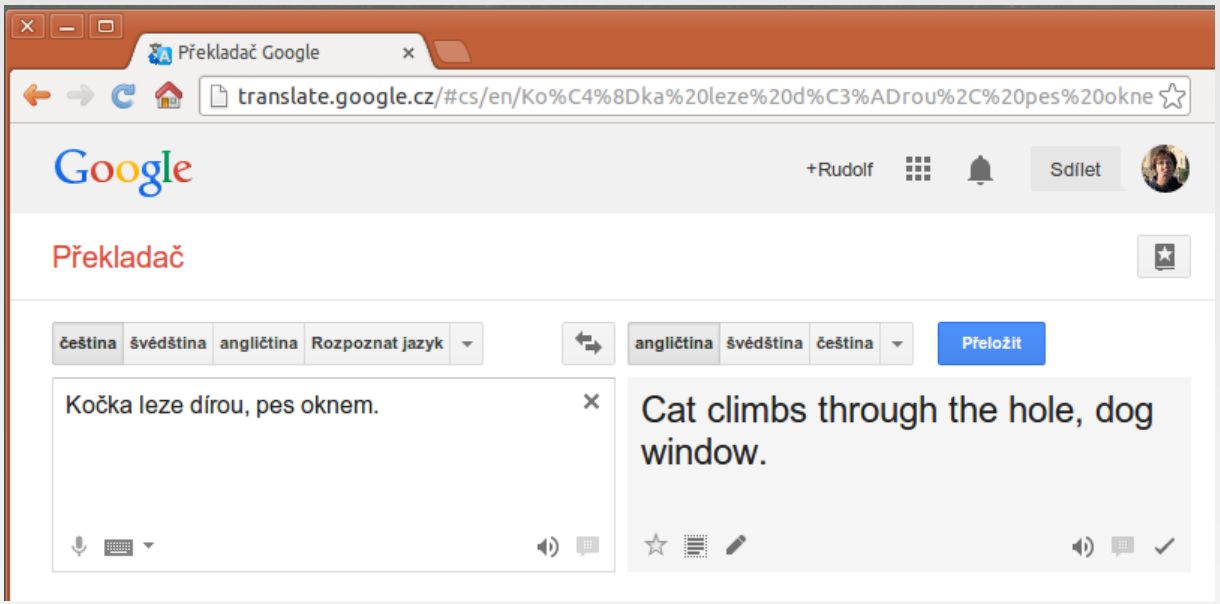
Mgr. Rudolf Rosa

Jak dělat strojový překlad lépe než Google Translate

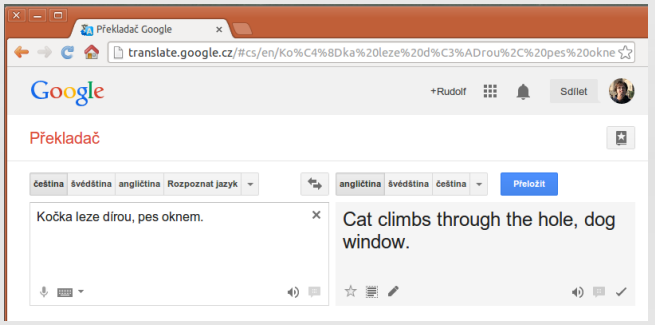


Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

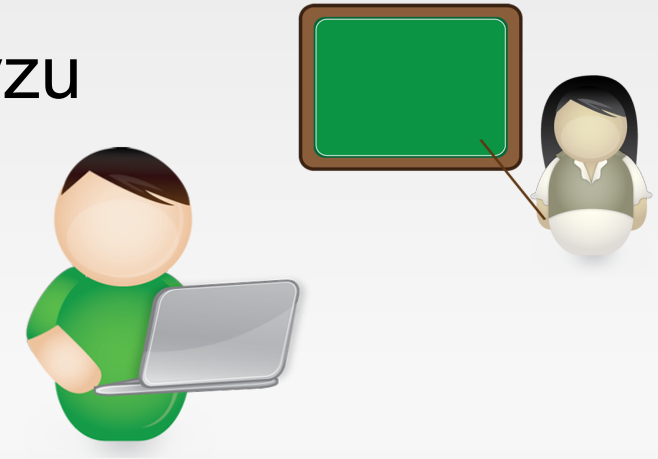
ProSŠ, Gymnázium Kladno, 23. října 2014

Co se dozvíte

- Jak funguje Google Translate?
  - slovník?
  - fráze z internetu?
  - problémy?



- Jde to i lépe?
  - moje diplomová práce na Matfyzu
  - proč se programátorovi hodí znát českou gramatiku
  - <u>všechny</u> problémy vyřešeny?



Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague

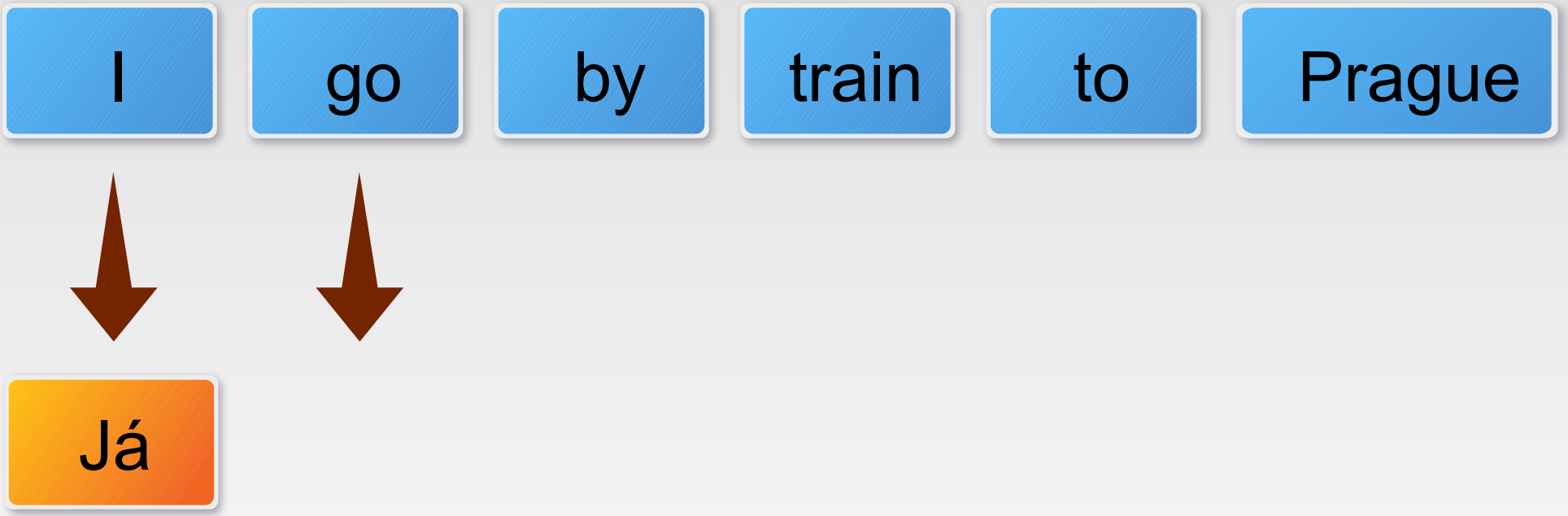
Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague

