Computer Science Journal of Moldova, vol.**28**, no.3(84), 2020

On Alignment of Textual Elements in a Parallel Diachronic Corpus

Tudor Bumbu

Abstract

This paper presents the description of some tools and resources for aligning diachronic parallel texts. We try to emphasize the idea of improving the lexicographical similarity between words and expressions used in old and modern texts, using a special technique of natural language processing namely the BLEU score. The overall result of the research is a package of language tools and resources, which will serve to the automatic alignment of the old Romanian text with the modern text.

Keywords: text alignment, parallel diachronic corpus, word alignment tools, BLEU score

1 Introduction

The aligning of an old text to its modern representation means translating it into a contemporary language by replacing outdated lexical variants with modern expressions.

Parallel texts are valuable linguistic resources in many fields of research, but also in practical applications. The best known application is statistical machine translation, or more recently, neural machine translation (using neural networks). Also, parallel texts are researched in disambiguating the meaning of the word, recognizing proper names, learning the language model, but also in the diachronic analysis of natural languages.

A set of parallel texts can form a parallel corpus. The central part in working with a parallel corpus is the alignment task. Alignment in

©2020 by CSJM; T. Bumbu

this sense is the process of linking the corresponding textual parts of parallel texts. An important feature of parallel texts is the property of having in itself a correspondence between two or more texts, for example, the equivalence of translation or paraphrasing. We assume that they connect with each other via their meaning.

In computational linguistics, the term parallel text, also refers to pairs of collections of texts in the same field that include translations of one and the same document. But in most cases, parallel texts refer to bilingual corpora [10].

A diachronic parallel corpus refers to a parallel corpus, where the parallel texts share the same language but their time periods are different.

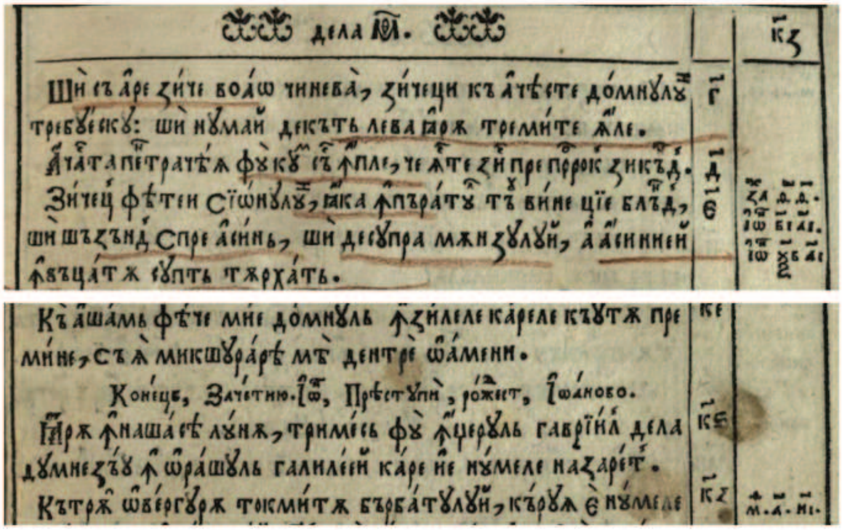
The purpose of our research is to translate historical texts into texts that use expressions and words from the modern dictionary of the Romanian language. Thus, the first beginning of the work is the elaboration of a parallel diachronic corpus with text in Romanian written hundreds of years ago, aligned with its modern variant. We took as primary resources text from the New Testament of the 17th century and New Testament of the 20th century, but an extension of the corpus follows with other texts from our cultural thesaurus.

2 Textual resources

Our main resource is the parallel diachronic corpus developed in Bumbu, 2019 [1], placed on the web with open access. One might say that this resource cannot be called a parallel corpus but only a parallel text. We are working to expand this resource by digitizing and iteratively adding new texts, so that in the end we get a large collection of parallel texts and we prefer to call it a parallel corpus.

The parallel corpus contains text from the book „Noulu Testament sau Înpacarea, au Leagea noao a lui Is. Hs.”, printed in 1648 at the Belgrade Fortress, Transylvania [2] and text from the electronic version of the New Testament (written and annotated by Bartolomeu Valeriu Anania, archbishop of the Archdiocese of Vadul, Feleac and Cluj in 1990) [3]. The volume of text in the corpus expressed in sentences is currently approx. 10,000 sentences.

Given the fact that the alphabet of the book printed in the seventeenth century is old and unused (the Romanian Cyrillic alphabet), it takes several steps to reach the editable text with the modern alphabet of the Romanian language. The first step is the optical character recognition (OCR). The OCR applied to the old book has an accuracy of up to 80%, and errors are corrected manually. Some issues that worsen the accuracy of OCR applied to the old book are: the pages in the book are worn (they have brownish spots and underlines below text lines); the text in the book has a lot of features specific to that period, such as: letters over other letters, words written together, references written in Slavonic with a smaller font, abbreviations, widespread use of accents, etc (Figure 1).



