TransLiTex: A Parallel Corpus of Translated Literary Texts

Amel Fraisse[1], Quoc-Tan Tran[1], Ronald Jenn[1], Patrick Paroubek[2], Shelley Fisher Fishkin[3]
[1]University of Lille (France), [2]LIMSI-CNRS (France), [3]Stanford University (USA)
{amel.fraisse, ronald.jenn}@univ-lille3.fr, quoc-tan.tran@etu.univ-lille3.fr, pap@limsi.fr, sfishkin@stanford.edu

Abstract

In this paper, we present our ongoing research work to create a massively parallel corpus of translated literary texts which is useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. Using a crowdsourcing approach, we identified and collected 29 translations of Mark Twain's *Adventures of Huckleberry Finn* published in 23 languages including less-resourced languages. We report on the current status of the corpus, with 5 chapter-aligned translations (English-Dutch, two English-Hungarian, English-Polish and English-Russian). We evaluated the correctness of chapter alignment by computing the percentage of common words between the English version and the translated ones. Results show high percentages that vary between 43% and 64% proving the high correctness of chapter alignment.

Keywords: parallel corpus, comparable corpus, translated literary texts

1. Introduction

Parallel corpora are a valuable resource for linguistic research and natural language processing (NLP) applications. Such corpora are often used for testing new tools and methods in Statistical Machine Translation (SMT), where large amounts of aligned data are often used to learn word alignment models between two languages (Och and Ney, 2003). The most widely used parallel corpora in computational approaches are the Canadian Hansards (Roukos et al., 1995) which are bilingual (English and French), the United Nations Parallel Corpus (6 languages) (Ziemski et al., 2016), or the European Parliament proceedings (21 languages) (Koehn, 2005). These resources belong to the legal and political sphere.

Another source of parallel corpora that has recently attracted attention is religious texts such as the Bible. This line of research, which entailed the compilation of many parallel texts, has broken new ground and allowed computational linguistics to handle vast corpora. Cysouw and Walchli (2007) introduced the notion of 'massively parallel corpora' for texts that have translations into a great number of languages (100+). Although there are not many such texts, those that are available offer an incredibly rich source for computational linguistic researchers. That is why our project taps into relatively unexplored sources for massively parallel corpora: translated literary texts.

A growing number of those texts are now available in electronic form on the internet and they are indexed by public online catalogues such as Wikisource[1], Archive.org[2], Project Gutenberg[3], etc. In this paper, we will report on our ongoing research work to compile such a massively parallel literary corpus. This paper is structured as follows. The next section gives an overview on related work on the construction of parallel corpora. Section 3 describes the Mark Twain translation corpus. Section 4 presents our method for data collection. Section 5 outlines the corpus alignment. Section 6 describes the corpus evaluation and discusses the results. Section 7 mentions the expected use of the corpus and in the last section, we conclude and underline future work.

2. Related work

Computational linguistics researchers have been exploring different sources for building parallel and comparable corpora. Resnik and Smith (2003) used the Web as parallel text to construct a significant parallel corpus for a low-density language pair.

In accordance with the fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for construction of parallel corpora (Tomas et al., 2008; Tufiș et al., 2013; Labaka et al., 2016). Other research works used Twitter as comparable corpus to build multilingual linguistic resources (Fraisse and Paroubek, 2014; Vicente et al., 2016).

There have also been research works which show the potential of the Bible as a source to compile massively parallel corpora (Resnik et al., 1999). Mayer and Cysouw (2014), based on freely available resources, created a Bible corpus with over 900 translations in more than 830 language varieties. Christodouloupoulos and Steedman (2015) built a massively parallel corpus based on 100 translations of the Bible, emphasizing difficulties in acquiring and processing the raw material.

There are also parallel corpora related to translated literary works (e.g. Harry Potter, Le Petit Prince, Master i Margarita) or translations from the web, mostly available for a set of closely related languages (Mayer and Cysouw, 2014; Cysouw and Walchli, 2007). Most of these texts, however, cannot be regarded as massively parallel texts, they are not freely available online, and they mainly concern well-endowed largely known languages.