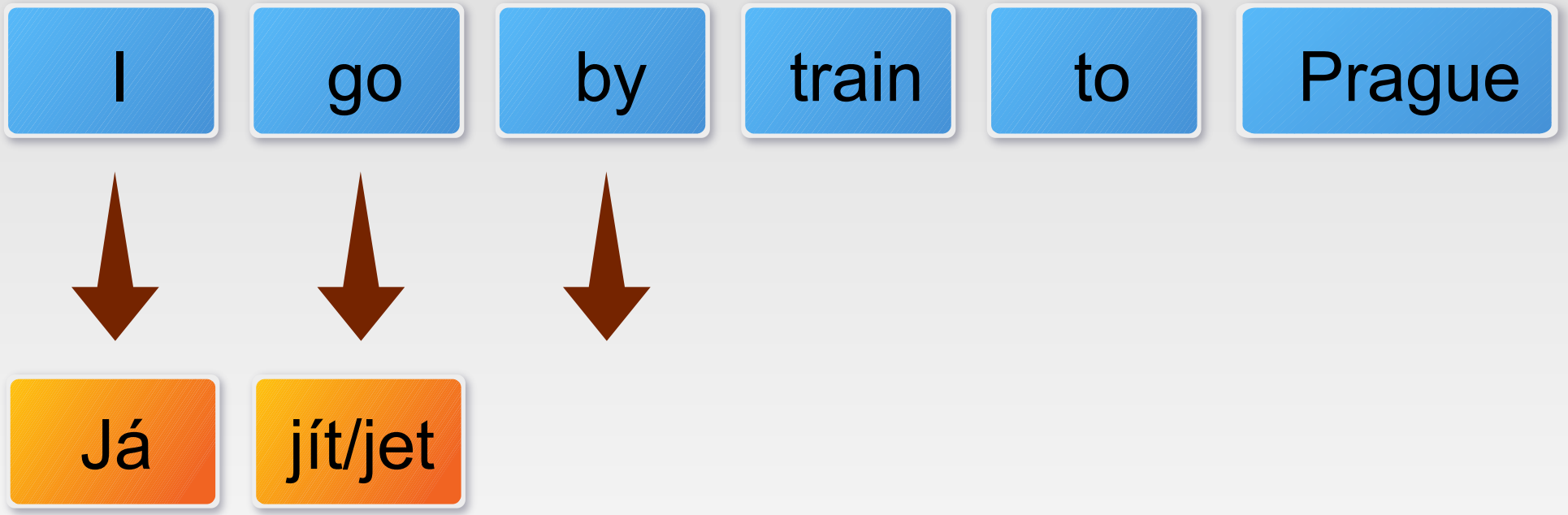
Jak překládá počítač? 1940 – 1990

- slovník 



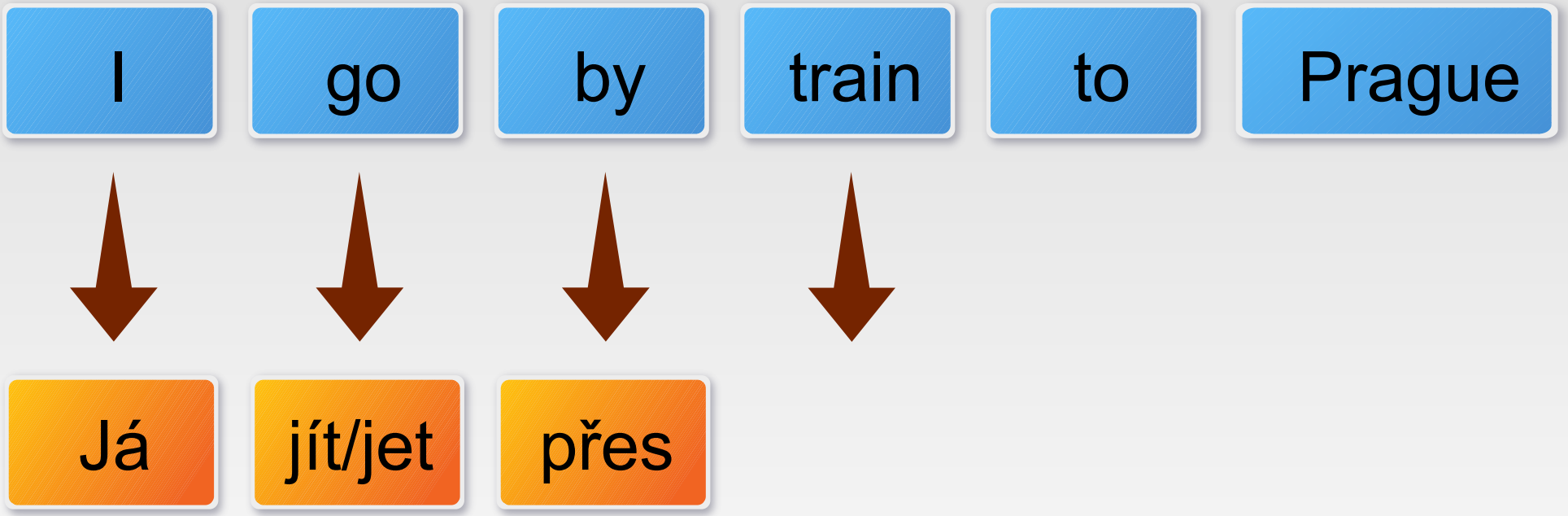
Jak překládá počítač? 1940 – 1990

- slovník 



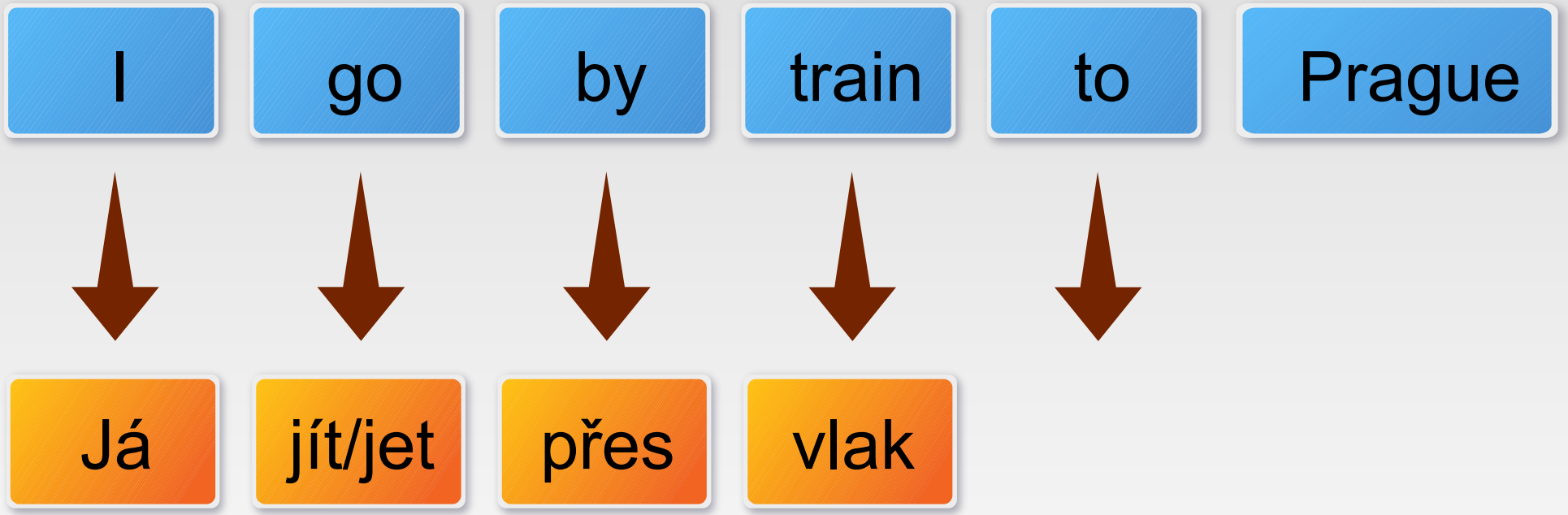
Jak překládá počítač? 1940 – 1990

- slovník 



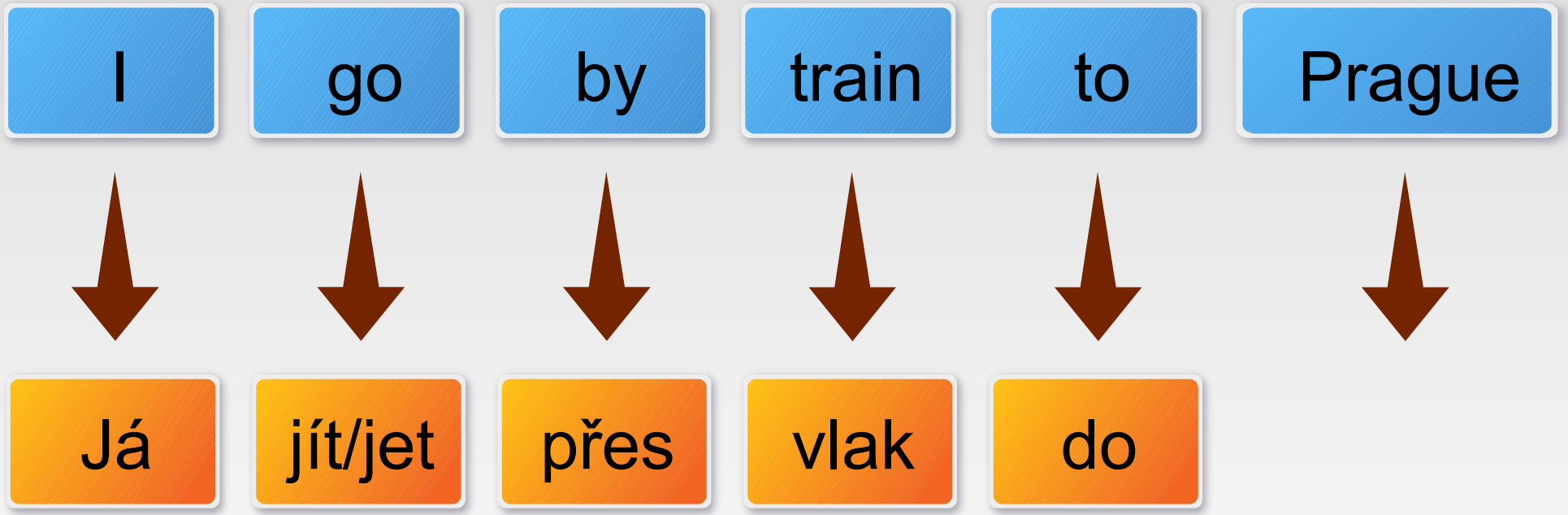
Jak překládá počítač? 1940 – 1990

- slovník 



Jak překládá počítač? 1940 – 1990

- slovník 



Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague
↓	↓	↓	↓	↓	↓
Já	jít/jet	přes	vlak	do	Praha

Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague
↓	↓	↓	↓	↓	↓
Já	jít/jet	přes	vlak	do	Praha

- slovník 

Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague
↓	↓	↓	↓	↓	↓
