Udělá za vás strojový překlad domácí úkol?

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

Obsah prezentace

- Úvod do strojového překladu:
  - Motivace k překladu.
  - Obtížnost překladu.
- Podrobněji: Dva přístupy k překladu.
  - Frázový překlad a jeho problémy.
  - Hloubkový překlad a jeho problémy.
- Ještě podrobněji: (slidy anglicky)
  - Co dělá překlad *statistickým*.
  - Formální definice, Bayesův zákon.
  - Stavový prostor částečných hypotéz.
- Proč studovat na MFF (a ÚFALu).

Strojový překlad je lákavý

Strojový překlad (machine translation, MT) zajímavý akademicky, komerčně i pro uživatele:

- Hřiště pro testování užitečnosti mnoha dílčích nástrojů zpracování jazyka.
- EU utrácí ročně 1 000 000 000 eur za překlady.
- USA investuje do překladu pro účely rozvědky.
- Automatický překlad umožňuje využít texty z webu bez ohledu na zdrojový jazyk.

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

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.

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.

Proč je překlad těžký?

- Víceznačnost a význam slov.
- Cílový slovní tvar.
- Pořádek slov (tj. i vzdálenost mezi slovy).
- Negace.
- Zájmena.
- 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.

Pořádek slov

- Anglicky: subject-verb-object (SVO)
- Japonsky: subject-object-verb (SOV)

IBM bought Lotus.
IBM Lotus bought.

Reporters said IBM bought Lotus.
Reporters IBM Lotus bought said.

- Německy: Satztklammer (SV_1OV_2, OV_1SV_2)

  Die Satzklammer oder Klammerform **stellt** den typischen Satzbau der deutschen Sprache **dar**.

- Kombinatorická exploze možností, nestihneme probrat všechny.

Negace

- 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).

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. Byl<u>a</u> černá<u>á</u>.

He saw a pen. *It was red.*

Viděl pero. Byl<u>o</u> černé<u>é</u>.

Idiomatická spojení

Kromě známého:

kick the bucket = natáhnout bačkory

a bone of contention = jablko sváru

jde i „obyčejná“ frázová slovesa:

run into = potkat

show up = přijít, ukázat se, stavit se

make up = vymyslet si

talk sb. into sth. = přemluvit někoho, aby ...

I lidé se překlad kazí...

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

I lidé se překlad kazí...

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

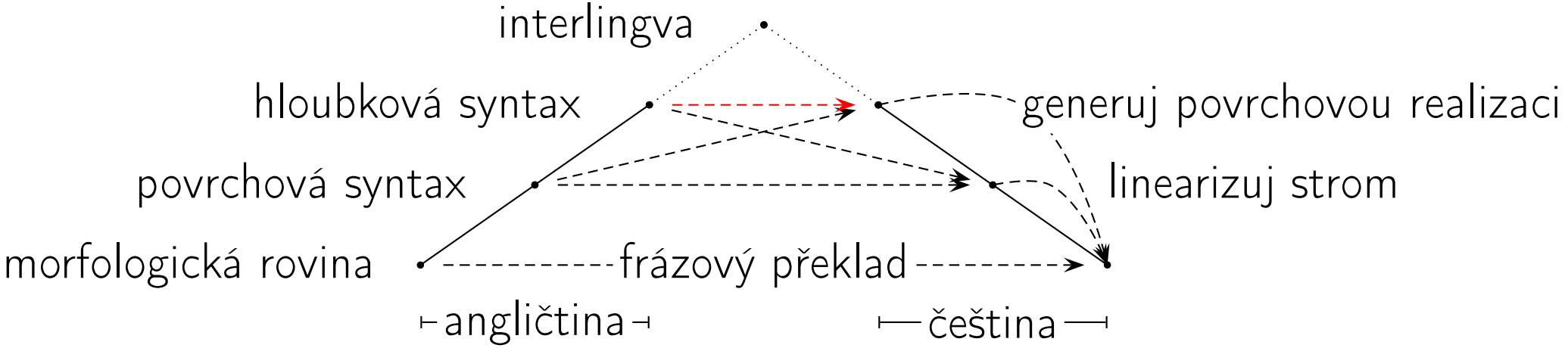
The Fentimans Logo is a based on Fearless, Thomas Fentiman's prize Alsatian, double winner of the Crufts Obedience Test.

