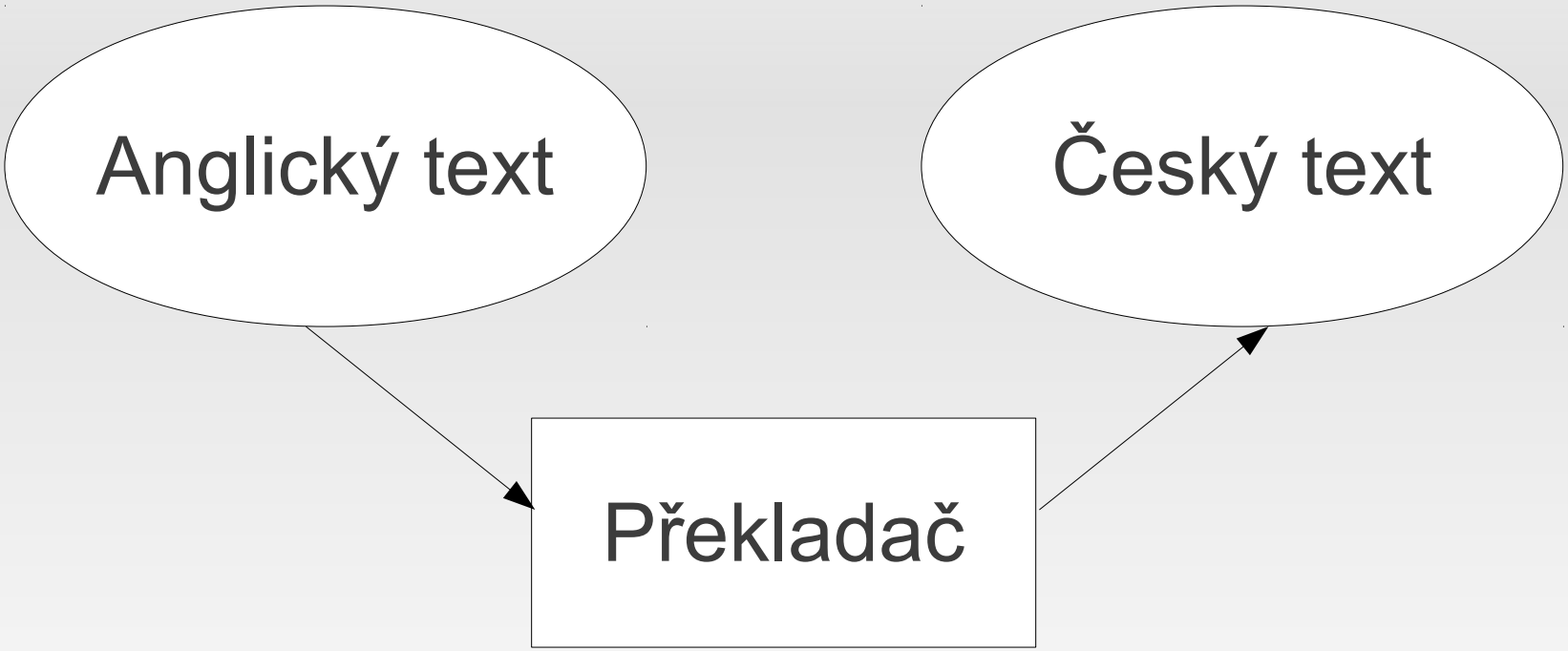
Rudolf Rosa

Strojový překlad
pojmenovaných entit
za pomoci Wikipedie

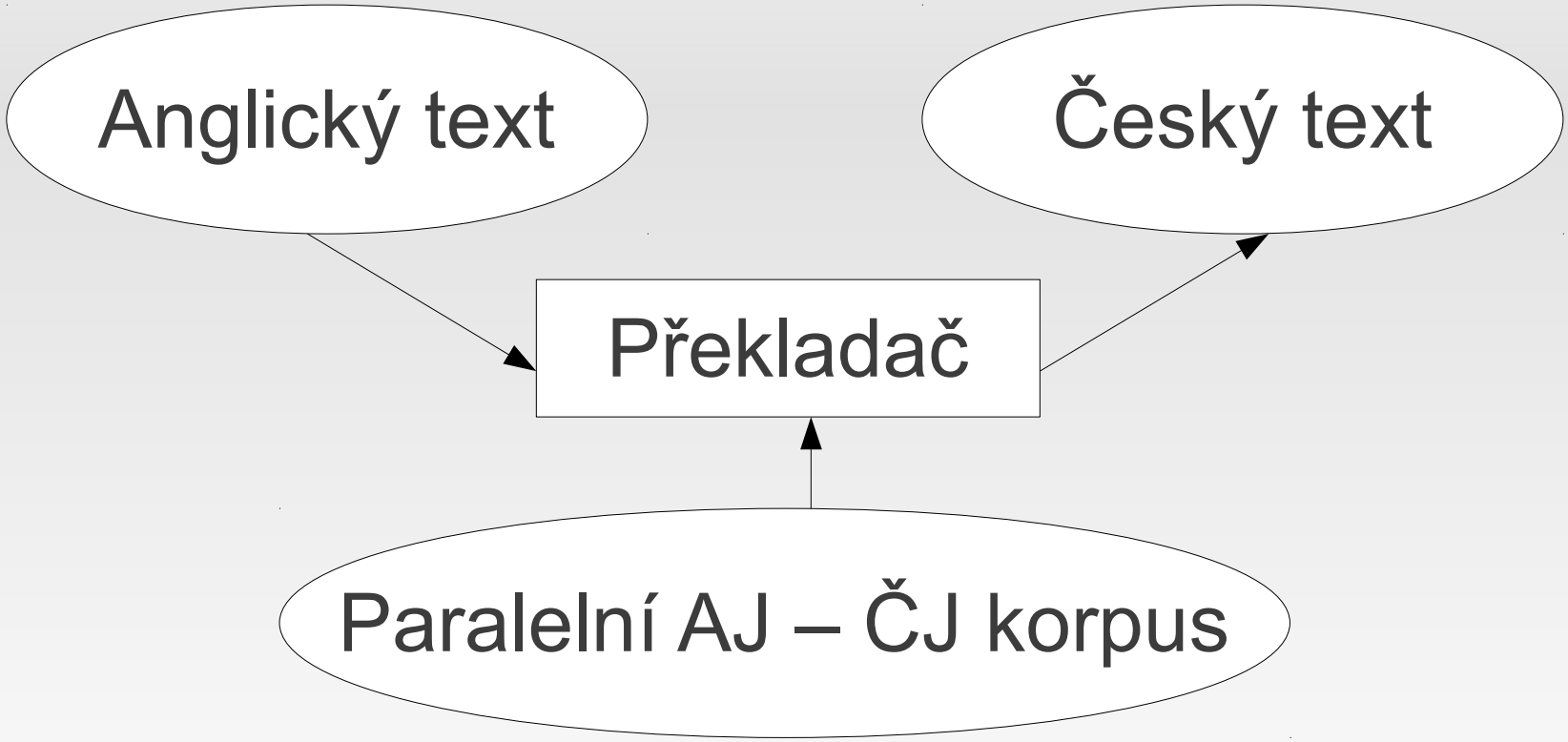
Obsah

- Strojový překlad
  - Statistický strojový překlad
  - Frázový statistický strojový překlad
  - Překlad pojmenovaných entit
- O. Hálek, R. Rosa, A. Tamchyna
  - Rozpoznání pojmenovaných entit
  - Překlad pojmenovaných entit
  - Průběžné výsledky

Strojový překlad



Statistický strojový překlad



Paralelní AJ – ČJ korpus (CzEng)

```
<s id='en-p29s2'>
  <w id='en-p29s2w1'>
    Everything</w>
  <w id='en-p29s2w2'>
    was</w>
  <w id='en-p29s2w3'>
    so</w>
  <w id='en-p29s2w4'>
    beautiful</w>
  <w id='en-p29s2w5'>
    !</w>
</s>
```

```
<s id='cs-p29s2'>
  <w id='cs-p29s2w1'>
    Všechno</w>
  <w id='cs-p29s2w2'>
    bylo</w>
  <w id='cs-p29s2w3'>
    tak</w>
  <w id='cs-p29s2w4'>
    krásné</w>
  <w id='cs-p29s2w5'>
    !</w>
</s>
```

Segmentace – jednotlivá slova

- Korpus (AJ)
  - Yesterday ~~I~~ was in ~~the~~ cinema.
  - He ~~is~~ going ~~to~~ sleep.
- Vstup (AJ)
  - He was going ~~to~~ ~~the~~ cinema.