Já	jít/jet	přes	vlak	do	Praha

- slovník + gramatika 

Jak překládá počítač? 1940 – 1990

- slovník 

I	go	by	train	to	Prague
↓	↓	↓	↓	↓	↓
Já	jít/jet	přes	vlak	do	Praha

- slovník + gramatika + … 

Jak překládá počítač? 1940 – 1990

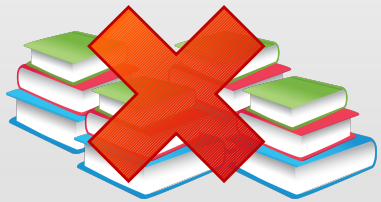
- slovník 

I	go	by	train	to	Prague
↓	↓	↓	↓	↓	↓
Já	jít/jet	přes	vlak	do	Praha

- slovník + gramatika + … + …? 

Jak překládá počítač? 1990 –

- zahodit slovník i gramatiku

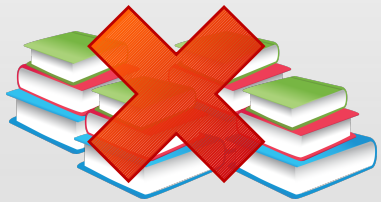


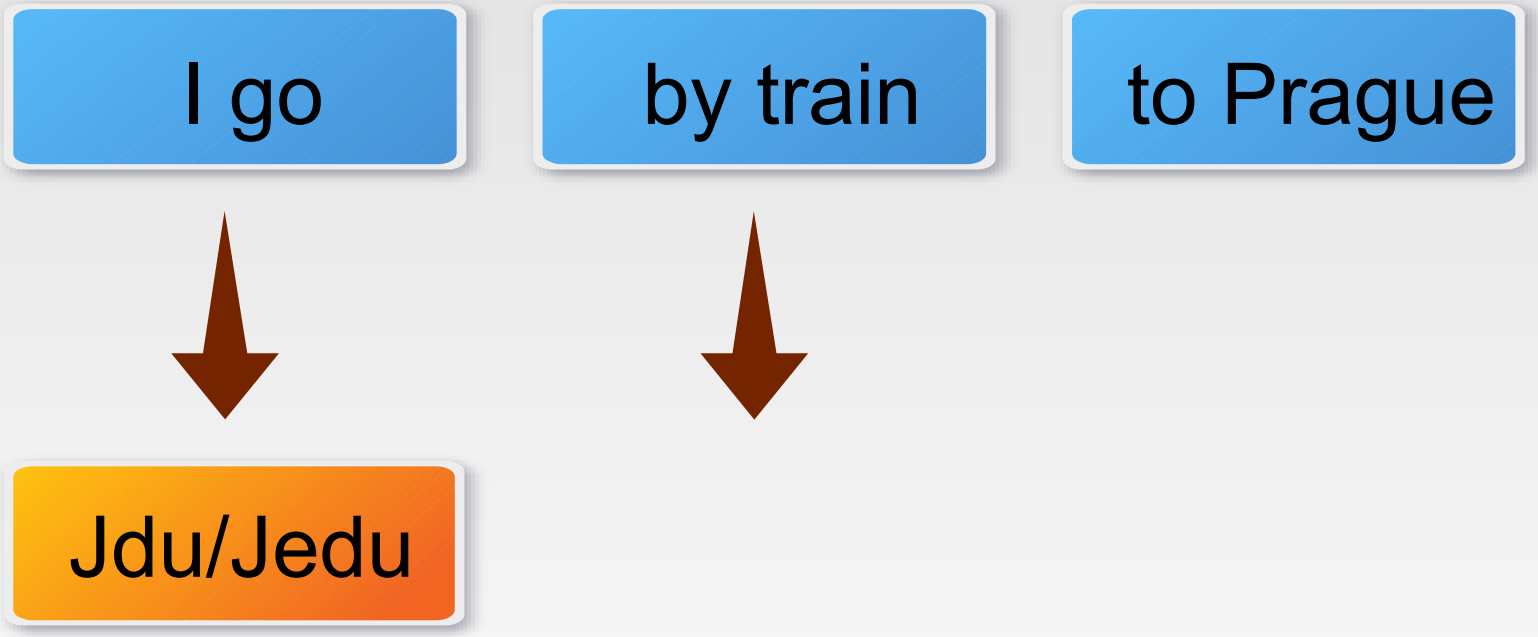
- překládat po frázích

I go | by train | to Prague

Jak překládá počítač? 1990 –

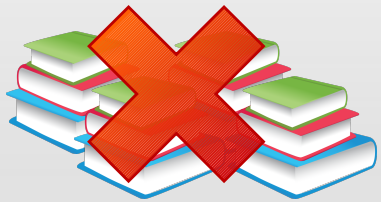
- zahodit slovník i gramatiku
- překládat po frázích

