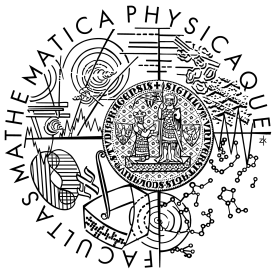


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

ÚFAL

Modelování závislostní syntaxe napříč jazyky

Hlavní řešitel: Mgr. Rudolf Rosa

Vedoucí: doc. Ing. Zdeněk Žabokrtský Ph.D.

Spoluřešitelé: Mgr. Jan Mašek
Mgr. Martin Popel

Pracoviště: Ústav formální a aplikované lingvistiky MFF UK

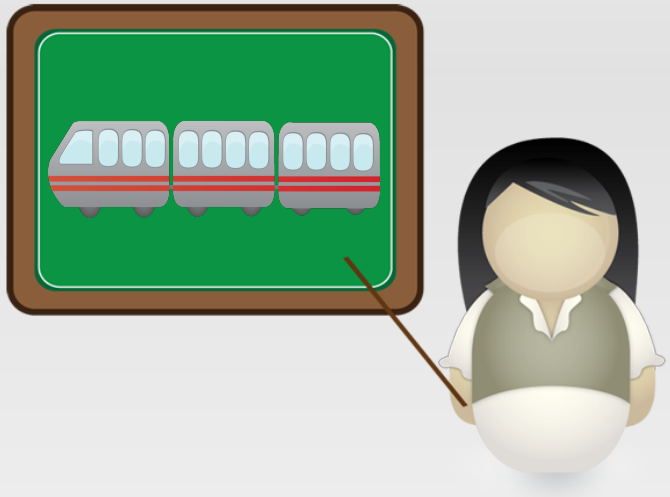
Kulatý stůl GAUK, Praha, 10. dubna 2018

Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

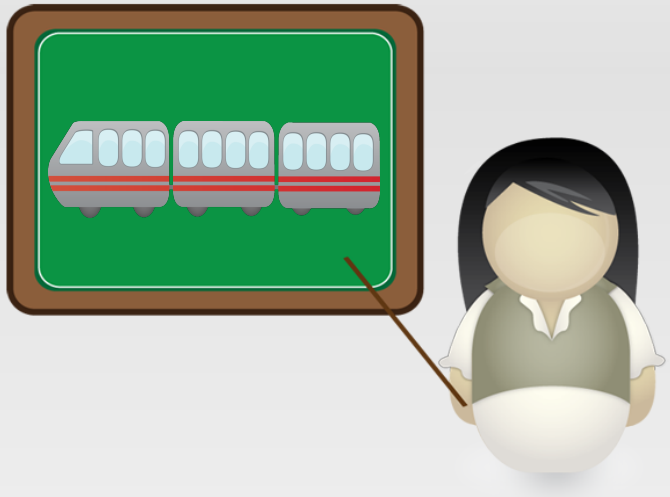


Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

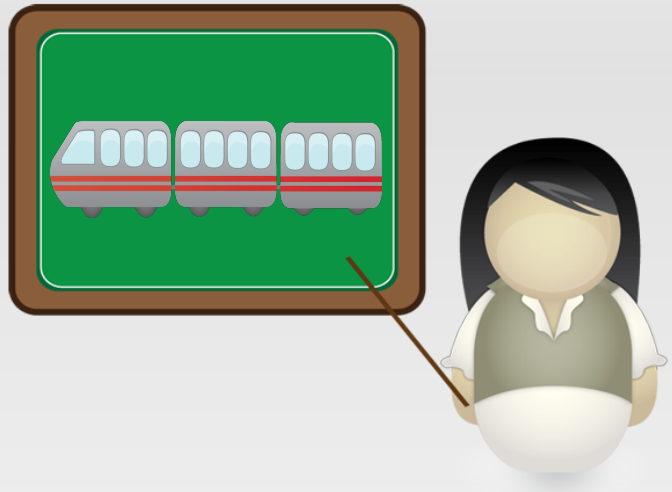
podmět přísudek slovesný

vlak ═══ přijel



Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.