- Korpus (ČJ)
  - Včera jsem byl v kině.
  - On bude spát.
- Výstup (ČJ???)
  - On jsem byl bude kině.

Frázový statistický strojový překlad

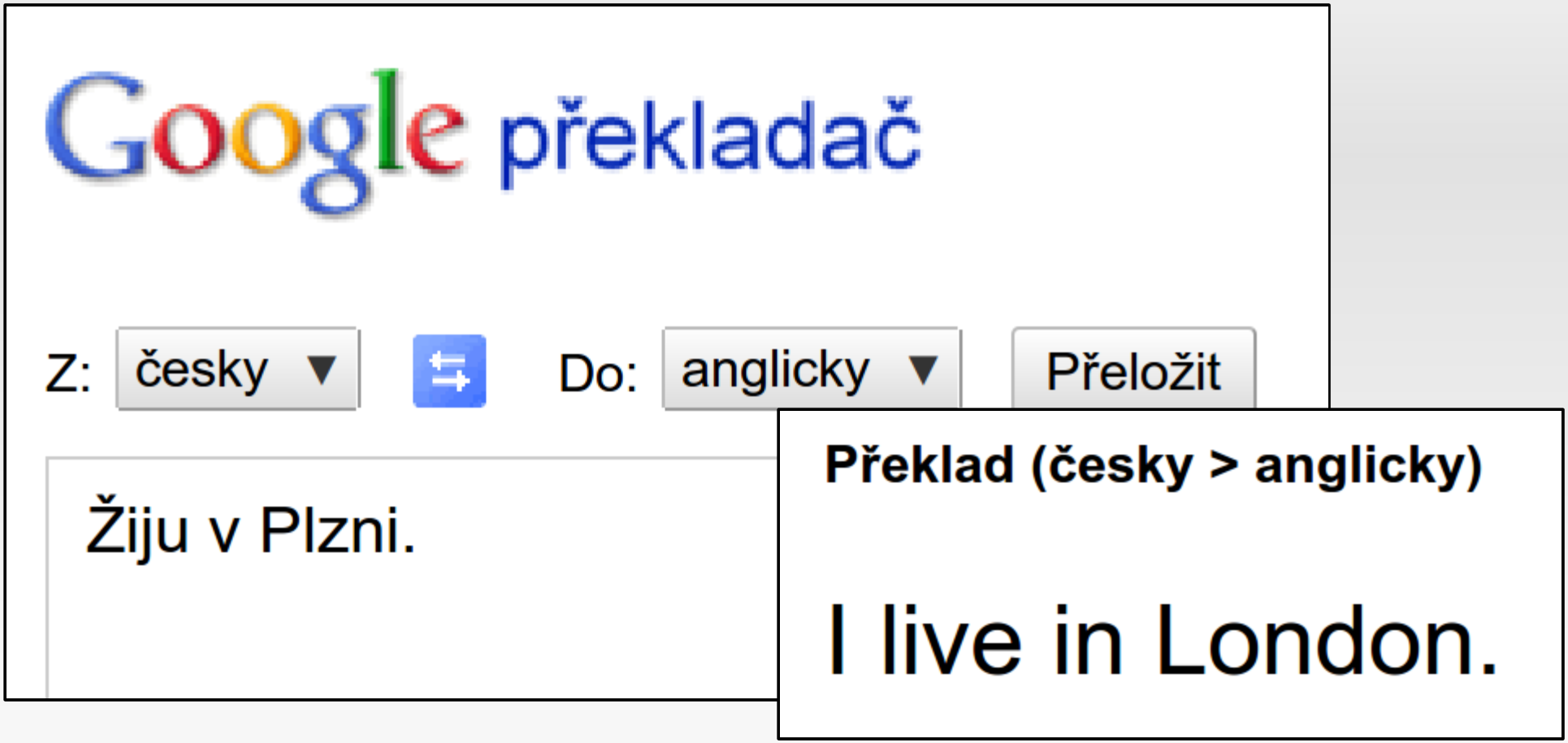
- Vstup (AJ)
  - Yesterday
  - I was
  - in the cinema
  - .
- Výstup (ČJ)
  - Včera
  - jsem byl
  - v kině
  - .