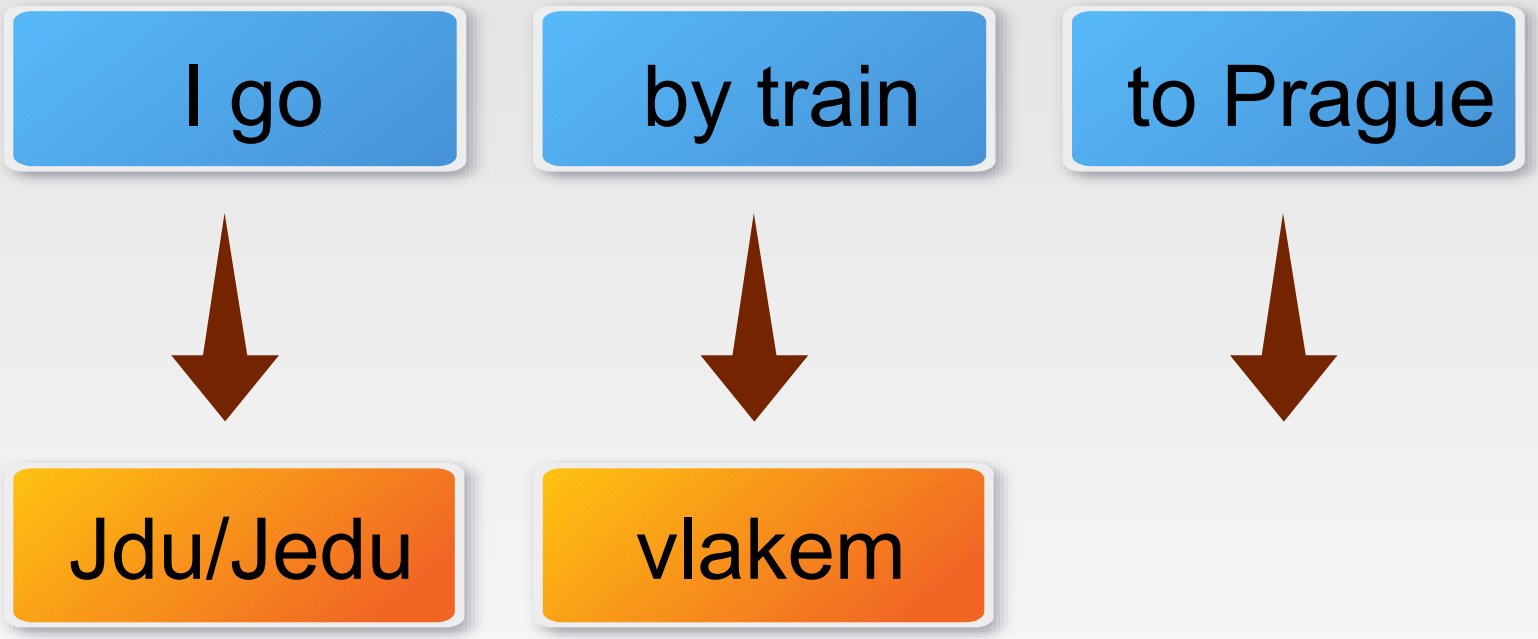




Jak překládá počítač? 1990 –

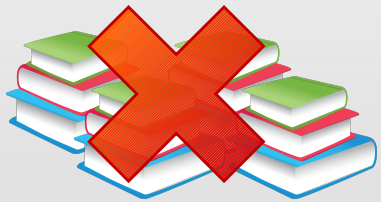
- zahodit slovník i gramatiku
- překládat po frázích

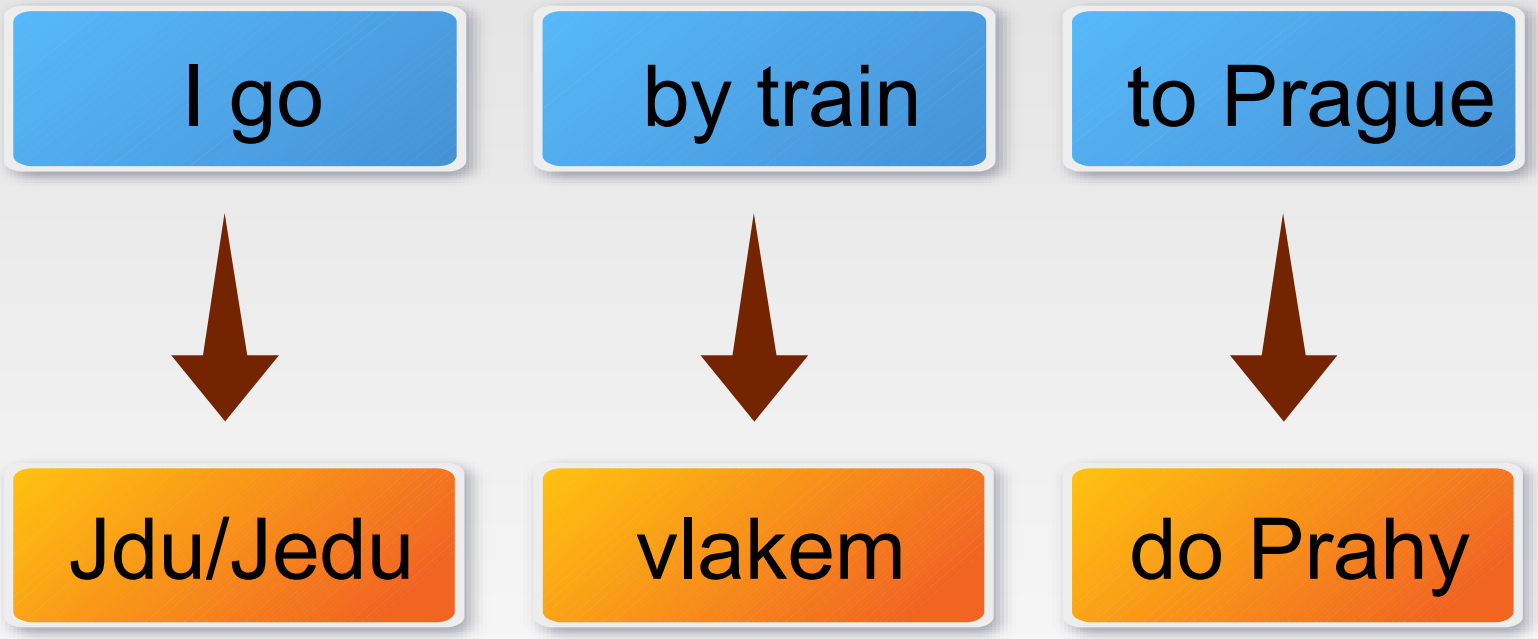




Jak překládá počítač? 1990 –

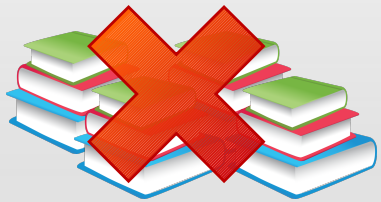
- zahodit slovník i gramatiku
- překládat po frázích





Jak překládá počítač? 1990 –

- zahodit slovník i gramatiku



- překládat po frázích

I go	by train	to Prague
↓	↓	↓
Jdu/Jedu	vlakem	do Prahy

- kde vzít ty fráze? slovník frází...?

Kde vzít frázovou tabulku

- filmové titulky, překlady knih, zákony EU...

Harry, you'll go to Hogwarts by train.	Harry, do Bradavic pojedeš vlakem.
But Anna did not intend to travel by train...	Ale Anna se nechystala vlakem cestovat...
Citizens are allowed to cross the border by train.	Občané smí překročit hranici vlakem.

Kde vzít frázovou tabulku

- filmové titulky, překlady knih, zákony EU...

Harry, you'll go to Hogwarts by train.	Harry, do Bradavic pojedeš vlakem.
But Anna did not intend to travel by train...	Ale Anna se nechystala vlakem cestovat...
Citizens are allowed to cross the border by train.	Občané smí překročit hranici vlakem.