podmět — přísudek slovesný

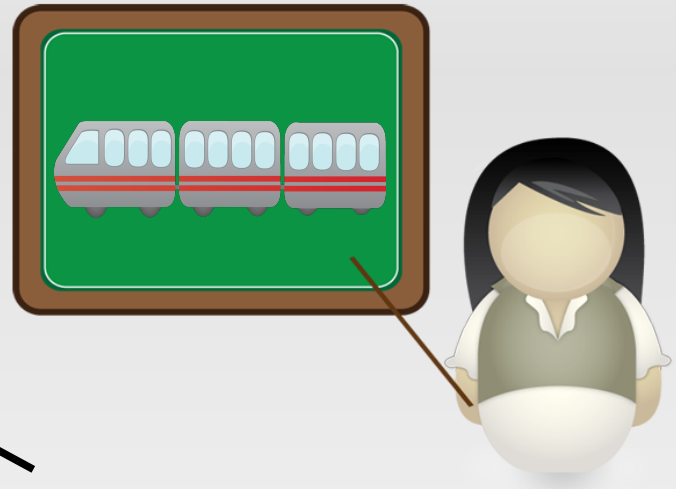
vlak ═══ přijel

přívlastek shodný

Dlouhý

Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.



podmět — vlak

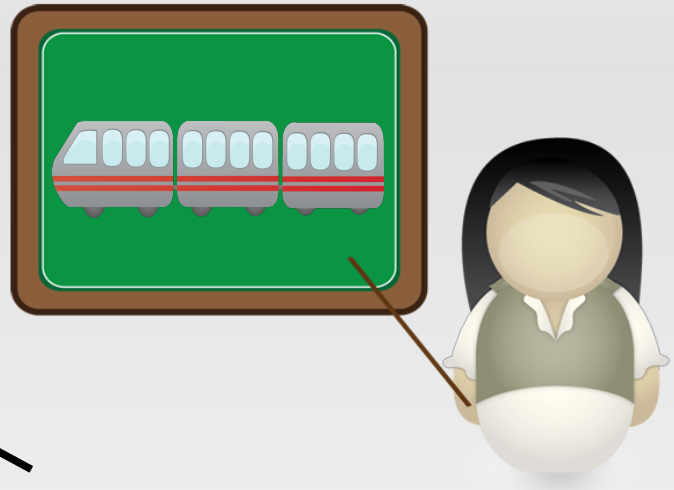
přísudek slovesný — přijel

přívlastek shodný — Dlouhý

příslovečné určení místa — na nádraží

Závislostní syntaxe = větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.



podmět — vlak

přísudek slovesný — přijel

přívlastek shodný — Dlouhý

příslovečné určení místa — na nádraží

přívlastek neshodný — v Berouně

Automatický větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

Automatický větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

Automatický větný rozbor

Dlouhý vlak přijel na nádraží v Berouně.

podmět
vlak

přísudek slovesný
přijel

přívlastek shodný
Dlouhý

příslovečné určení místa
na nádraží

přívlastek neshodný
v Berouně

Problém multilingvality

- strojové učení
  - potřebuje ručně vytvořená data
  - tisíce větných rozborů
  - tisíce hodin práce jazykových odborníků

Problém multilingvality

- strojové učení
  - potřebuje ručně vytvořená data
  - tisíce větných rozborů
  - tisíce hodin práce jazykových odborníků

- dostupná data
  - 2014: 30 jazyků
  - 2018: 60 jazyků

Problém multilingvality

- strojové učení
  - potřebuje ručně vytvořená data
  - tisíce větných rozborů
  - tisíce hodin práce jazykových odborníků

- dostupná data
  - 2014: 30 jazyků
  - 2018: 60 jazyků
- odhad počtu živých jazyků
  - asi 7000 jazyků

Přenos nástrojů napříč jazyky

- čeština
- španělština
- ruština
- turečtina
- hindština
- urdština
- …

Přenos nástrojů napříč jazyky

- čeština
- španělština
- ruština
- turečtina
- hindština
- urdština
- …

- hornolužická srbština
- kazaština
- bengálština
- telugština
- bambarština
- jorubština
- …

Přenos nástrojů napříč jazyky

- čeština
- španělština
- ruština
- turečtina
- hindština
- urdština
- …

- hornolužická srbština
- kazaština
- bengálština
- telugština
- bambarština
- jorubština
- …

Přenos nástrojů napříč jazyky

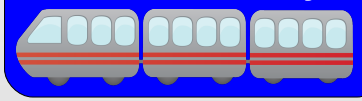
- čeština
- španělština
- ruština
- turečtina
- hindština
- urdština
- …

- hornolužická srbština
- kazaština
- bengálština
- telugština
- bambarština
- jorubština
- …

Které jazyky jsou si podobné?