3. Mark Twain corpus

Mark Twain's books are some of the most well-travelled texts on the planet. As the UNESCO Index Translationum shows the American writer is ranked 15 in the top-50 of the most translated authors worldwide. His works have been

[1] https://wikisource.org
[2] https://archive.org
[3] https://www.gutenberg.org

translated into almost every language in which books are printed (Rodney, 1982). The novel *Adventures of Huckleberry Finn* (Twain, 1885) is one of the most commonly translated of his books. Rodney (1982) identified 375 translations as of 1976. As UNESCO's Index Translationum[4] suggests, hundreds of additional translations have been published in the four decades since Rodney completed his survey. But these two sources are both significantly out of date and incomplete. (For example, UNESCO's Index Translationum lists 15 translations of the novel in Chinese, but Lai-Henderson (2015) documented 90 Chinese translations). The scores of language into which the book has been translated include Afrikaans, Albanian, Arabic, Assamese, Bengali, Bulgarian, Burmese, Catalan, Chinese, Chuvash, Czech, Danish, Dutch, Estonian, Farsi, Finnish, French, German, Georgian, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Kirghiz, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Oriya, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Sinhalese, Slovak, Slovenian, Spanish, Swedish, Tamil, Tatar, Telugu, Thai, Turkish, Ukranian, and Uzbek. In many of these languages, there have been multiple translations over time, reflecting different moments in history, and different ideological perspectives on the part of the translators or publishers, as well as different attitudes towards the US, towards childhood, towards minorities and minority dialects, towards race and racism, etc. Of all the existing translations of Mark Twain's works, *Adventures of Huckelberry Finn* stands out because of its rich intercultural content and the great number of translations. That's why we decided to focus on this particular novel for our project.

4. Data collection

To collect our raw data, we proceeded in two steps. First, we built a seed corpus by crawling existing databases and digital archives such as Gutenberg Project , Unesco's Index Translationum, Wikisource, etc. For this seed corpus we collected the original English text of *Adventures of Huckelberry Finn* as well as the French, German, Polish, Russian, Dutch and Hungarian translations. Then, as the example of Chinese translations demonstrates (Lai-Henderson, 2015), the existing sources and databases show cracks that only the power of the crowd can help us fill. That is why we use a crowdsourcing-based approach to discover and collect translations from other languages that hadn't been indexed in those above-mentioned databases.

4.1. Crowdsourcing experiment

Due to the significant amount of existing translations and the growing number of digital versions made available online, the crowdsourcing allowed us to gather data that would have otherwise been beyond our reach. Crowdsourcing helped reduce the amount of time spent on the task, increase the variety and the range of the data covered (such as identifying translations which are not indexed in public databases). We used the CrowdFlower[5] platform. The parametrization of the experiment was as follows: as we are looking for translations over the world, we have not limited the geographic location of the contributors. Each task consisted in a set of 9 questions (i.e. units in the CrowdFlower terminology) and completing the task will earn 0.25 $ (instead of 0.15 $ recommended by CrowdFlower). In fact, the task that the workers had to go through to complete the job was complex. First we asked people to use search engines or online catalogs to look for existing translations in their native language. Then we asked them if they could find the translator's name, the first year of publication, the publishing house, the URL of the cover, the bibliographic record, and available public digital versions.

Because of the complexity of the task, the crowdsourcing approach did not look like the best option. We assumed that the cultural background of crowdsourcing workers would not allow them to complete the task efficiently but it turned out that they managed to provide us with valuable and reliable information. One week after launching the job on CrowdFlower, we received 710 judgements covering 31 different languages. On top came Spanish (163 responses), Arabic (76), Malay and Indonesian (47), German (46), French (45), Greek (43) and Turkish (39). After data cleaning we collected 29 translations in 23 languages of different formats (html, text, pdf, epub).

4.2. Full-text acquisition

Before collecting the full-text, we checked the reliability of the collected translations. First, we use a Python script and Google Translate to verify if the translated titles are equivalent to the original one. Among 710 given titles, we eliminated 25 incorrect responses (such as "The Adventures of Tom Sawyer").

Secondly, we verified the copyright by checking the full-text URLs in order to know whether they came from a national or public institution that has the right to distribute the digital versions. Based on information gathered by the crowd, we crawled further into national archives and digital libraries to get the full-text versions such as Wikisource[6], DBNL[7] (Digital Library for Dutch Literature), Archive.org, Lib.ru[8] (also known as Maksim Moshkow's Library and Russia's Project Gutenberg), MEK[9] (Hungarian Electronic Library), etc.