Frázová tabulka nestačí

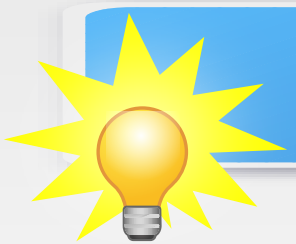
I go		by train	to Prague
Jdu	Jedu	vlakem	do Prahy

Frázová tabulka nestačí

I go	by train	to Prague
Jdu ? Jedu	vlakem	do Prahy

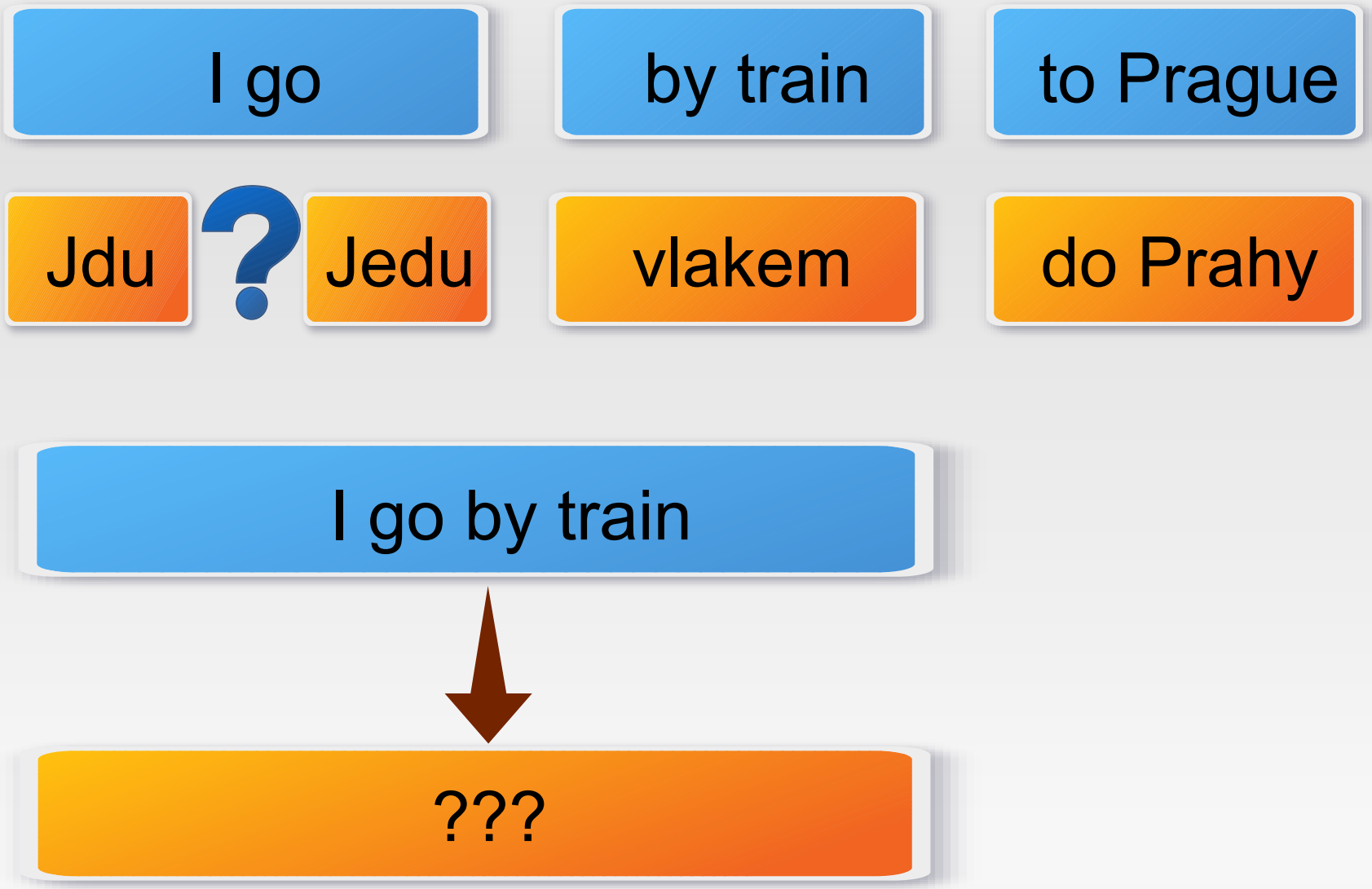
Frázová tabulka nestačí

I go	by train	to Prague
Jdu / Jedu	vlakem	do Prahy



I go by train

Frázová tabulka nestačí



Frázová tabulka nestačí

I go	by train	to Prague
Jdu ? Jedu	vlakem	do Prahy

I go by train

???

Frázová tabulka nestačí

I go	by train	to Prague
Jdu ? Jedu	vlakem	do Prahy

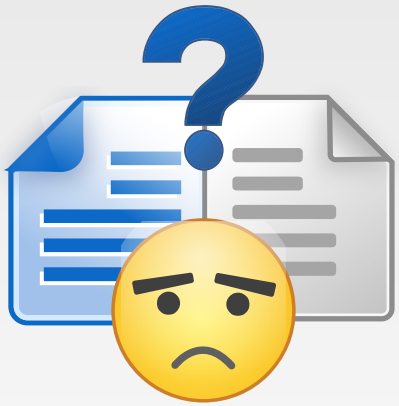
I go by train

???

Frázová tabulka nestačí

I go	by train	to Prague
Jdu ? Jedu	vlakem	do Prahy

I go by train



???

Frázová tabulka nestačí

I go	by train	to Prague
Jdu ? Jedu	vlakem	do Prahy

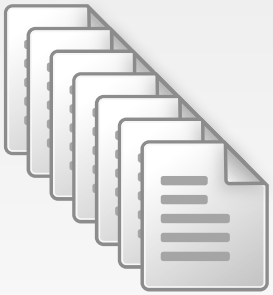
příliš dlouhá fráze!

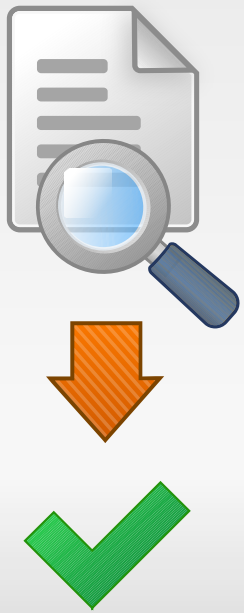
Frázová tabulka nestačí

- paralelních dat je málo

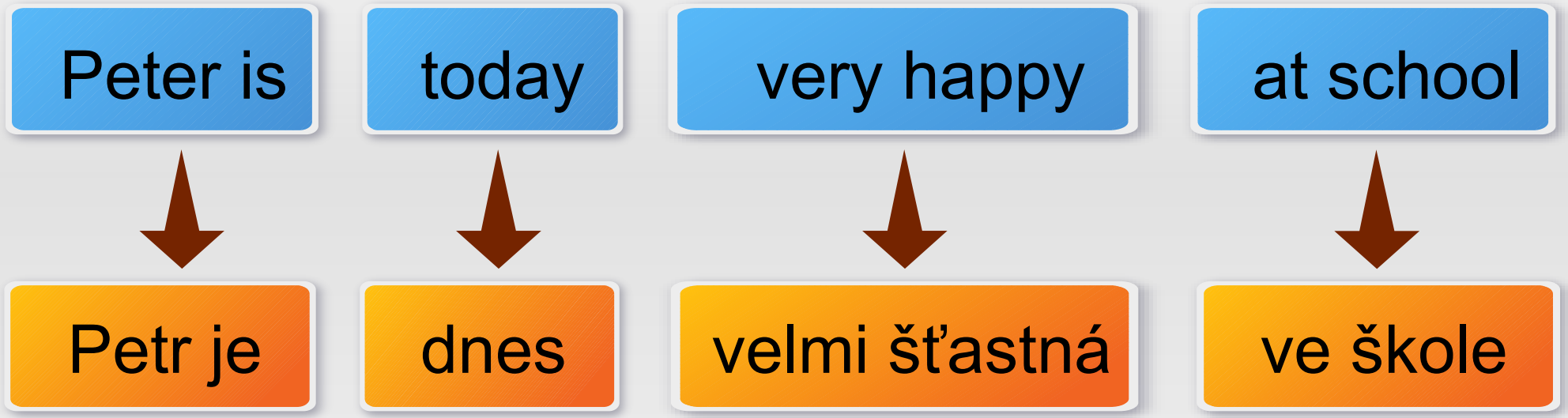
Jdu vlakem ? Jedu vlakem do Prahy

- …ale na tohle přece nepotřebuju paralelní data!
- monolingválních dat jsou spousty





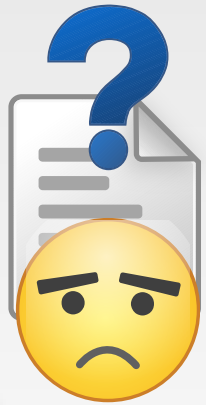
Už to takhle stačí?