čeština

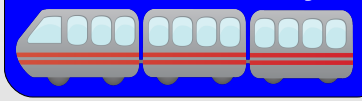
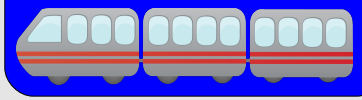
začátek věty



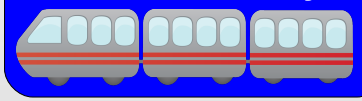
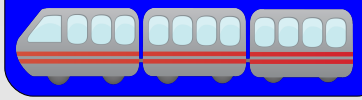
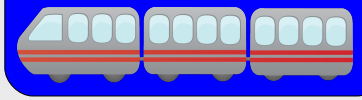
přídavné jméno
dlouhý

podstatné jméno
vlak

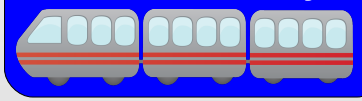
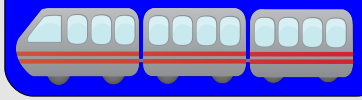
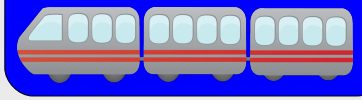
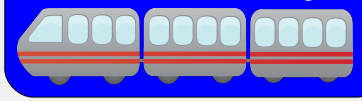
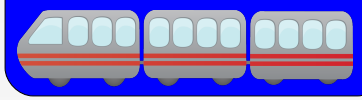
Které jazyky jsou si podobné?

čeština	začátek věty 	přídavné jméno dlouhý	podstatné jméno vlak	
angličtina	začátek věty 	člen a	přídavné jméno long	podstatné jméno train

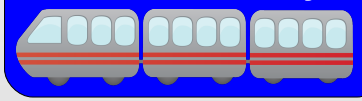
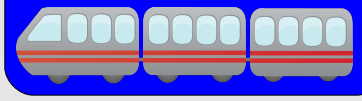
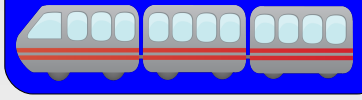
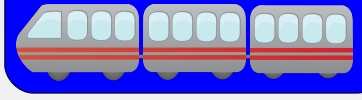
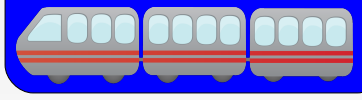
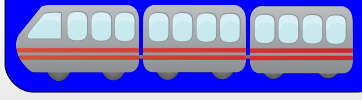
Které jazyky jsou si podobné?

čeština	začátek věty 	přídavné jméno dlouhý	podstatné jméno vlak	
angličtina	začátek věty 	člen a	přídavné jméno long	podstatné jméno train
němčina	začátek věty 	člen ein	přídavné jméno langer	podstatné jméno Zug

Které jazyky jsou si podobné?

čeština	začátek věty 	přídavné jméno dlouhý	podstatné jméno vlak	
angličtina	začátek věty 	člen a	přídavné jméno long	podstatné jméno train
němčina	začátek věty 	člen ein	přídavné jméno langer	podstatné jméno Zug
italština	začátek věty 	člen un	přídavné jméno lungo	podstatné jméno treno
	začátek věty 	člen un	podstatné jméno treno	přídavné jméno lungo

