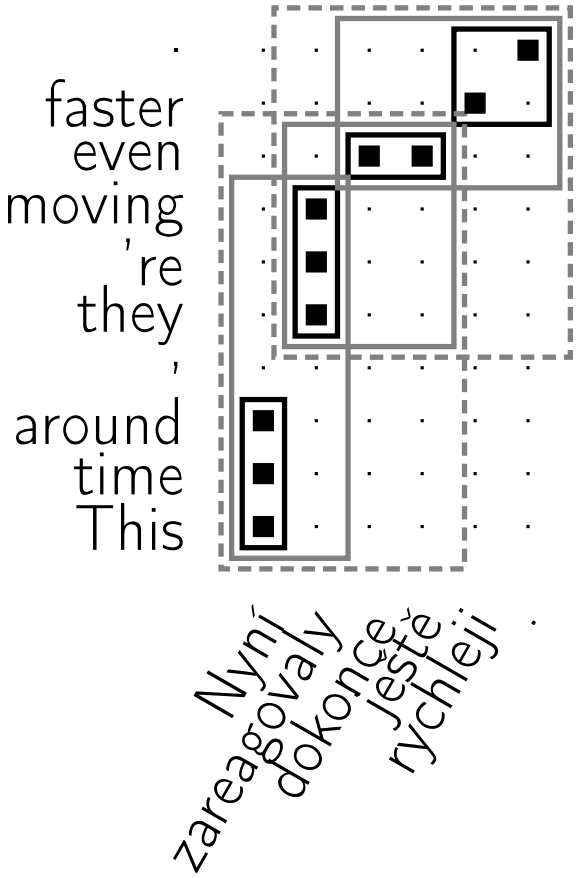
Přístupy ke strojovému překladu



- 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.

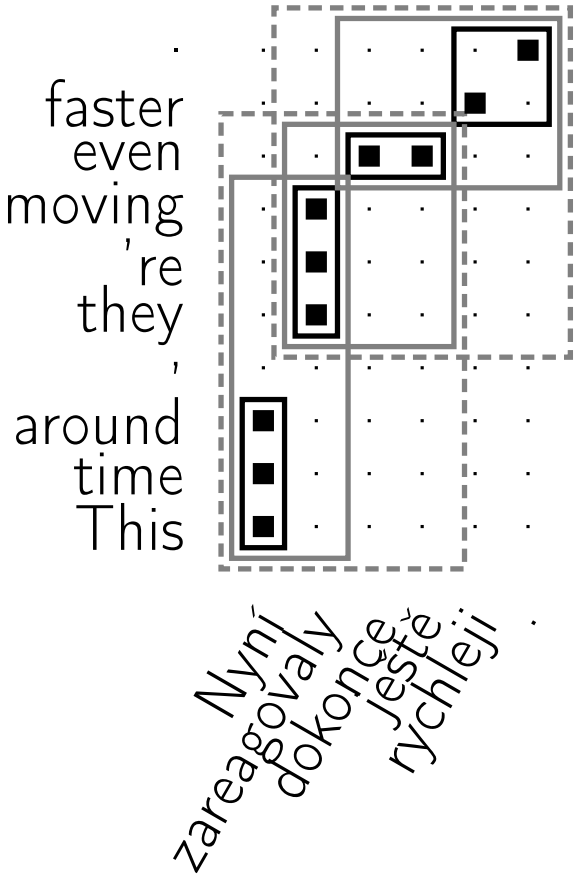
Frázový překlad



Trénovací data:

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

Frázový překlad

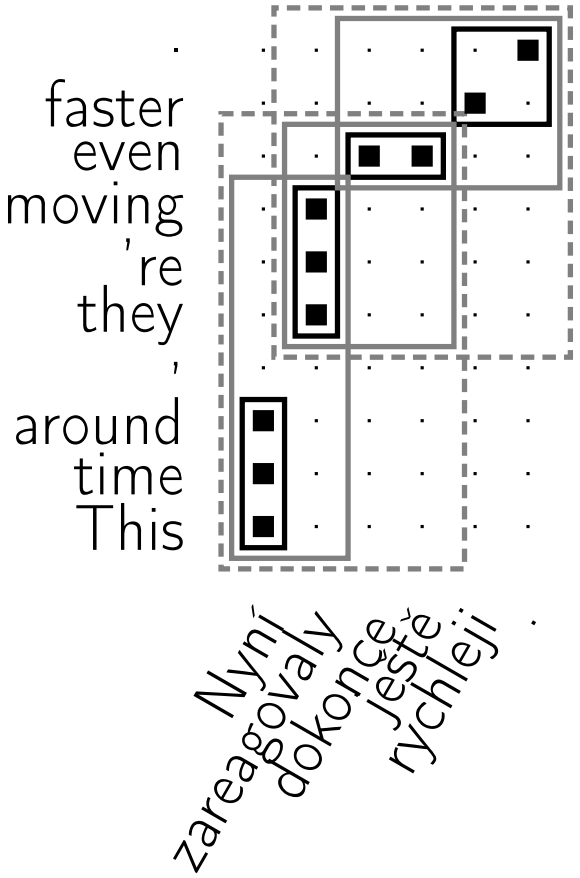


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Trénovací data:

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

Frázový překlad



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:

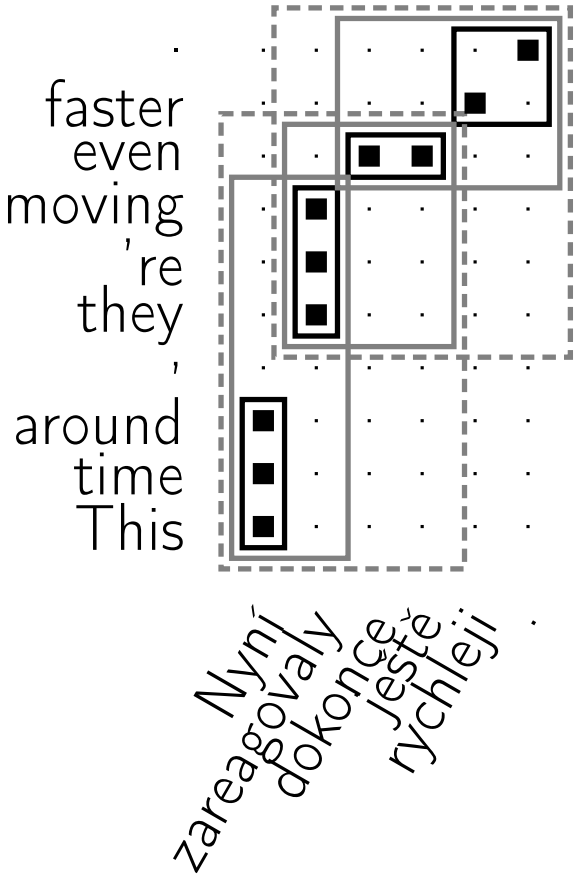
- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo)

Při samotném překladu hledáme:

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

aby byl výstup co nejpravděpodobnější.

Frázový překlad



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: ... 15 mil. párů vět

- paralelní korpus (česká věta = anglická věta)
- automatické zarovnání slov (české slovo ∼ anglické slovo) ... >200 mil. slov v každé řeči

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.	✗

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.	✗

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.	×

...zkusme tedy překlad dělat pořádně.

Formální popis češtiny

zákony

udělejte

pro

lidi

Formální popis češtiny

Morfologická rovina:

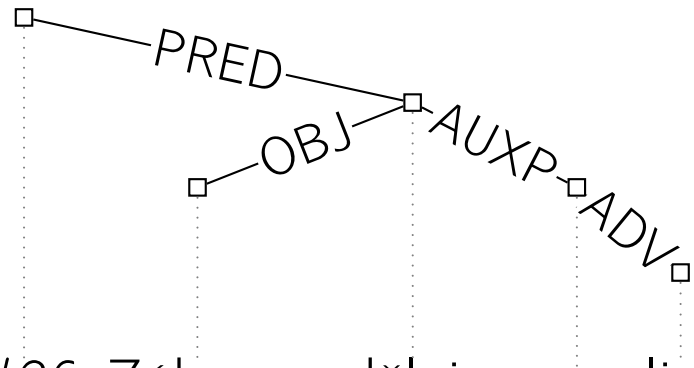
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Formální popis češtiny

Morfologická rovina:

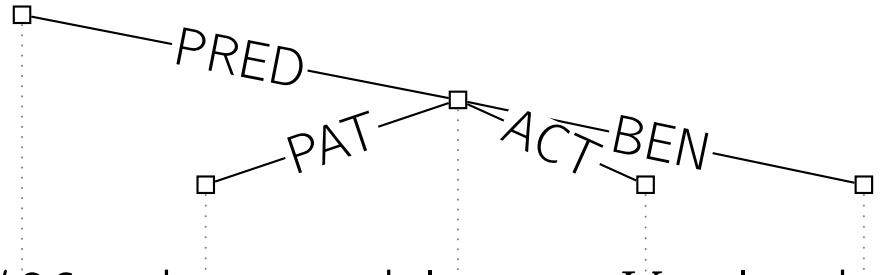
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