I lidé se překlad kazí...

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.

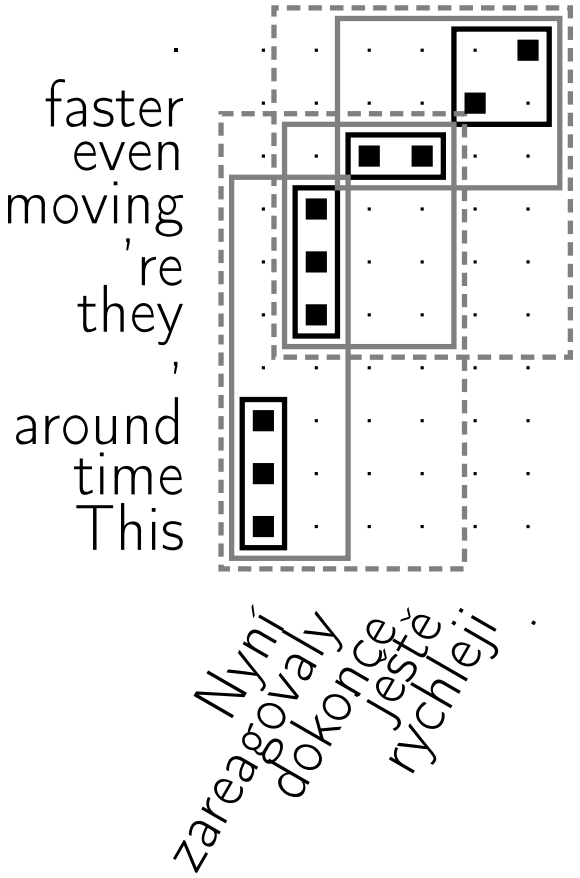
Přístupy ke strojovému překladu



- 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.

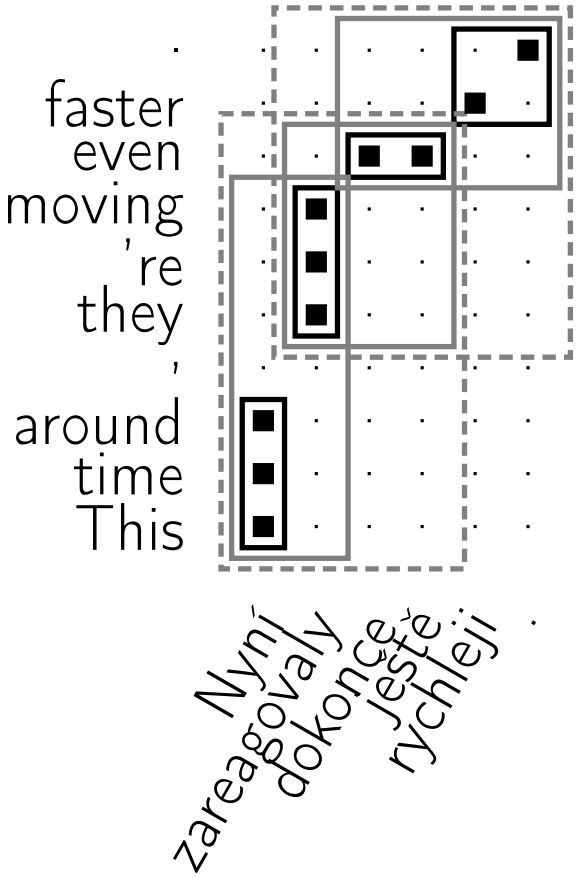
Frázový překlad



Trénovací data:

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

Frázový překlad

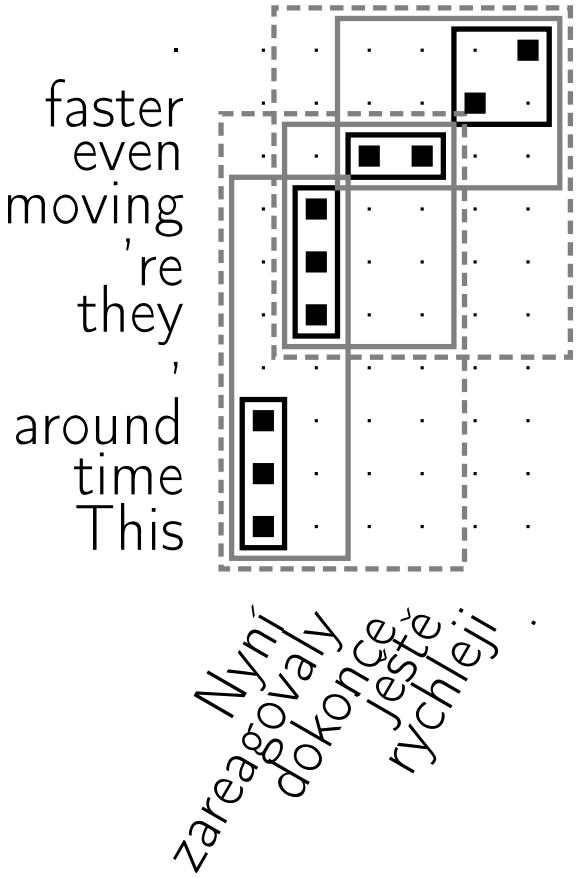


This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

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 v praxi

Můj aktuální model na letošní soutěž:

- vychází z paralelního korpusu CzEng 1.0:
  - 15 milionů paralelních vět,
  - 200/230 milionů českých/anglických slov,
  - cca 3 měsíce čištění.
- slovní zarovnání běželo 52 hodin (2 vlákna) a zabralo 24 GB RAM,
- extrakce frází trvala 15 hodin,
- překladový model:
  - 8 GB tabulka frází,
  - 3 GB tabulka slovosledných změn,
  - 3 GB jazykové modely,
  - ladění vah trvalo 3 hodiny (15 vláken).

(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.	✗

Fráze jsou pevné délky...

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

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

Fráze jsou pevné délky...

...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.	✓

Fráze jsou pevné délky...

...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.	~

Fráze jsou pevné délky...

...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.	×

Fráze jsou pevné délky...

...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:

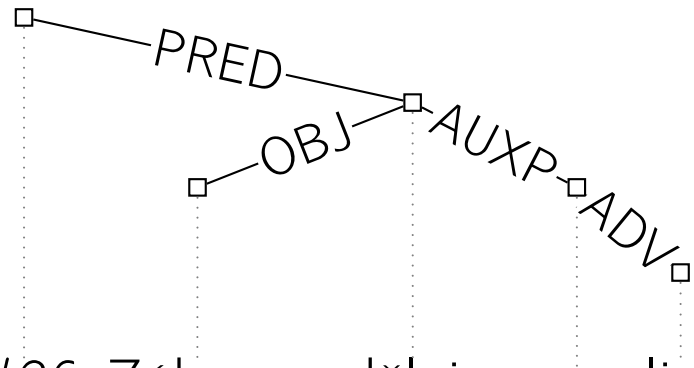
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk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

Formální popis češtiny

Morfologická rovina:

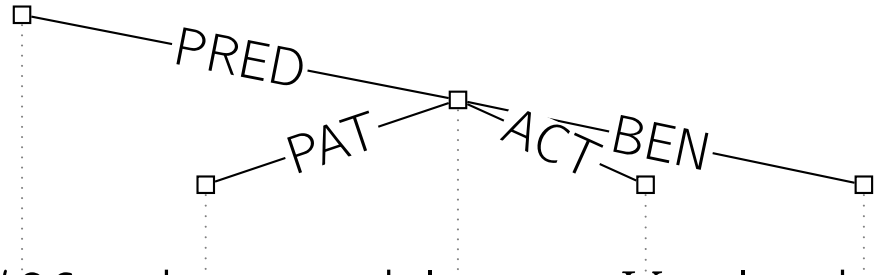
Slovo	Lema	Morfologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk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