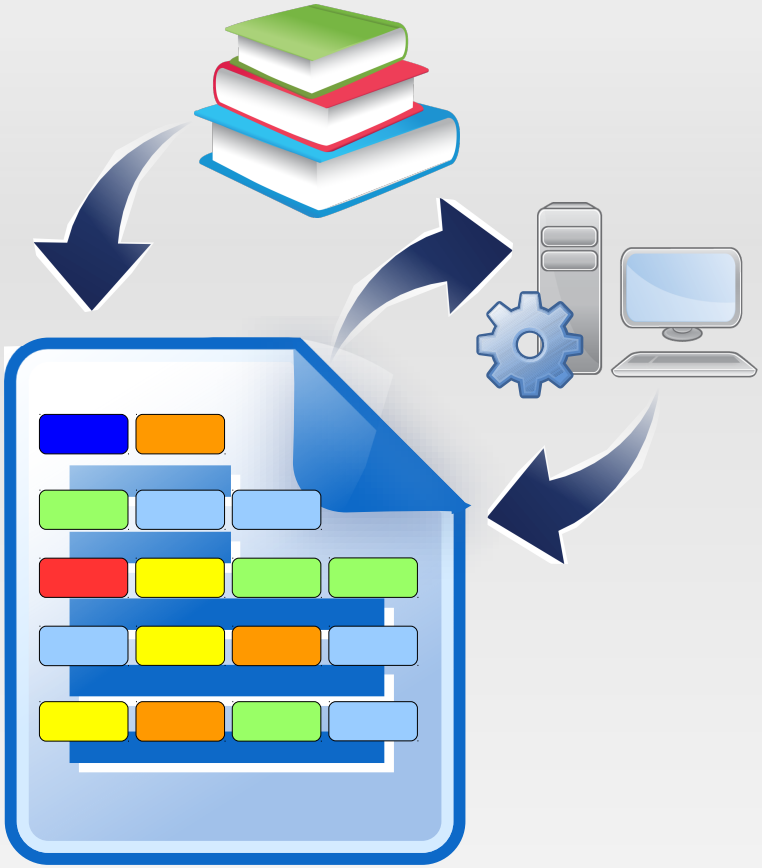
Které jazyky jsou si podobné?

čeština	začátek věty 	přídavné jméno dlouhý	podstatné jméno vlak	
angličtina	začátek věty 	člen a	přídavné jméno long	podstatné jméno train
němčina	začátek věty 	člen ein	přídavné jméno langer	podstatné jméno Zug
italština	začátek věty 	člen un	přídavné jméno lungo	podstatné jméno treno
	začátek věty 	člen un	podstatné jméno treno	přídavné jméno lungo
turečtina	začátek věty 	přídavné jméno uzun	člen bir	podstatné jméno tren

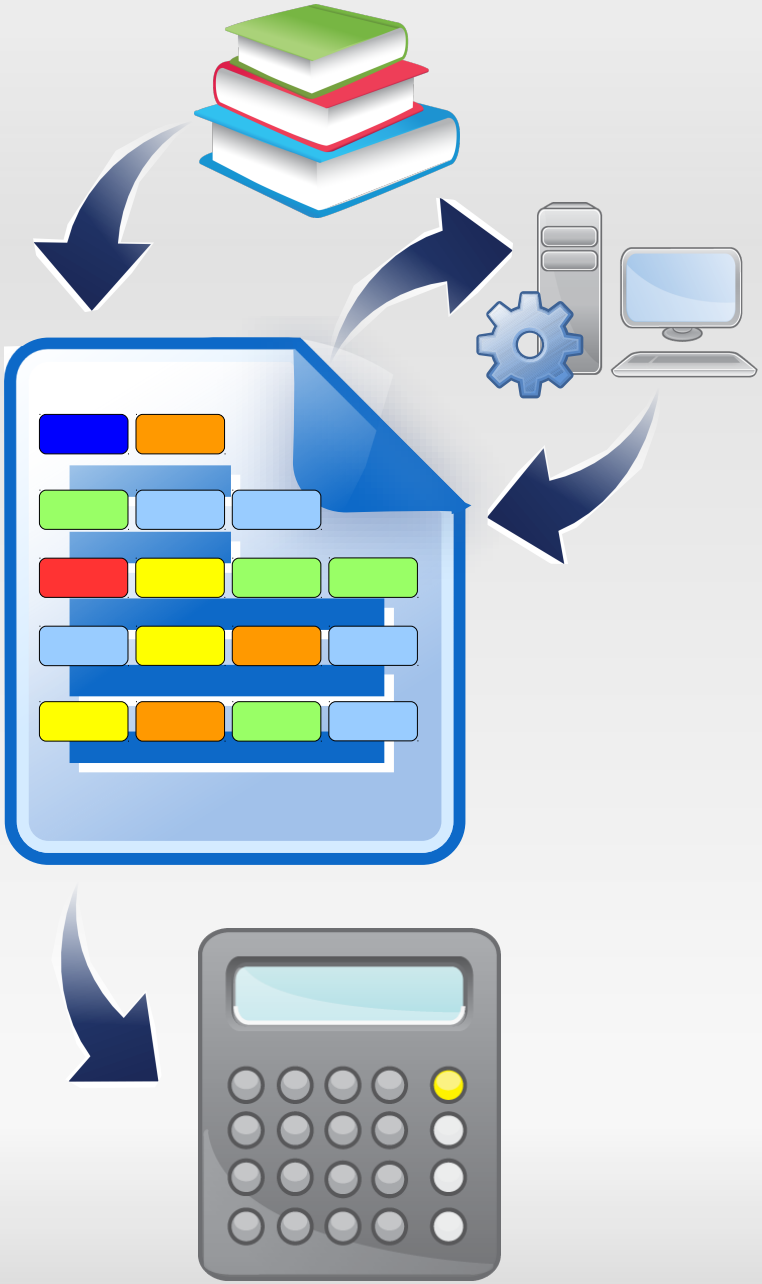
Které jazyky jsou si podobné?