5. Corpus alignment

The original version of the novel as well as most of the collected translations are already structured by chapter and by paragraph. We kept the original structure for the alignment process. Each translation contains chapter (<CHAPTER>), and paragraph (<P>) mark-ups on separate lines (Figure 1). In this work, we performed an alignment at chapter level by using the mark-up <CHAPTER> as a marker to extract and align chapters of translations that have the same number of chapters as the English source version (43 chapters). In total, we aligned 5 translations (English-Dutch, two English-Hungarian, English-Polish and English-Russian). Transla-

[4] http://www.unesco.org/xtrans/

[5] https://www.crowdflower.com/

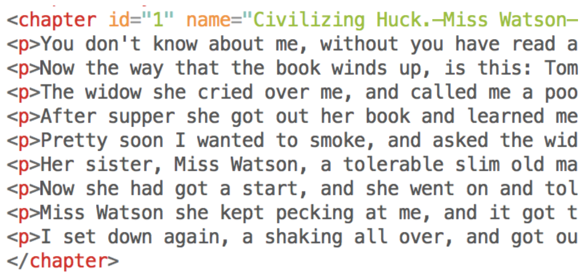
[6] https://fr.wikisource.org/

[7] http://www.dbnl.org

[8] http://az.lib.ru/

[9] http://mek.oszk.hu

tions with different numbers of chapters, will be aligned and included in a further version of the corpus. Table 1 describes the number of paragraphs and sentences for each aligned translation.



```
<chapter id="1" name="Civilizing Huck.—Miss Watson—
<p>You don't know about me, without you have read a
<p>Now the way that the book winds up, is this: Tom
<p>The widow she cried over me, and called me a poo
<p>After supper she got out her book and learned me
<p>Pretty soon I wanted to smoke, and asked the wid
<p>Her sister, Miss Watson, a tolerable slim old ma
<p>Now she had got a start, and she went on and tol
<p>Miss Watson she kept pecking at me, and it got t
<p>I set down again, a shaking all over, and got ou
</chapter>
```

Figure 1: Format of the released corpus. Extract from the chapter 1 of the English version.

6. Corpus evaluation and results

In order to evaluate the quality of our parallel corpus, we wanted to determine what degree of similarity between the original text and the translations was. As the alignment unit used for this version of the corpus was chapter–not paragraph or sentence–we evaluated the semantic similarity between two chapters as a whole. The goal of the evaluation is to find out how similar the parallel chapters are. We consider two aligned chapters as similar if they contain a significant percentage of words with the same semantic meaning. Firstly, we identified for each text the direction of translation–that is to say, whether they were directly translated from English or whether they went through another target language first. In fact, the source language has an important influence on the nature of its translation. A manual survey of the five translated versions studied in this work confirmed that they have English as their source language. (English-Polish, English-Russian, two English-Hungarian, English-Dutch). Secondly, we used Google Translate to acquire the English literal translation of each collected target text and compared it to the original. The comparison consists in computing the percentage of common words between each chapter of the literal and the original version (the stop-words are excluded). We used the Stanford tokenizer (Manning et al., 2014) to extract tokens from both texts.

The Figure 2 shows that the percentage of common words ranges between 43% and 64% according to chapters and target languages. For the Polish translation, the lowest score is 43% in chapters 13 and 43, and the highest score is 56% in chapter 1. For Russian, the lowest score is 46% in chapter 43 and the highest score is 60% in chapter 11. In the first Hungarian translation the lowest score is 49% in chapter 21 and the highest score is 64% in chapter 11. In the second Hungarian translation the lowest score is 46% in chapter 43 and the highest score is 60% in chapters 11 and 31. In the Dutch translation the lowest score is 51% in chapters 23, 29 and 43 and the highest score is 61% in chapter 11 .

Although these scores consider only literal and strict translation as common words, they show that the collected translations are similar and staying fairly to the original and could be considered as parallel in the strict sense of the term.

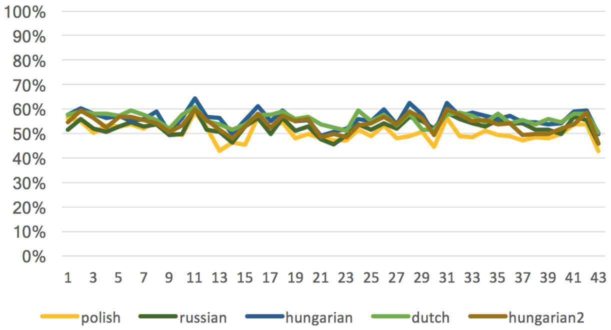


Figure 2: Percentage of common words with the English version by language and chapter.