Už to takhle stačí?

Peter is	today	very happy	at school
↓	↓	↓	↓
Petr je	dnes	velmi šťastná	ve škole

Petr je dnes velmi šťastná



příliš dlouhá fráze!

Proč překladači vadí skloňování

- překlad do angličtiny – OK
  - *Peter is today very **happy**,*
    *Mary is today very **happy**,*
    *puppy is today very **happy**...*

- překlad do češtiny – problém
  - *Petr je dnes velmi šťastn**ý**,*
    *Jana je dnes velmi šťastn**á**,*
    *dítě je dnes velmi šťastn**é**,*
    *vodník je dnes velmi šťastn**ý**,*
    *štěně je dnes velmi šťastn**é**...*

Proč překladači vadí skloňování

- *angličtina*
  - *happy*

- *čeština*
  - *šťastný, šťastná, šťastné, šťastní, šťastného, šťastnému, šťastných…*

Proč překladači vadí skloňování

- *angličtina*
  - *happy*
- *čeština*
  - *šťastný, šťastná, šťastné, šťastní, šťastného, šťastnému, šťastných…*

- *happy, unhappy, happier, unhappier, happiest, unhappiest*

= 6

- ?

Proč překladači vadí skloňování

- 11 různých tvarů
  - *šťastný šťastného šťastnému šťastném šťastným*
    šťastná šťastnou šťastné šťastní šťastných šťastnými
- + stupňování, negace (*šťastnější nejnešťastnější*)
- + jmenná přídavná jména:
  - *šťasten šťastna šťastno šťastni šťastny šťastnu*
- + nespisovné tvary:
  - *šťastnej šťastnějšim nejšťastnějšíma...*

Proč překladači vadí skloňování

- 11 různých tvarů
  - *šťastný šťastného šťastnému šťastném šťastným šťastná šťastnou šťastné šťastní šťastných šťastnými*
- + stupňování, negace (*šťastnější nejnešťastnější*)
- + jmenná přídavná jména:
  - *šťasten šťastna šťastno šťastni šťastny šťastnu*
- + nespisovné tvary:
  - *šťastnej šťastnějším nejšťastnějšíma...*
- celkem **82** tvarů: http://ufal.mff.cuni.cz/morphodita

Co s tím?

- když jste Google Translate
  - tak děláte překlad mezi 81 jazyky (3240 párů)
  - nemáte prostor se češtinou zabývat extra
  - prostě zkusíte nasbírat JEŠTĚ VÍC dat
- když jste student počítačové lingvistiky (= já)
  - můžete zkusit speciálně pro češtinu něco vymyslet
  - co třeba se vrátit k té zahozené gramatice?
    - využít to co funguje (frázový překlad)
    - jenom opravit některé chyby v gramatice

Depfix

- moje diplomová práce
- vylepšení strojového překladu z angličtiny do češtiny

1. frázový strojový překlad (jako Google Translate)
2. automatická pravidlová post-editace
   I. jazykový rozbor věty (existující nástroje)
   II. oprava gramatických chyb

Jak Depfix opravuje chyby

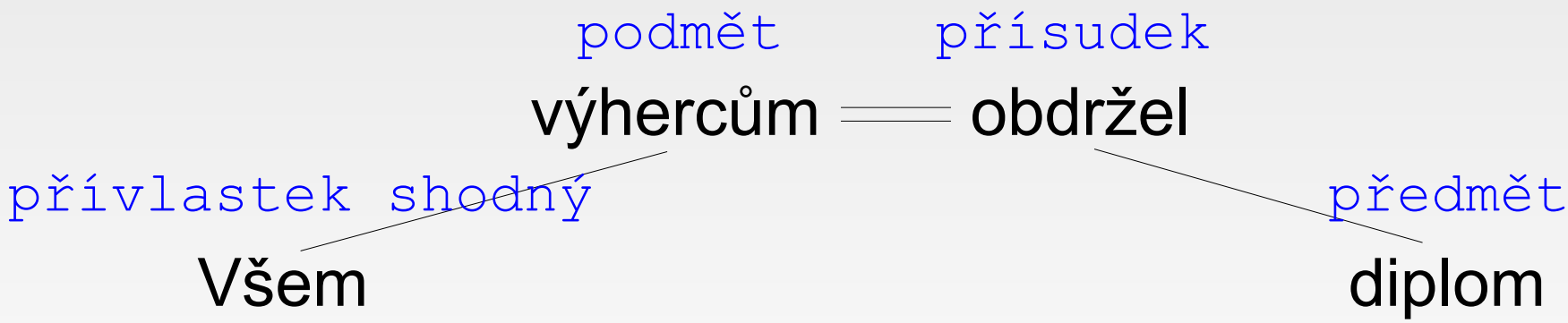
- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí frázového překladače:
  - *Všem výhercům obdržel diplom.*

Jak Depfix opravuje chyby

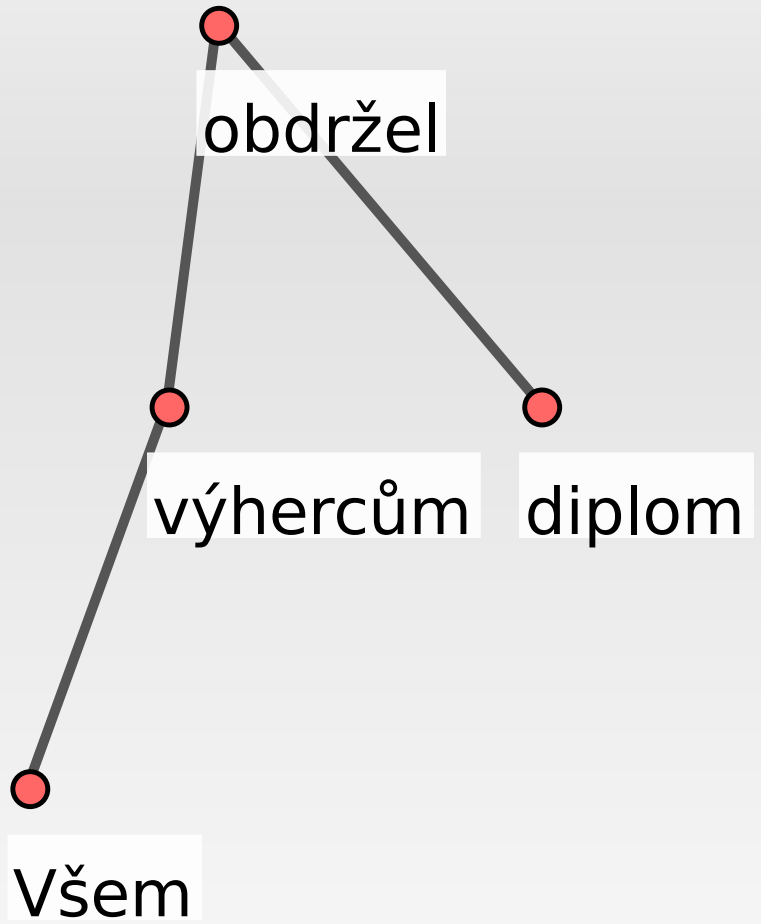
- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí frázového překladače:
  - *Všem výhercům obdržel diplom.*
- oprava pomocí systému Depfix:
  - *Všichni výherci obdrželi diplom.*