Překlad pojmenovaných entit

- **Rice University** is at 6100 **Main Street**.
- **Steven Bird** passed on the editorship...
- Exit at **Government Plaza Station** on **5th Street**.
- **fork()** creates a new process.

- **Univerzita rýže** je v 6100 **hlavní ulici**.
- **Steven pták** přenesl na editorship...
- Konec **vlády plaza na nádraží** v **páté třídě**.
- **vidlička()** vytváří nový proces.

Google překladač



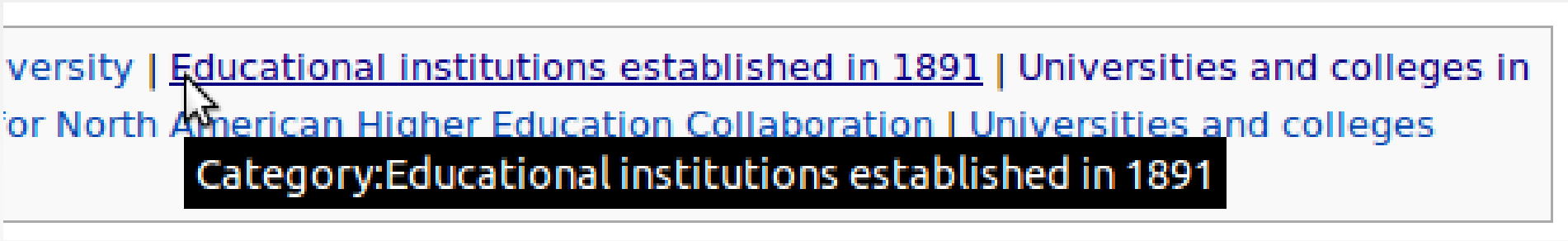
O. Hálek, R. Rosa, A. Tamchyna

- Strojový překlad pojmenovaných entit za pomoci Wikipedie
  - překlad z angličtiny do češtiny
- Rozpoznání pojmenovaných entit
  - podle kategorií anglického článku na Wikipedii
- Překlad pojmenovaných entit
  - podle titulku odpovídajícího českého článku

Rozpoznání pojmenovaných entit

- Vybrat fráze, které mohou být pojmenovanou entitou
  - **Rice University** is at 6100 **Main Street**.
- Zjistit kategorie článku na Wikipedii
- Prohledat (do šířky) nadřazené kategorie
  - Ručně vytvořený seznam kategorií obsahujících pojmenované entity

Zjištění (všech) kategorií



Zjištění kategorií – WikiMedia API

- http://en.wikipedia.org/w/api.php?action=query&prop=**categories**&redirects&clshow=!hidden&format=**xml**&titles=**Rice_University**

-
```
<?xml version="1.0"?>
<api><query><pages>
  <page pageid="25813" ns="0"
    title="Rice University">
    <categories>
      <cl ns="14" title="Category:Association
        of American Universities" />
      <cl ns="14" title="Category:Educational
        institutions established in 1891" />
```
...

Prohledání nadřazených kategorií

- **Educational institutions established in 1891**
- Educational institutions established in the 1890s
- Educational institutions established in the 19th century
- Educational institutions by year of establishment
- Organizations by year of establishment
- **Organizations**

Kategorie pojmenovaných entit

- Places („Místa“ – není na české Wikipedii)
- People (Lidé)
- Organizations (Organizace)
- Companies (Firmy)
- Software (Software)
- Transport infrastructure (Dopravní stavby)