7. Expected use and availability

One major achievement will be to provide statistical machine translation systems with a rich parallel corpus. This current version of our corpus displays 5 languages (English, Dutch, Hungarian, Polish and Russian) and other languages are being processed so that the corpus will grow over time. One of our ultimate goals is to reach out to the less-resourced languages such as Finnish, Latvian, Malay, Turkish, etc.

Another goal is to engage scholars in the field of digital humanities as well as languages and Translation Studies specialists to address a number of fundamental questions. What happen in translations? What is the impact of the linguistic and cultural transfer of the novel on its textual and iconic nature? An aligned digital corpus would allow them to evaluate the modifications and adaptations set up by translators and the translation process. It will make available to them a stable and reliable corpus to conduct their own research. This research work will raise awareness of corpora and how they can benefit academics both in their research and their teaching in various humanities areas.

The corpus is available online and accessible on Github at the URL: `https://github.com/amelfraisse/TransLiTex/releases/tag/v1.1`.

8. Conclusion and future works

In this work, we provided a parallel corpus of translated literary texts of Mark Twain's *Adventures of Huckleberry Finn*. The aim is to support interdisciplinary research that benefits from the convergence of knowledge in computational linguistics and Translation Studies. On the one hand, it explores new directions in which parallel corpora of literary texts can help produce statistically reliable results. On the other hand, it provides digital humanities scholars with materials for extraction and acquisition of new knowledge. For raw data collection, we resorted to crowdsourcing to discover translations that had not been indexed in public databases such as the UNESCO's Index Translationum and particularly when they were published in less-resourced

Version	Num. of Chapters	Num. of Paragraphs	Num. of Sentences
English	43	2155	6190
Dutch	43	2150	6134
Russian	43	2214	7486
Polish	43	2293	8339
Hungarian 1	43	2237	6503
Hungarian 2	43	2162	6608

Table 1: Characteristics of the realized parallel corpus: numbers of chapters, paragraphs and sentences in different translations.

languages. After verifying the copyright of full-text translations, we aligned them by chapter. We report on the current status of the corpus, with 5 aligned translations in 5 languages (English-Dutch, two English-Hungarian, English-Polish and English-Russian). We evaluated the semantic similarity between aligned chapters by computing the percentage of common words between each chapter of the literal translated and the original version. Results show that the percentage of common words ranges between 43% and 64% proving the high correctness of chapter alignment between the 5 translations. In a future work, we plan to perform the alignment at the paragraph and the sentence level and extend this version to other languages.

Acknowledgments

This research work is conducted within the framework of the "Global Huck" project funded by the MESHS (Maison Européenne des Sciences de l'Homme et de la Société) in Lille, France. It is a partnership with Stanford University. We are also grateful to the Center for Mark Twain Studies in Elmira, N.Y.

9. Bibliographical references

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Cysouw, M. and Walchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.

Fraisse, A. and Paroubek, P. (2014). Twitter as a comparable corpus to build multilingual affective lexicons. In *Proceedings of the 7th International Workshop on Building and Using Comparable Corpora at LREC 2014*, pages 17–21.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit 2005*, page 79–86, Phuket, Thailand.

Labaka, G., Alegria, I., and Sarasola, K. (2016). Domain adaptation in mt using titles in wikipedia as a parallel corpus: Resources and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lai-Henderson, S. (2015). *Mark Twain in China*. Stanford University Press, Stanford, CA.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Resnik, P., Olsen, M. B., and Mona, D. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Rodney, R. M. (1982). *Mark Twain International: A Bibliography and Interpretation of his Wordwide Popularity*. Greenwood Press, Westport, CT.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard french/english. In *Philadelphia: Linguistic Data Consortium*.

Tomas, J., Bataller, J., and Casacuberta, F. (2008). Mining wikipedia as a parallel and comparable corpus. *Language Forum*, 34(1).

Tufiș, D., Ion, R., Ștefan Daniel, and Ștefănescu, D. (2013). Wikipedia as an smt training corpus. In *Proceedings of the 9th conference RANLP*.

Twain, M. (1885). *Adventures of Huckelberry Finn*. Charles L. Webster and Company, San Mateo, CA.

Vicente, I. S., Alegria, I., Espana-Bonet, C., Gamallo, P., Oliveira, H. G., Garcia, E. M., Toral, A., Zubiaga, A., and Aranberri, N. (2016). Tweetmt: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 3530–3534, Portorož, Slovenia.