# Between Comparable and Parallel: English-Czech Corpus from Wikipedia

Adéla Štromajerová, Vít Baisa, Marek Blahuš

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xstromaj,xbaisa,xblah}@fi.muni.cz

**Abstract.** We describe the process of creating a parallel corpus from Czech and English Wikipedias using methods which are language independent. The corpus consists of Czech and English Wikipedia articles, the Czech ones being translations of the English ones, is aligned on sentence level and is accessible in Sketch Engine corpus manager.[1]

**Key words:** parallel corpora, comparable corpora, Wikipedia

## 1 Introduction

Wikipedia is now available in almost 300 languages, 13 of them having more than one million articles. The largest is the English Wikipedia, which contains more than 5 million articles. For each article, Wikipedia stores information about all the editing: the editor, the time of the editing and the changes made in the article. It is also possible to view any previous version of the article.

New articles are either written from scratch, or an article from another language version of Wikipedia is translated. However, translations do not need to cover the whole original article. Translations in Wikipedia are mostly from English to other languages. The *Translated page* template should be always added into such article so that it is clear that it is a translation and to identify what article and what language version of Wikipedia have been used for the translation.[2]

We used this information and extracted the translated articles as a basis for an English-Czech parallel corpus. There are more than 37,000 such articles[3] and they cover a wide range of topics. As the whole Czech Wikipedia contains about 350,000 articles, the proportion of the English-Czech translations is quite high, i.e. approximately 10%, when the number of articles (not their length) is considered.

[1] https://ske.fi.muni.cz

[2] https://cs.wikipedia.org/wiki/Wikipedie:WikiProjekt_P%C5%99eklad/Rady

[3] https://cs.wikipedia.org/wiki/Kategorie:Monitoring:%C4%8Cl%C3%A1nky_p%C5%99elo%C5%BEen%C3%A9_z_enwiki

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 3–8, 2016. 

## 2 Related work

There have been a few attempts at creating corpora from Wikipedia, e.g. [1]. There is a huge comparable corpus *Wikipedia Comparable Corpora*[4]. This consists of monolingual corpora, every one of them containing all articles from a particular language version of Wikipedia. These corpora are then document-aligned, i.e., the corpora consist of document pairs of articles on the same subject. There are also some parallel corpora based on Wikipedia, for example, a Chinese-Japanese parallel corpus created to help to improve the SMT between these two languages [2].

The field of web parallel corpora is of great interest nowadays, and there are many projects dealing specifically with Wikipedia (Besides the already mentioned ones, e.g. a Persian-English parallel corpus [3]. However, there is no parallel corpus made out of English and Czech articles from Wikipedia.

## 3 Exploiting Wikipedia Article Translations

The workflow was the following: a) to identify which Czech articles were created by translating English articles; b) to find out which version of the Czech article was the first, original translation; c) to identify the English articles from which the Czech ones were translated; d) to determine the version of the English article from which the Czech one was translated; and e) to download the texts of the Czech articles and the corresponding texts of the English articles.

First, it was necessary to determine which Czech articles were translated from English. This was quite simple as it is required to include the Translated page template in all the translated pages in Wikipedia. This template is supposed to be inserted into the References section and the structure of the Czech template, called Šablona:Překlad, is as follows: `{{Překlad|jazyk= |článek= |revize= }}` or, in short, `{{Překlad|en|article|123456}}`, where the second field denotes the language of the original article (represented by a language code, e.g., en for English), the third gives the name of the original article and the last field stands for the version identifier of the revision of the original article from which the Czech one was translated. The version identifier is a number and can be found at the total end of the permanent link to a given article/version, preceded by `oldid=`.

Second, it had to be identified which version of the Czech article was the first, original translation. If the current version of the Czech article was downloaded, the English and Czech texts could differ substantially as the articles change over time and some parts of the original translation would be edited or deleted, while some others would be added. Therefore, it would be then a more difficult task for a sentence aligner to extract parallel sentences from such different texts. Therefore, the revisions of Czech articles had to be searched and the revision where the full Translated page template appeared for the first time, had to be downloaded.

[4] `http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/`

For accessing its data, meta-data and features, Wikipedia provides users with various APIs connected to MediaWiki. MediaWiki[5] is a free software open source wiki package written in PHP, which was originally designed for Wikipedia, but nowadays it is used by many other wikis. Its most known API is the *MediaWiki action API*. This provides a direct access to the data in MediaWiki databases via a URL. Clients then request particular `action` parameter to get the desired information.[6] Using `action=query` module, it is possible to get meta information about the wiki, properties of pages, or lists of pages matching certain criteria[7]. IDs of all the articles translated from the English Wikipedia were retrieved with the help of this module.