Analytická rovina (povrchová syntax):



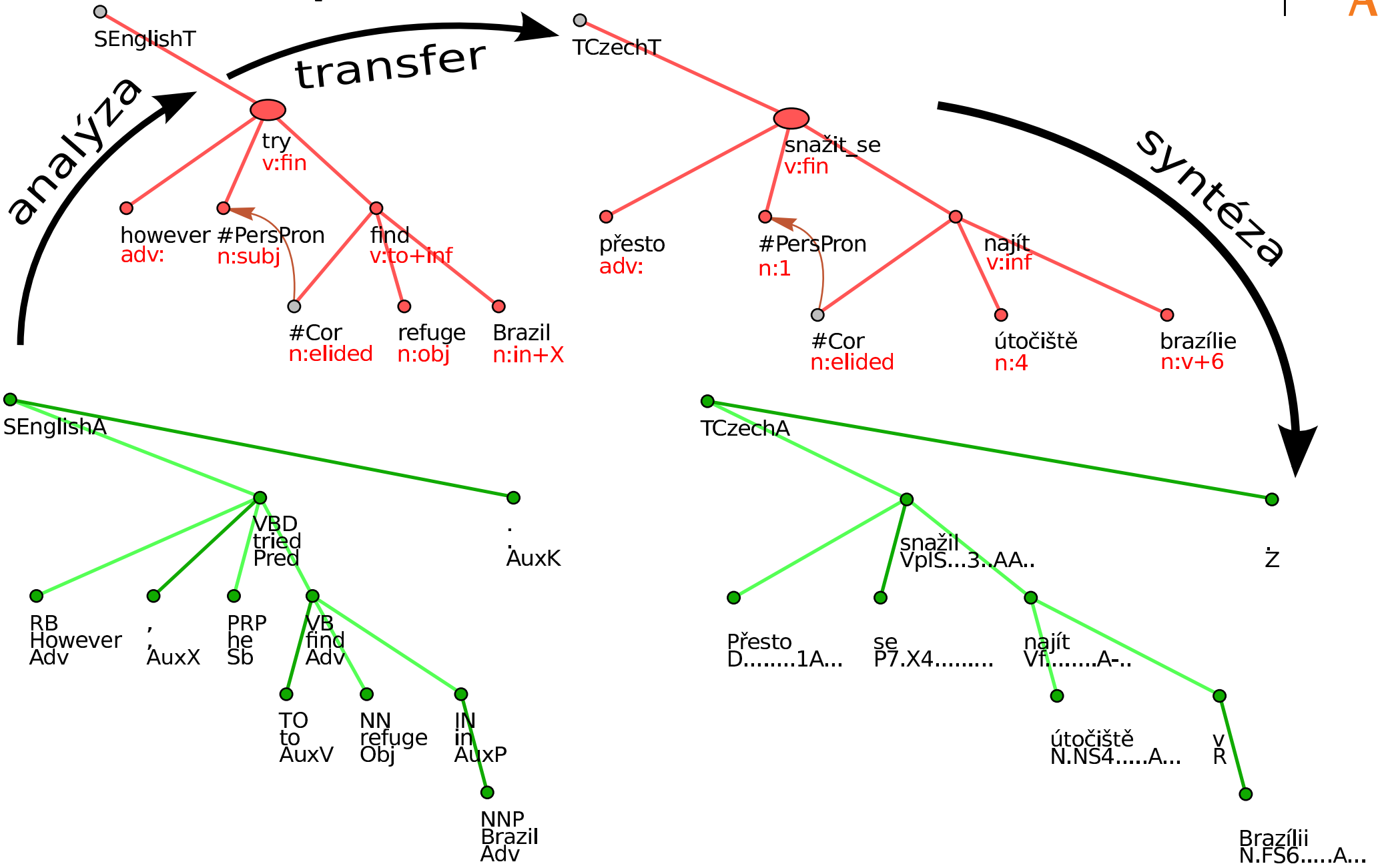
#36 Zákony udělejte pro lidi

Tektogramatická rovina (hloubková syntax):



#36 zákon$_{Pl}$ udělat$_{imp}$ Vy člověk$_{Pl,pro}$

Překlad přes hloubkovou rovinu



Proč je překlad těžký?