Analytická rovina (povrchová syntax):



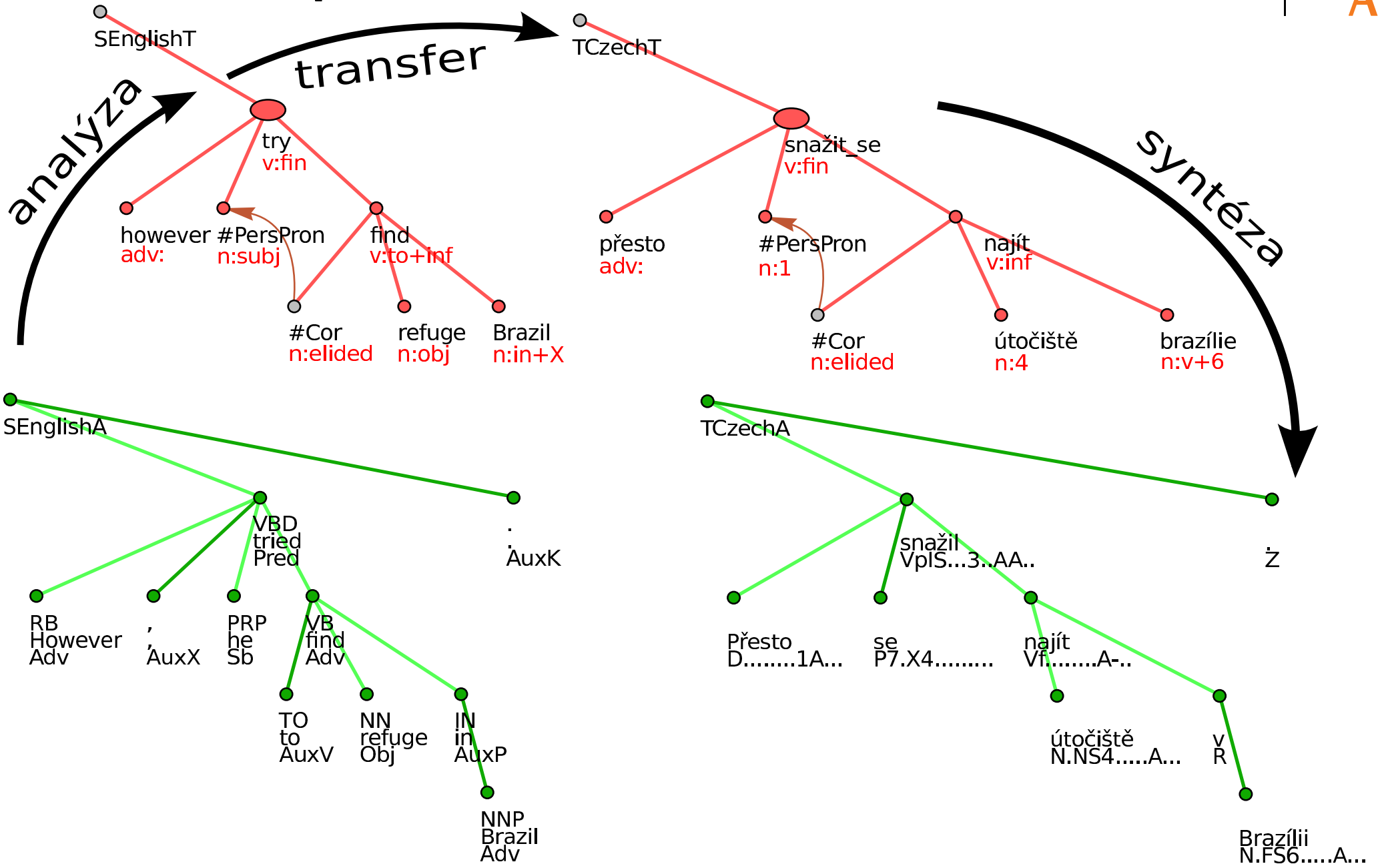
#36 Zákony udělejte pro lidi

Tektogramatická rovina (hloubková syntax):



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

Překlad přes hloubkovou rovinu



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.)