We decided to retrieve Wikipedia pages in HTML format. The process of downloading a language pair of a Czech and an English article was the following. Taking the page ID from the ID list retrieved before, the revisions of the particular Czech article were accessed. They were listed in the order from the oldest to the newest. For every revision, its ID, timestamp and content were listed. The content of them was then searched for the *Translated page* template using regular expressions. The target of the search was only the full template containing not only the language and the name of the article, but also the revision ID. There were some articles which contained only an incomplete template without the revision ID. These articles were not downloaded as it was not possible to determine exactly from which version the Czech translation was made[8]. The ID of the identified revision was then taken and this particular version of the Czech article was retrieved.

We then used *jusText*[9] [4]. It removes all the boilerplate content, such as navigation links, headers and footers, and it preserves the text in the form of a list of paragraphs.

Another step was to remove the final sections of Wikipedia articles which were needless in the text, i.e., References, External links, etc. Finally, other unnecessary parts were removed from the page, e.g., reference numbers and note numbers. After this, the title and then the rest of the Czech text was written into a document with a name consisting of a number followed by the code for the Czech language, i.e., "cs".

It has to be noted that we could work with MediaWiki format of articles, but there is no suitable method of converting MediaWiki format into HTML reliably in a large scale.

The English pages were then processed in a similar way.

---

[5] `https://www.mediawiki.org/wiki/MediaWiki`

[6] `https://www.mediawiki.org/wiki/API:Main_page`

[7] `https://www.mediawiki.org/wiki/API:Query`

[8] The amount of such articles was, however, very low, and together with other download errors caused by inconsistencies in metadata, etc., it was lower than 1% of the total sum of 37,00 articles for both Czech and English.

[9] `http://corpus.tools/wiki/Justext`

## 4 Parallel Sentences Extraction

The retrieved texts cannot be considered parallel. Therefore, they had to be processed before including them in the corpus. With the help of a sentence aligner, sentences in the corresponding texts were aligned. These aligned sentences can then be considered parallel and can be used as data for a parallel corpus.

There are many sentence aligners available. We used *hunalign* [5] because it is easy to use, it is quick, it is a free software, and it supports not only one-to-one alignments, but also m:n mapping.

Sentence aligners use mostly three types of alignment methods: length-based, dictionary or translation-based and partial similarity-based. Hunalign is a hybrid sentence aligner whose alignment method is length-based and dictionary-based. These methods are simpler than, e.g., the translation-based method. Sentence aligners using the translation-based method, such as Bleualign [6], first make a machine translation of the source text and then compare this translation to the target text. Because of this machine translation included, translation-based alignment methods are quite complex and time-consuming. There are also sentence aligners that first need big corpora to be trained on (e.g., Gargantua [7]), which, again, makes their use more demanding both of resources and time.

Hunalign combines the length-based and the dictionary-based methods and aligns the sentence segments according to scores from both methods together. If there is no dictionary provided, it first aligns according the sentence length, and then, based on this alignment, makes an automatic dictionary and realigns the texts according to this dictionary.

As hunalign does not come with an English-Czech dictionary, the next step was to get such a dictionary in the format suitable for hunalign. This dictionary was created from monolingual dictionaries provided with *LF Aligner*. LF Aligner is a hunalign wrapper written by Andras Farkas[10]. It comes with built-in monolingual dictionaries which are then during the process of aligning combined into bilingual dictionaries according to the languages used. The English and the Czech dictionary were taken and they were combined into an English-Czech dictionary according to the structure of hunalign dictionaries, which is `target_language_phrase @ source_language_phrase` per line. The resulting dictionary was then provided to hunalign to achieve a better alignment.

During the alignment process, it was found out that hunalign is not able to produce alignment of files which differ in sizes considerably. In this case, the alignment is not produced. However, such files are unalignable in general, regardless of the used sentence aligner. The amount of files not aligned was quite big, i.e., about 20%. However, the remaining data were still enough for building a corpus of a reasonable size.

The examples of extracted parallel sentences can be found in Table 1.

