Parallel Corpus from Wikipedia

Adéla Štromajerová, Vít Baisa, Marek Blahuš

Faculty of Arts, Faculty of Informatics, Masaryk University

December 2, 2016

Aim

- to create a parallel Czech-English corpus from Wikipedia
- Czech articles translated from English Wikipedia and the corresponding English articles

Steps

- download Czech and English articles
- extract parallel sentences
- prepare data for Sketch Engine

Retrieval of Wikipedia Articles

- identification of the required articles
  - *MediaWiki API* (wikidumps not very suitable), *mwclient*
  - Czech articles: template "Šablona: Překlad"
  - English articles: revision ID
- retrieval of the articles
  - html format
  - *jusText* for boilerplate removal
  - unnecessary parts removed (References, External links, ...)

Parallel Sentences Extraction

- sentence segmentation
- *hunalign*, a statistical dictionary used
- score of the alignment > 0.5
- manually assessed quality of translations: $> 90\%$
- variable proportion of the translated texts

Preparation of Data for the Corpus

- vertical text
  - prevertical files
  - tokenization with *unitok*
  - part-of-speech tagging and morphological disambiguation – *majka* and *Desamb*, *TreeTagger*
- configuration files
  - general info, location, structure, display, word classes and lemmas, dynamic attributes, etc.
- mapping file .align, m:n alignments
  - conversion from the *hunalign* format to the format required for the corpus

Result I

- *Czech Wikipedia Parallel Corpus*
- *English Wikipedia Parallel Corpus*
- Size of the corpora
  - tokens: English 46,238,455, Czech 18,785,688
  - sentences: English: 5,061,518, Czech: 2,693,657
- Size of the aligned content
  - words: English – 7,275,092, Czech – 6,414,841
  - sentences: English – 1,200,833, Czech – 1,189,910

Result II

Metadata

- Author 1,687
- Categories 45,214
- Date 3,308
- Revision ID 34,429
- Title 34,429
- URL 34,429

Categories: Málo dotčené druhy, Američtí herci, Americké herečky, Hudební skupiny 2000-2009, Členové Knesetu, ATP World Tour 2015, Hudební skupiny 1990-1999, Americké televizní seriály, Údržba:Specifikovat cíl odkazu v šabloně Wikidruhy, Hudební skupiny 2010-2019, Alba v angličtině, ATP World Tour 2014, Narození 1987, ATP World Tour 2013, Narození 1985, WTA Tour 2014, Americké filmové komedie, Americké dramatické filmy, Narození 1986, WTA Tour 2013, Ministři vlád Izraele, Epizody seriálu Glee, LGBT práva podle zemí, Města ve Washingtonu, Izraelští Židé, Americké zpěvačky, Američtí zpěváci, Narození 1988, Americké hudební skupiny, Narození 1990, Brouci, ...

Result III

Query **matka, mother** **1,225** (65.20 per million)

Page 1 of 62 Go Next | Last

Czech Wikipedia Parallel Corpus	English Wikipedia Parallel Corpus
Kevin je kardiochirurg a její **matka** , Carolyn byla inženýrkou životního	cardiothoracic surgeon and her **mother** , Carolyn, was formerly an environmental
seriálech jako Jak jsem poznal vaši **matku** ,Lovci duchů a Greek. Také se	popular TV series How I Met Your **Mother** , Supernatural, and Greek. She
v seriálu Jak jsem poznal vaši **matku** (Epizoda: „Girls Versus Suits	Barney Stinson on How I Met Your **Mother** (Episode: "Girls Versus Suits
Burton jako Donna Ralston, Aronova **matka** Skutečný Aron Ralston se na chvíli	Burton as Donna Ralston, Ralston's **mother** . Aron Ralston himself makes a
Port St. Lucie na Floridě, ale **matka** ho vychovávala v městě Deerfield	May 11, 1988 and raised by his **mother** in Deerfield Beach, part of Broward
Conchords (HBO) Jak jsem poznal vaši **matku** (CBS) Kancl (NBC) Tráva (Showtime	Conchords (HBO) How I Met Your **Mother** (CBS) The Office (NBC) Weeds
v seriálu Jak jsem poznal vaši **matku** (Epizoda: „Benefits") (CBS) Jack	Barney Stinson on How I Met Your **Mother** (Episode: "Benefits") (CBS) Jack
Stinson, Jak jsem poznal vaši **matku** Rachel Berry a Santana Lopez,	Barney Stinson, How I Met Your **Mother** Rachel Berry & Santana Lopez,
Scherbatsky, Jak jsem poznal vaši **matku** Caroline Channing a Max Black	Robin Scherbatsky, How I Met Your **Mother** Caroline Channing & Max Black
Lopez (Naya Rivera) a Santaninu **matku** Maribel Lopez (Gloria Estefan	Lopez (Naya Rivera), and Santana's **mother** Maribel Lopez (Gloria Estefan
Underwood, Jak jsem poznal vaši **matku** Demi Lovato, Chirurgové Neil	Carrie Underwood, How I Met Your **Mother** Demi Lovato, Grey's Anatomy Neil
zajmout krále Karla IX. a jeho **matku** v Meaux, huguenoti obsadili několik	capture King Charles IX and his **mother** at Meaux. The Huguenots do capture
Katy Perry, Jak jsem poznal vaši **matku** Kristin Chenoweth, Glee Michael	Office Katy Perry, How I Met Your **Mother** Kristin Chenoweth, Glee Michael
Efendı. Adile Sultan ztratila **matku** , když byla ještě malá a vychovávala	Istanbul). Adile Sultan lost her **mother** at a very young age, and was
velšské básni stojí, že Áedánova **matka** byla dcerou krále Dumnaguala	Welsh poem states that Áedán's **mother** was a daughter of King Dumnagual
princ nižší královské třídy (jeho **matka** nebyla šlechtického rodu), si	a lesser-class prince (as his **mother** was a commoner), thought that
budoucí manželku prince Mahidola a **matku** dvou budoucích králů Thajska.	future wife of Prince Mahidol and **mother** of two future kings of Thailand
šlechtice Æ (comes Agelberhtus) a **matka** Edmunda, Æ, Eadwig and Eadgyth	Æthelberht (comes Agelberhtus) and the **mother** of Edmund, Æthelstan, Eadwig
centuriony a zavražděn v rukou své **matky** Iulie Domny. Vzájemné odtažité	Caracalla's army and slain in his **mother** Julia's arms. Julia is said thereafter
Guayaquilu, v Ekvádoru, zatímco její **matka** je Kanaďanka s německými, anglickými	Guayaquil, Ecuador, while her **mother** is a Canadian who has German,

Page 1 of 62 Go Next | Last

Further use of the corpus

- Master studies of Translation of English
- Research of the quality of Wikipedia translations[1]
- Output: Advice concerning what to do and what to avoid when translating articles in Wikipedia

[1] Adéla successfully defended her master thesis just a few days ago.