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 (cokoli, co v trénovacích stromech nebylo).

$\Rightarrow$ Zatím funguje lépe frázový překlad.

$\Rightarrow$ Syntaktický překlad má ale potenciál řešit těžší problémy.

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/`).

Quotes on Statistical MT

Warren Weaver (1949):

> I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that is has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.

Noam Chomsky (1969):

> ...the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Frederick Jelinek (80's; IBM; later JHU and sometimes ÚFAL)

> Every time I fire a linguist, the accuracy goes up.

Hermann Ney (RWTH Aachen University):

> MT = Linguistic **M**odelling + Statistical Decision **T**heory

The Statistical Approach

(Statistical = Information-theoretic.)

- Specify a probabilistic model.
  = How is the probability mass distributed among possible outputs given observed inputs.
- Specify the training criterion and procedure.
  = How to learn free parameters from training data.

Notice:

- Linguistics helpful when designing the models:
  – How to divide input into smaller units.
  – Which bits of observations are more informative.

Statistical MT

Given a source (foreign) language sentence $f_1^J = f_1 ... f_j ... f_J$,
Produce a target language (English) sentence $e_1^I = e_1 ... e_j ... e_I$.

Among all possible target language sentences, choose the sentence with the highest probability:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \tag{1}$$

We stick to the e_1^I, f_1^J notation despite translating from English to Czech.

Brute-Force MT

Translate only sentences listed in a "translation memory" (TM):

Good morning.	=	Dobré ráno.
How are you?	=	Jak se máš?
How are you?	=	Jak se máte?

$$p(e_1^I|f_1^J) = \begin{cases} 1 & \text{if } e_1^I = f_1^J \text{ seen in the TM} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- Not a probability. There may be f_1^J, s.t. $\sum_{e_1^I} p(e_1^I|f_1^J) > 1$.

  $\Rightarrow$ Have to normalize, use $\frac{\text{count}(e_1^I, f_1^J)}{\text{count}(f_1^J)}$ instead of 1.

- Not "smooth", no generalization:

Good morning.	$\Rightarrow$	Dobré ráno.
Good evening.	$\Rightarrow$	$\emptyset$

Bayes' Law

Bayes' law for conditional probabilities: $p(a|b) = \frac{p(b|a)p(a)}{p(b)}$

So in our case:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \qquad \text{Apply Bayes' law}$$

$$= \underset{I, e_1^I}{\operatorname{argmax}}\, \frac{p(f_1^J | e_1^I) p(e_1^I)}{p(f_1^J)} \qquad p(f_1^J) \text{ constant}$$

$$\Rightarrow \text{irrelevant in maximization}$$

$$= \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I)$$

Also called "Noisy Channel" model.

Motivation for Noisy Channel

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I) \tag{3}$$

Bayes' law divided the model into two components:

$p(f_1^J | e_1^I)$ Translation model ("reversed", $e_1^I \rightarrow f_1^J$)

...is it a likely translation?