[10] `https://sourceforge.net/projects/aligner/`

Table 1: Examples of aligned sentences.

| | |
|---|---|
| In April 1944 the Squadron moved back to the UK and re-assembled at North Weald on 23 April. | V dubnu 1944 byla peruť přeložena do Spojeného království a 23. dubna reaktivována na základně RAF North Weald. |
| ABAKO and Kasavubu spearheaded ethnic nationalism there and in 1956 issued a manifesto calling for immediate independence. | ABAKO a Kasavubu zde razili cestu etnickému nacionalismu a v roce 1956 vydali prohlášení volající po okamžité nezávislosti. |
| The band's music has a combination of influences: reggae, Latin, rock and hip hop, which is performed in a minimalistic folk style limited to vocals, beatboxing, and acoustic guitar. | Hudba 5'nizze je kombinací vlivů reggae, latinskoamerické hudby, rocku a Hip hopu v minimalistickém folkovém stylu omezeném na vokály, beatboxování a akustickou kytaru. |
| The Book of Abramelin tells the story of an Egyptianmage named Abramelin, or Abra-Melin, who taught a system of magic to Abraham of Worms, a GermanJew presumed to have lived from c.1362 - c.1458. | Abramelinova kniha vypráví příběh egyptského mága jménem Abramelin, nebo Abra-Melin, který předal svou nauku o magii Abrahamovi z Wormsu, německému Židu, o kterém se předpokládá, že žil v letech 1362-1458. |
| Her father, Kevin, is a cardiothoracic surgeon and her mother, Carolyn, was formerly an environmental engineer before becoming a homemaker. | Její otec, Kevin je kardiochirurg a její matka, Carolyn byla inženýrkou životního prostředí, než začala být ženou v domácnosti. |
| Accola appeared in her first movie, Pirate Camp, in 2007. | Její filmový debut přišel v roce 2007 ve filmu Pirate Camp. |
| Although the comet was next expected at perihelion on 1997 April, no observations were reported. | Ačkoli byla kometa znovu očekávána v periheliu roku 1997, nebyly hlášeny žádná pozorování. |

After the alignment, we processed the data into vertical format required by corpus indexing system manatee and corpus manager Sketch Engine [8]: namely we a) added metadata to each document, segmented paragraphs, tokenized texts, tagged them for part of speech, prepared configuration files, prepared mapping files from the output of hunalign and compiled them.

The tools used in this chapter are available through the Natural Language Processing Centre at the Faculty of Informatics, Masaryk University[11] and at `http://corpus.tools`. E.g. unitok [9] was used for tokenization of plain texts.

Hunaling provides a score for each pair of aligned sentences. We decided to keep only the alignments with the score higher than 0.5.[12].

[11] `https://nlp.fi.muni.cz`

[12] We did not find the range of hunalign scores, the threshold was chosen heuristically.

## 5 Conclusion

The English corpus contains 46,238,455 tokens and the Czech corpus contains 18,785,688 tokens. The size of the aligned content is then 7,275,092 words in the English corpus and 6,414,841 words in the Czech one.

The corpus was made public and it is available at the Sketch Engine site of the Faculty of Informatics. The individual Czech and English corpora can be found under the names "Czech Wikipedia Parallel Corpus" and "English Wikipedia Parallel Corpus". The corpora are published under the CC BY-SA 4.0 license[13].

The corpus is accessible for all students and members of staff of Masaryk University. It can be used both in the field of NLP and in the field of linguistics, providing information about the language to teachers, lexicographers and translators.

**Acknowledgments.** This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2016* (MUNI/A/0863/2015) and by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

## References

1. Davies, M.: The Wikipedia corpus: 4.6 million articles, 1.9 billion words. Adapted from Wikipedia. Accessed February **15** (2015)
2. Chu, C., Nakazawa, T., Kurohashi, S.: Constructing a Chinese-Japanese Parallel Corpus from Wikipedia. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (may 2014)
3. Mohammadi, M., GhasemAghaee, N.: Building bilingual parallel corpora based on wikipedia. In: Computer Engineering and Applications (ICCEA), 2010 Second International Conference on. Volume 2., IEEE (2010) 264–268
4. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic (2011)
5. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005. (2005) 590–596
6. Sennrich, R., Volk, M.: Mt-based sentence alignment for ocr-generated parallel texts. In: The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado. (2010)
7. Braune, F., Fraser, A.: Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics (2010) 81–89
8. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1**(1) (2014) 7–36
9. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In Horák, A., Rychlý, P., eds.: RASLAN 2014, Brno, Czech Republic, Tribun EU (2014) 71–75

[13] `https://creativecommons.org/licenses/by-sa/4.0/`