Jazykový rozbor věty

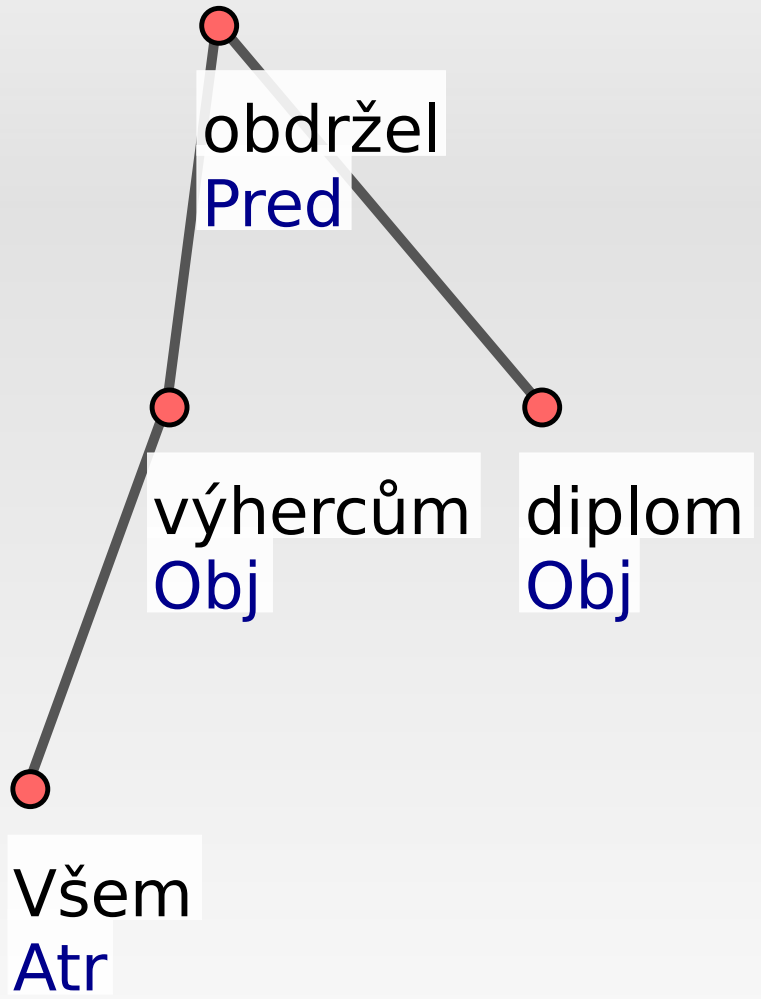
- slovní druhy, mluvnické kategorie
  - *Všem **výhercům** obdržel diplom.*
  - lemma *výherce*, podst. jm., 3. p., mn. č., rod m. živ.
- větný rozbor, větné členy



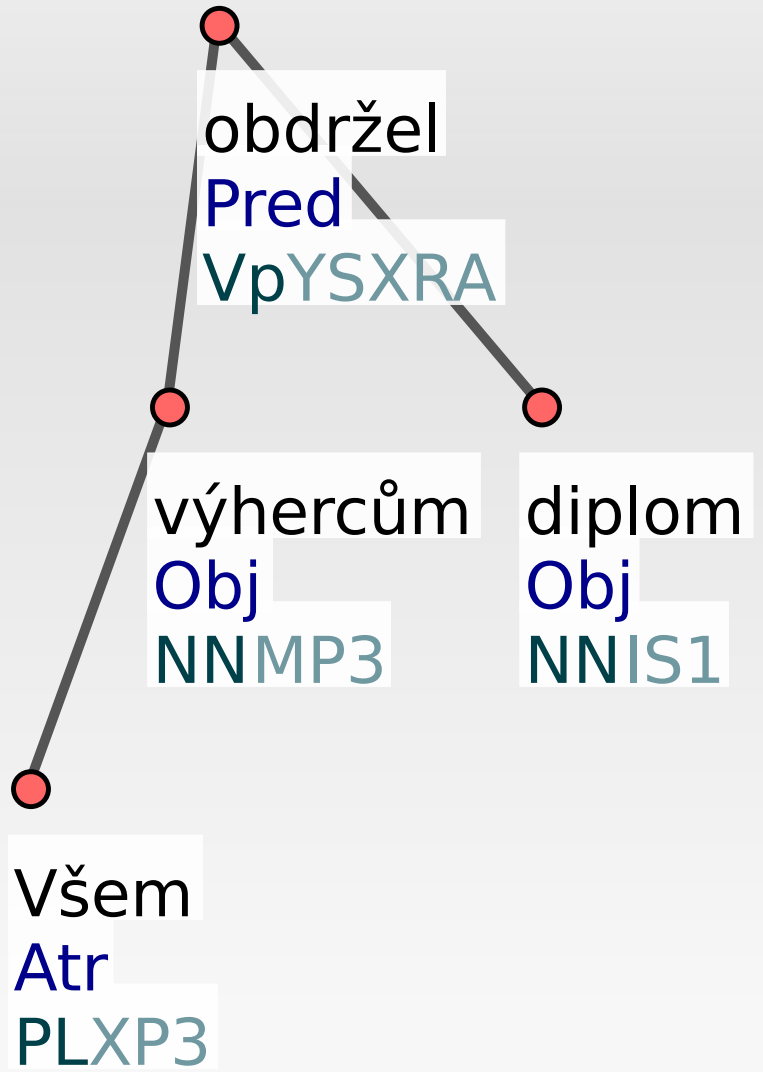
Všem výhercům obdržel diplom.



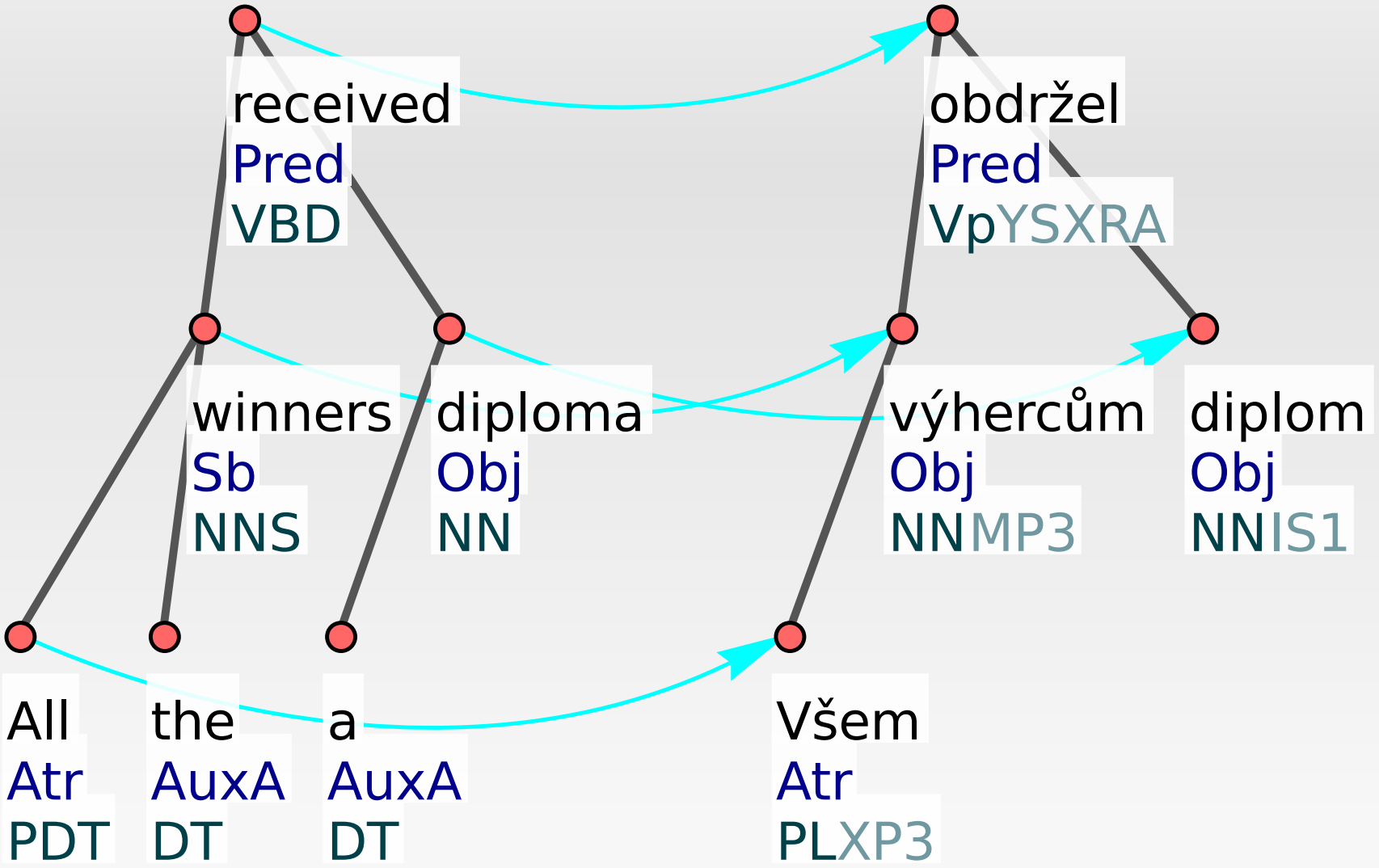
Všem výhercům obdržel diplom.