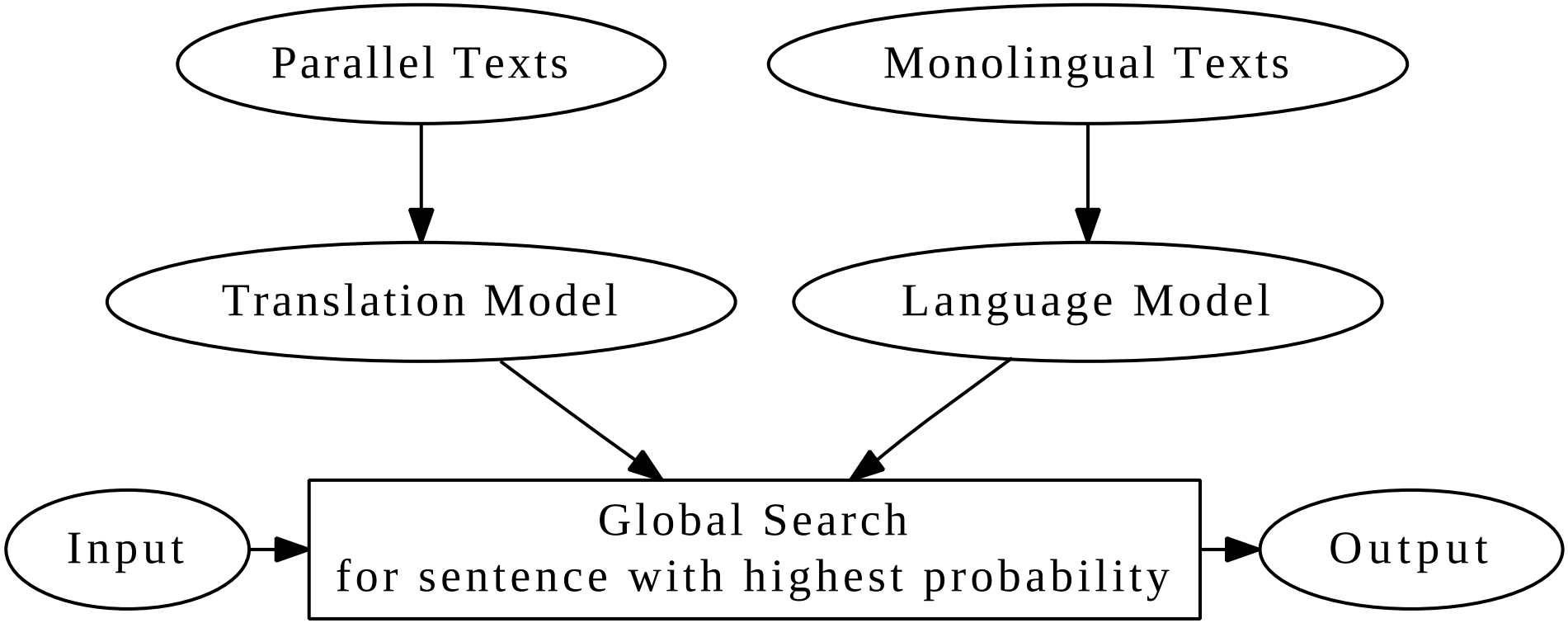
$p(e_1^I)$ Language model (LM)

...is the output a likely sentence of the target language?

- The components can be trained on different sources.

  There are far more monolingual data ⇒ language model more reliable.

Without Equations



Search Space of PBMT

... see the slides by Philipp Koehn and Barry Haddow.

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 nehledali jehly 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, ...

Shrnutí

- Dva přístupy ke strojovému překladu.
- Obtížnost překladu jako taková.
  - Problematické vstupy pro frázový i syntaktický překlad.
- Frázový překlad podrobně:
  - Věty a jejich pravděpodobnost.
  - Bayesův vzorec.
  - Prohledávání prostoru hypotéz.

... na Matfyzu si sáhnete na nejžhavější novinky hardwarové i softwarové.

Odkazy

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

→ Video Recordings

→ Tools (→ překladový systém Moses)

Další ukázky frázového překladu:

http://demo.statmt.org/

http://tool.statmt.org/

Summary of Language Models

- $p(e_1^I)$ should report how "good" sentence e_1^I is.
- We surely want $p(\text{The the the.}) < p(\text{Hello.})$
- How about $p(\text{The cat was black.}) < p(\text{Hello.})$?

  ...We don't really care in MT. We hope to compare synonymic sentences.

LM is usually a 3-gram language model:

$$p(\text{↱ ↱ The cat was black . ↰ ↰}) = \begin{array}{lll} p(\text{The}|\text{↱ ↱}) & p(\text{cat}|\text{↱ The}) & p(\text{was}|\text{The cat}) \\ p(\text{black}|\text{cat was}) & p(\text{.}|\text{was black}) & p(\text{↰}|\text{black .}) \\ p(\text{↰}|\text{. ↰}) & & \end{array}$$

Formally, with $n = 3$:

$$p_{\text{LM}}(e_1^I) = \prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}) \tag{4}$$