Měření podobnosti jazyků

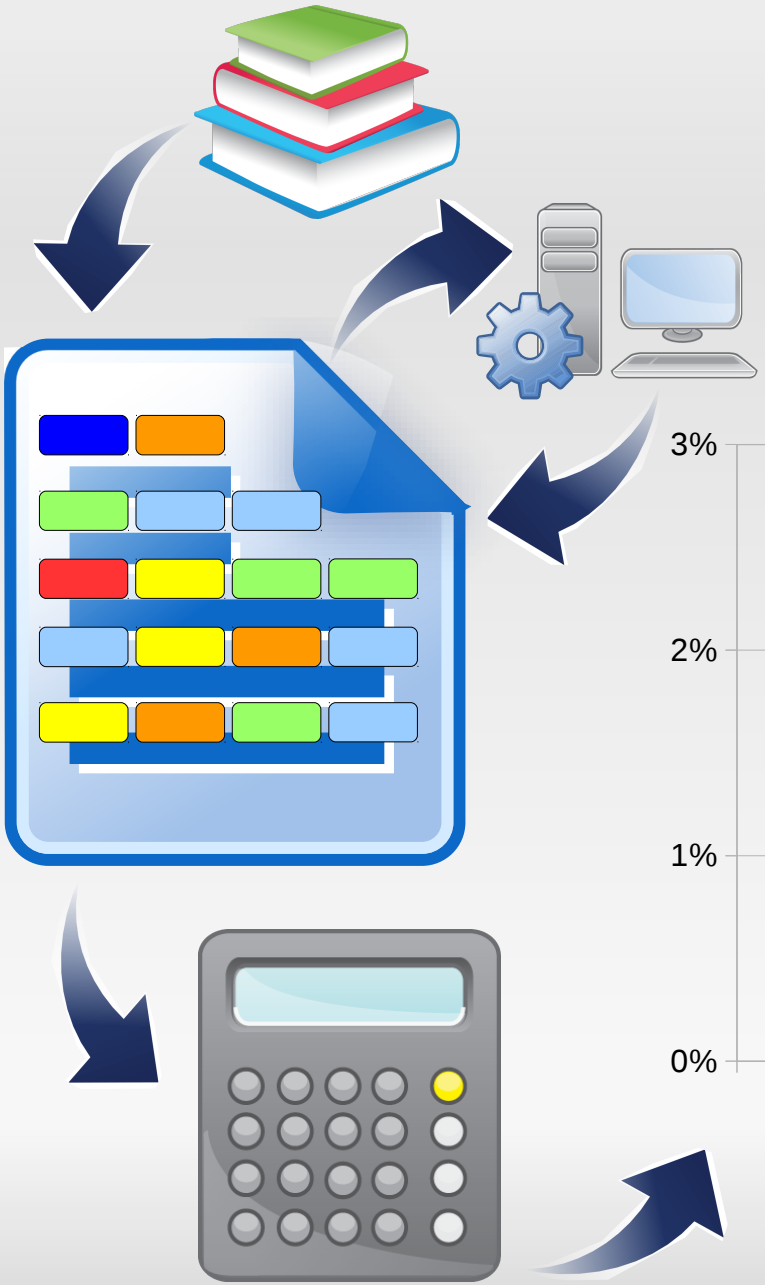
Měření podobnosti jazyků

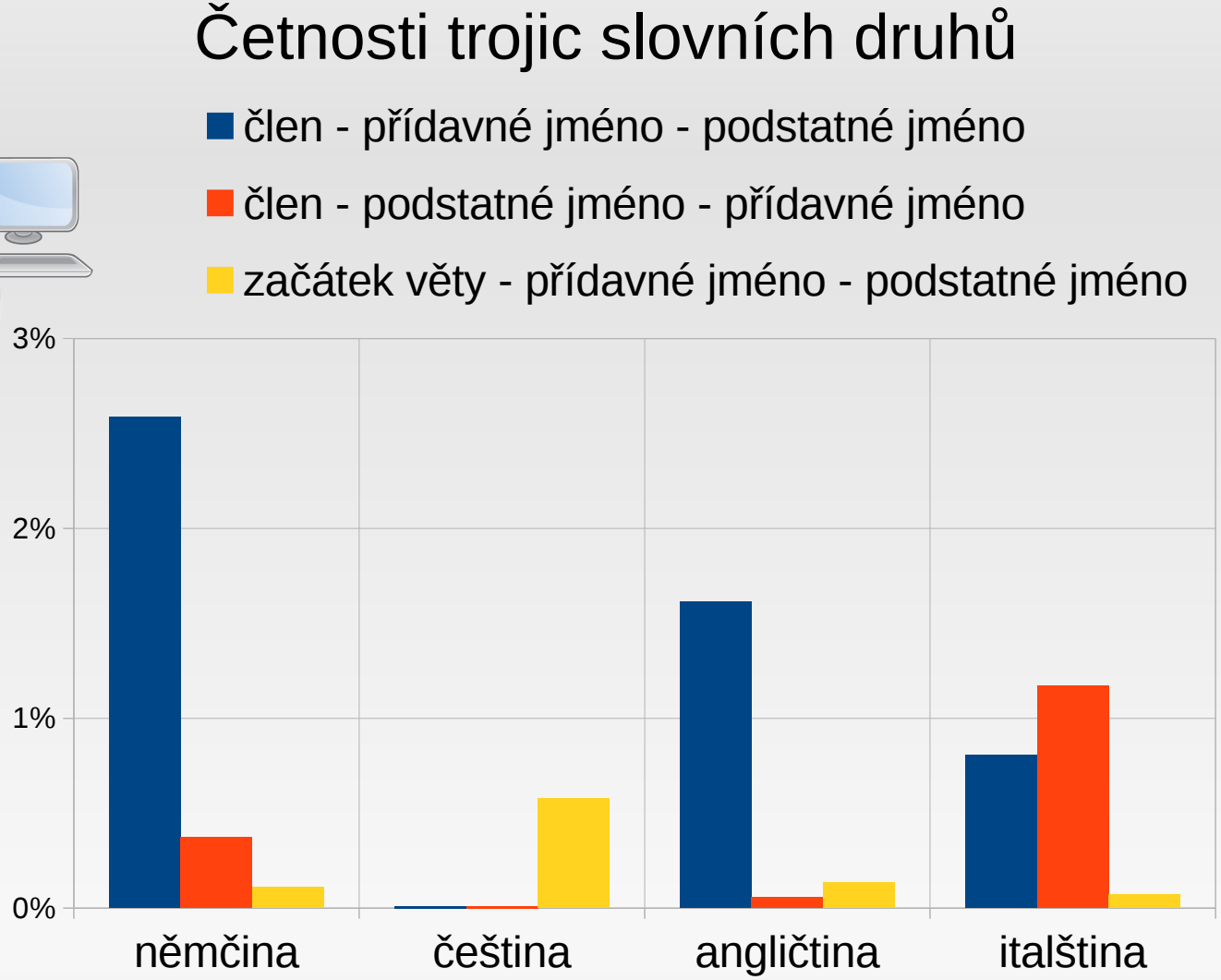


Měření podobnosti jazyků

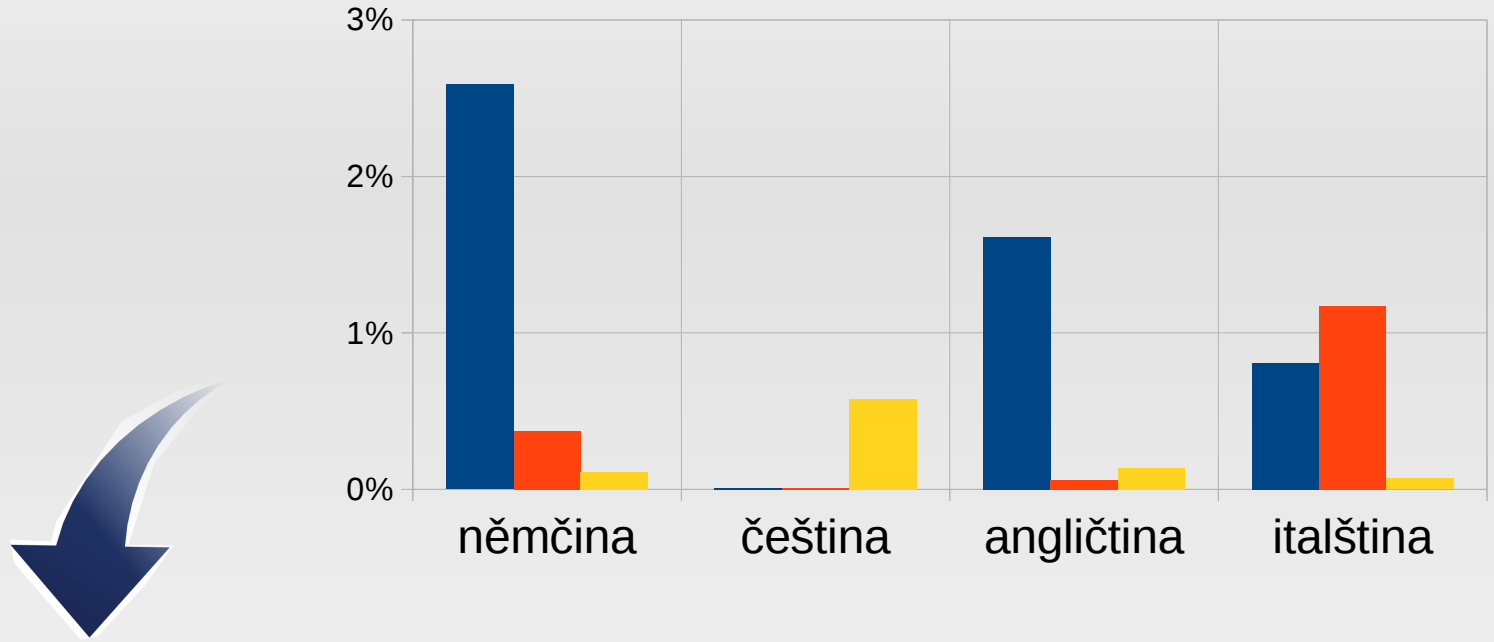


Měření podobnosti jazyků





Měření podobnosti jazyků



$$KL_{cpos^3}(tgt, src) = \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log\left(\frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)}\right)$$

- Kullback-Leiblerova divergence rozložení četnosti trojic slovních druhů

Úspěchy projektu

- nová metoda měření podobnosti jazyků
- inovovaná metoda kombinace jazyků