Všem výhercům obdržel diplom.



Všem výhercům obdržel diplom.



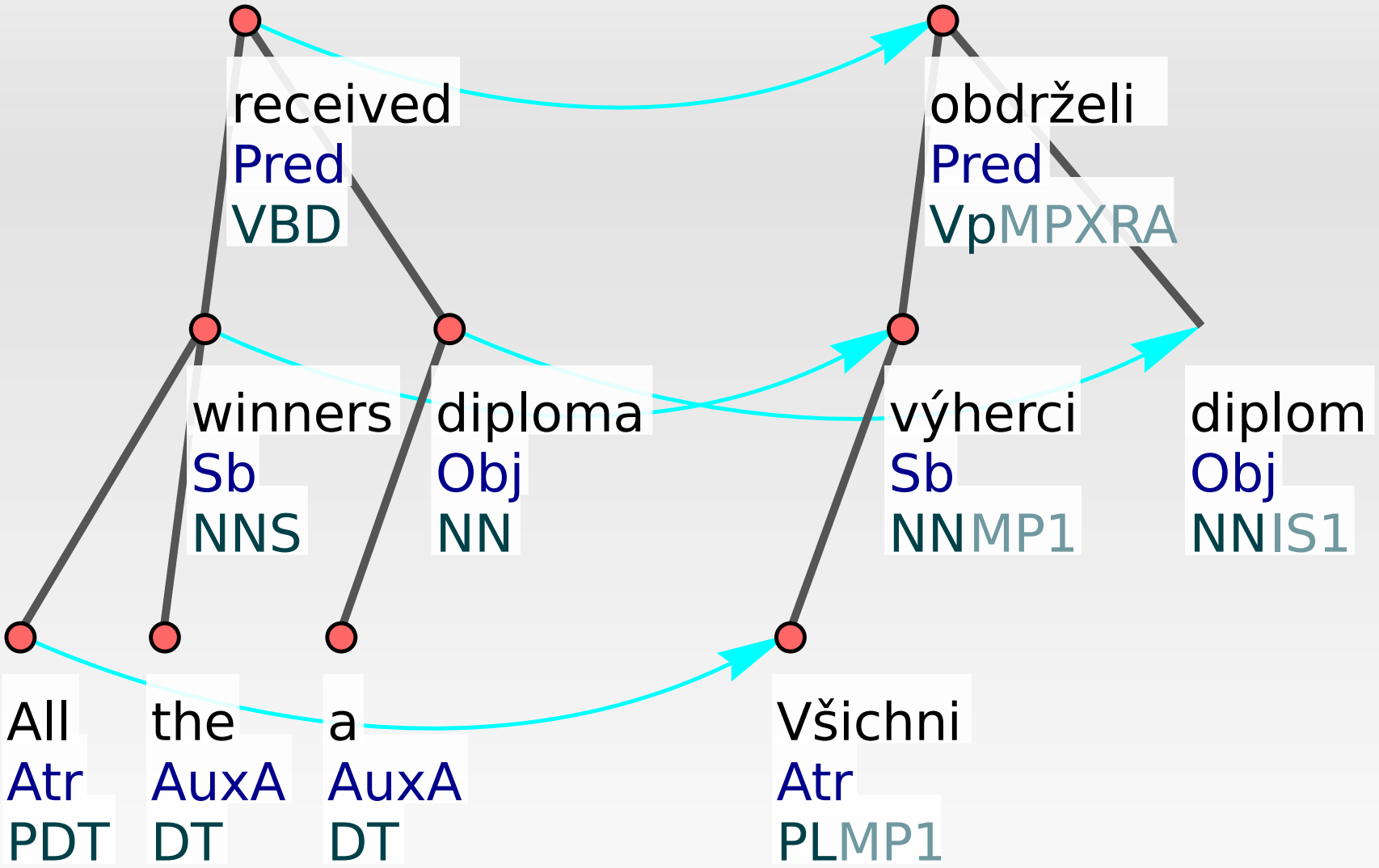
Oprava gramatických chyb

- podmět musí být v 1. pádě
- shoda přívlastku s podstatným jménem
- shoda podmětu s přísudkem
- oprava chybějící negace
- oprava zvratnosti
- oprava slovesné a jmenné valence
- zachování slovesného času
- …

Oprava gramatických chyb

- **podmět musí být v 1. pádě**
- **shoda přívlastku s podstatným jménem**
- **shoda podmětu s přísudkem**
- oprava chybějící negace
- oprava zvratnosti
- oprava slovesné a jmenné valence
- zachování slovesného času
- …

Všichni výherci obdrželi diplom.



Jak Depfix opravuje chyby

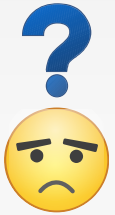
- anglická věta:
  - *All the winners received a diploma.*
- překlad pomocí frázového překladače:
  - *Všem výhercům obdržel diplom.*
- oprava pomocí systému Depfix:
  - *Všichni výherci obdrželi diplom.*

Máme nejlepší překlad do češtiny!

	Země	Instituce	Systém	Skóre
1.	ČR	Matfyz	Moses + Depfix	0,371
2.	UK	Edinburgh Uni	Moses	0,356
3.	ČR	Matfyz	Moses	0,333
4.	USA	Google	Google Translate	0,169
5.	USA	Microsoft	Bing Translator	0,030
6.	ČR	Microton	Eurotran	-0,534

Výsledky soutěže ve strojovém překladu WMT 2014
(Workshop on Statistical Machine Translation),
překlad z angličtiny do češtiny, kvalita překladu hodnocena lidmi.

Závěr

- Vytvoření strojového překladače je snadné
  - překlad po frázích 
  - texty postahované z internetu 
- Depfix: Pokud chceme dobře překládat do češtiny, vyplatí se zapojit i gramatiku 
  - moje diplomová práce a zaměstnání 
  - prezentoval jsem na konferencích (USA, Korea, Bulharsko, Itálie...) 
- Spousta problémů zůstává nedořešených... 

Motivace pro vás

- Tohle a mnoho jiného se dělá na Matfyzu
  - hlasové dialogové systémy, droni, roboti, hry...
  - informatika, matematika, fyzika, učitelství M/F/I
- Přijďte na Den otevřených dveří 26.11.
  - a na další akce (viz letáky/web)
  - zapojte se do korespondenčních seminářů
- Nebojte se k nám jít studovat! 
  - je to náročné, ale zajímavé – stojí to za to

Děkuji za pozornost

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Jak dělat strojový překlad lépe než Google Translate

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

Tato prezentace a další informace:

http://ufal.mff.cuni.cz/rudolf-rosa/