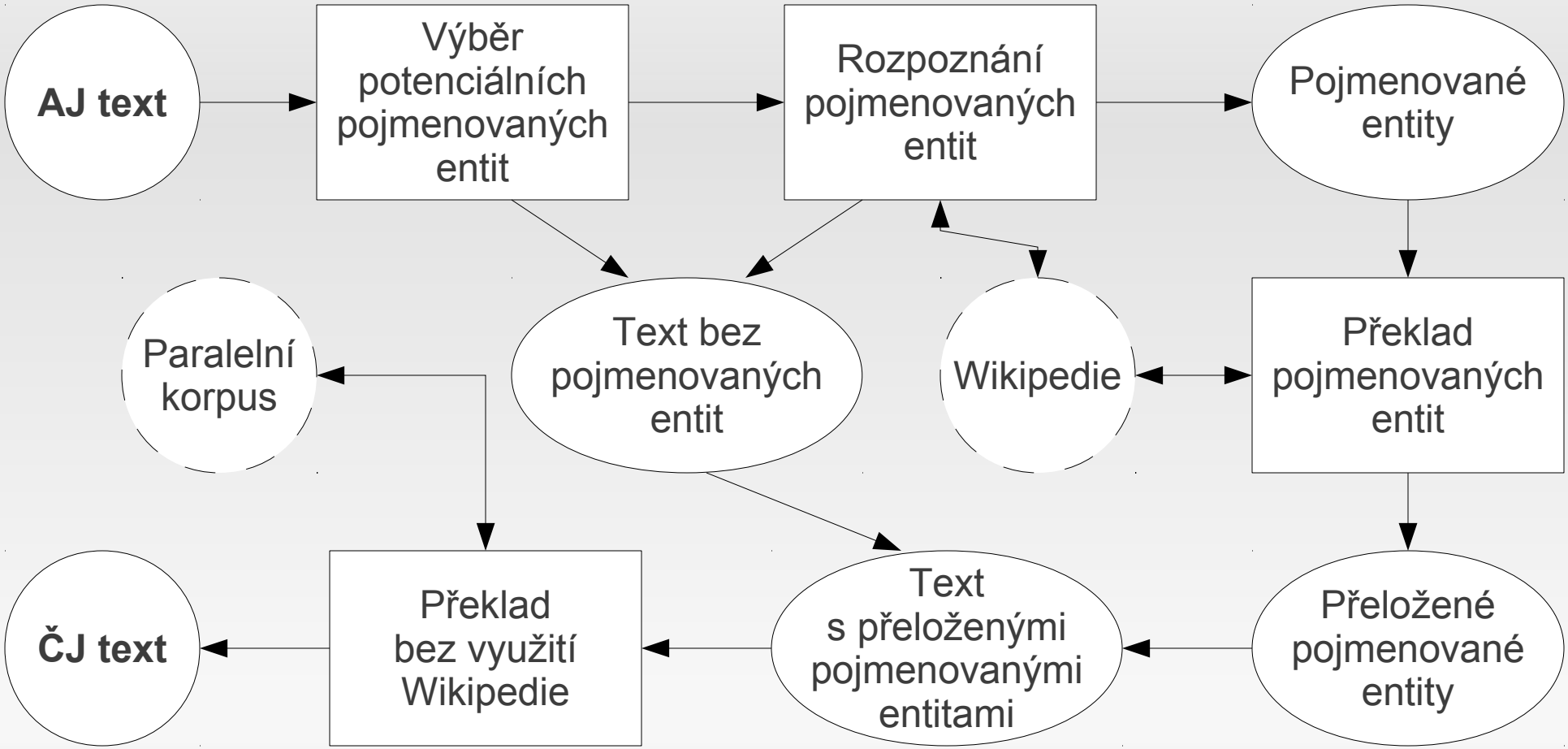
Překlad pojmenovaných entit

Předpokládáme, že jde o pojmenovanou entitu

- Zjistit, zda existuje článek na anglické Wikipedii
- Podívat se, zda existuje jeho český ekvivalent
- Použít název českého článku jako překlad anglické pojmenované entity

Překlad entity „Spain“

Hálek-Rosa-Tamchyna – shrnutí



Průběžné výsledky

Pojmenované entity bez českého článku	Překlad entit s českým článkem	BLEU skóre
Ponechat anglicky	Vždy dle Wikipedie	20,95
	Wikipedie/korpus	21,46
Pokusit se přeložit v paralelním korpusu	Vždy dle Wikipedie	21,49
	Wikipedie/korpus	21,82
Původní překlad (bez použití Wikipedie)		22,55

Reference

- Ondřej Bojar: *NPFL087 Statistický strojový překlad*
  - http://www1.cuni.cz/~obo/vyuka/
- Wikipedia, The Free Encyclopedia: *Named entity recognition*
  - http://en.wikipedia.org/wiki/Named_entity_recognition
- MediaWiki: *MediaWiki API documentation*
  - http://www.mediawiki.org/wiki/API:Main_page
- Ondřej Bojar, Zdeněk Žabokrtský: *CzEng, Large Parallel Treebank with Rich Annotation*
  - http://ufal.mff.cuni.cz/czeng/

Děkuji za pozornost

Tato prezentace je dostupná na adrese
http://mff.nikde.eu/