- Víceznačnost a význam slov.
- Cílový slovní tvar.
- Negace.
- Zájmena.
- (Koordinace.)

A též již 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

Reálné příklady: ...ze schůze sněmovny vypadl horní zákon. (Týden 40/2009)

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

Reálné příklady: ...ze schůze sněmovny vypadl horní zákon. (Týden 40/2009)

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.

„Úvaha“ frázového překladu

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			...	...		

„Úvaha“ zahrnuje členění vstupu

- Frázový překlad současně volí:
  - segmentaci věty na jednotlivé fráze,
  - překlady jednotlivých frází.
- Zda užívat fráze delší či kratší závisí zhruba na tom, jak dobrý se očekávaná překryv mezi trénovacími a testovacími daty.
- U jednotlivých vět segmentace ale úzce souvisí s významem:

Překlad \ Vstup	But	that	was	not	until	Sunday	.
Delší fráze	To ale nebylo				až do neděle.		
Spíše doslovný	Ale	to	bylo	až		v neděli	.

Negace

- Francouzská negace je okolo slovesa:

  Je ne parle pas français.

- Česká negace bývá zdvojená:

  Nemám žádné námitky.

Zdvojená negace při trénování vede ke ztrátě negace při překladu:

Trénovací	Nemám | žádného | psa.	=	I have | no | dog.	
data	Vidím | kočku.	=	I see | a cat.	
Vstup	Nemám | kočku.	=	I have | a cat.	×

- V severní a jižní Itálii se prý jízdenka v MHD procvaknutím:

  zneplatňuje nebo učiní platnou (in/validare).

Zájmena

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

Četl knihu. = He read a book.

Spal jsem. = I slept.

- Rod českého zájmene musí odpovídat odkazovanému slovu:

He saw a book. It was red.

Viděl knihu. Byla červená.

He saw a pen. It was red.

Viděl pero. Bylo červené.

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. ×

Shrnutí problémů překladu

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).
  $\Rightarrow$ Nutno studovat jen ty nadějné.
- Je těžké poznat lepší možnosti.
  (I lidé se neshodnou v tom, jak něco přeložit.)

Všechny dobré překlady

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

Všechny dobré překlady

Příklady ze 71 tisíc správných překladů anglické věty:

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdobná.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podobná.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

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. ...nebo to, co v trénovacích stromech nebylo.
- Má potenciál řešit těžší problémy, např. koreference.

Který přístup vítězí? Nevíme.

Angličtina→čeština	ÚFAL		Komerční	
	Frázový	Hloubkový	Google	PC Trans.
Oficiální WMT10: Seřaďte hypotézy od nejlepší po nejhorší. Shody povoleny.				
> ostatní	**45.0**	44.1	49.1	49.4
>= ostatní	**65.6**	60.1	70.4	62.1
Neoficiální WMT10: Člověk zkusil výstup MT opravit bez znalosti originálu.				
Je to dobrý překlad? (%)	**40**	34	55	43
Neoficiální: MT přeložil krátký text. Dokážete správně zodpovědět kontrolní otázky?				
% správných odpovědí	73.6	80.6	78.7	80.2

- Pravidelné soutěže (`http://www.statmt.org/wmt12/`).

Shrnutí

- Dva přístupy ke strojovému překladu.
  - Frázový a syntaktický.
- Obtížnost překladu jako taková.
  - Problematické jazykové jevy obecně.
  - Vstupy, které rozloží frázový i syntaktický překlad.
- Obtížnost rozhodnout, který překlad je lepší.

Odkazy

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

→ Research → Prague Czech-English Dependency Treebank 2.0

→ Data: Ukázkové české a anglické povrchové i hloubkové rozbory

→ Video Recordings

Ukázky frázového překladu:

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

http://demo.statmt.org/

http://tool.statmt.org/

http://studuj-matfyz.cz/

Proč studovat na MFF a ÚFALu

Můžete se naučit mj.:

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

  Text na českém webu ∼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 >200 CPU, počítače s 32 GB RAM a jeden s 0.5 TB RAM.
- Soutěžit na mezinárodní úrovni v překládání, analýzách, generování, ...

Lingvistické nástroje

- Identifikace kódování dokumentu a jazyka.
- 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.

  Rice University ≠ univerzita rýže
- Koreference (mj. identifikace, co zastupují zájmena).

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

⇒ Lze využít v programech (pravidlových i statistických).

Lingvistické aplikace

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