- světová úroveň mezijazyčného přenosu
  - vítězství v soutěži VarDial 2017
  - publikace na mezinárodních konferencích
    - včetně nejvýznamnější oborové konference (ACL)

- podíl na projektu Universal Dependencies
  - 200 členů, data a nástroje pro 60 jazyků

Pozitiva a negativa GAUKů

Pozitiva GAUKů

- radost
- vlastní peníze na cestování
  - konference (8)
  - studijní pobyty (1)
  - letní školy (0)
- stipendium
- nenáročná správa a podmínky

Problémy GAUKů: málo informací

- oficiální informace velice stručné
  - nejsou ukázky žádostí, posudků, co lze, co nelze...
  - neoficiální pokoutné informace
  - já: vše na webu – žádost, zprávy...
- problémy s nedočerpaným cestovným
  - článek nepřijat na konferenci → peníze zbydou
  - hospodárné: vrátit – není preferováno
  - nehospodárné: utratit za „hlouposti”
  - rozumné ale málo známé: nákup vybavení, studijní pobyt na univerzitě, přeúčtování spoluřešitelům...

Děkuji za pozornost

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Modelování závislostní syntaxe napříč jazyky

Univerzita Karlova
Matematicko-fyzikální fakulta
Ústav formální a aplikované lingvistiky

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