дела ІѠ. КЗ

Шѝ съ а҆ре зи́че воа́ѡ чинева̀, зи́чеци къ а҆чѣсте до́мнꙋлꙋ Г
требꙋеско: шѝ нꙋма́й де́къть ле ва трѣж тремите а҆ле.
Ача́та а тоѿ фост, ка съ сѧ пли́не, че а҆ зи прѣ пророк зикъ́нд. Д
Зичец фѣтеи Сїѡ́нꙋлꙋ, ѥ҆ка а҆мпъра́тꙋ тъ вине цїе блъ́нд, Е
шѝ шъзъ́нд спре а҆сѣнь, шѝ деасꙋпра мѫнзꙋлꙋй, а а҆сѣниней
а҆вꙋца́тж сꙋпть тжрха́ть.

Кꙋ а҆ша́мь фѣ́че мїе до́мнꙋль а҆зилеле ка́реле кꙋтж пре КЕ
мі́не, съ а҆ микшꙋра́рѣ мѣ̀ дентрѐ ѡа́мени.
Коне́цꙋ, Зачѣтию. Ꙇ҆ѡ, Прѣстꙋпѝ, рожест, Ꙇ҆ѡа́ново.
Ꙗ҆рж а҆ наша́се лꙋнж, тримі́сь фꙋ̀ а҆нџерꙋль Гаврїи́л дела КЅ
дꙋмнезъꙋ а҆ ѡра́шꙋль Галиле́и ка́ре не нꙋмеле Наза́рет.
Кътрж ѡвергꙋрж токми́тж бърба́тꙋлꙋй, кърꙋѧ̀ е нꙋмеле КЗ

Figure 1. Two passages from the New Testament book of 1648

The next step after OCR is to transliterate the text from the Cyrillic alphabet into the Latin alphabet. At this stage we use a transliteration tool developed at the Institute of Mathematics and Computer Science “V. Andrunachievici” [5]. An example of transliterated text is shown in Table 1.

Table 1. Fragment of transliterated text of the New Testament of 1648

OCRed text	Text after transliteration
Ну џюдекарець касъ ну фиць џюдекац	Nu giudecareț casă nu fiț giudecaț
Къ ку че џюдекатъ веци џюдека, џюдекавъвець	Că cu ce giudecată veți giudeca, giudecavăveț

After transliterating the text we proceed to verse alignment. The verses from the Old New Testament (NTV) were aligned with the verses from the Modern New Testament (NTM) (Table 2) using the Levenshtein distance. The verses are placed in the corpus according to their order in the book. Respectively, 1006 verses from the Gospel according to Matthew, 675 verses from the Gospel according to Mark, 1151 and 668 verses from the Gospel according to Luke and John.

Table 2. Aligned verses from NTV and NTM

NTV verses	NTM verses
Și fu îzilele acelia, veni Iisus de în nazareful galileei, şi să boteză dela Ioan în Iordan.	Și în zilele acelea, Iisus a venit din Nazaretul Galileii și S'a botezat în Iordan de către Ioan.
Și aciiş eşind de în apă, văzu deşchise ceriurele, şi duhul ca un porumbu, pogorănd spre El.	Și îndata, ieșind din apă, a vazut cerurile deschise și Duhul ca un porumbel pogorându-Se peste El.

In the next section some word alignment tools are briefly described, including a tool we are currently working on.

3 Alignment Tools

To reach the automatic alignment of old texts to the modern texts, there are still an indefinite number of steps to be done. An intermediate step that appears is the alignment of old words to modern ones. The text elements here are words. Word alignment tools are usually special software that helps an expert to map source text words to target text words. One of these tools is the Berkeley Word Aligner (BWA) [7], a program written in Java, used in many cases for machine translation.

BWA aligns the words in a parallel corpus aligned at the sentence level, using a Hidden Markov Model (HMM) [8]. To train an HMM syntactic alignment model, a third parallel text is also needed for each pair of aligned text. All texts should be annotated with syntactic information. Considering the fact that we do not have any third parallel text, nor the annotation of the text with the syntactic information, this tool is useless to us at this stage.

Another tool pack for word alignment is GIZA ++ [9]. This tool also makes extensive use of hidden Markovian models to align texts. Giza ++ works directly with the aligned sentences from two parallel texts, without asking for any additional linguistic information attached to the text. Using an expectation maximization algorithm, the results of the alignment of the final words can be obtained after the software trains itself with a parallel corpus, several iterations, from source text to target text and vice-versa.

Given the fact that we are dealing with parallel diachronic texts, we decided to create an alignment tool that will satisfy special needs. Some examples of necessity will be: the calculation of the BLEU score between texts, sentences, expressions; interactive viewing of n-grams; viewing the word/expression coverage tree; working with more than two parallel texts at the same time etc.

Therefore, a parallel word alignment editor, which is still under construction, helps us to prepare the results of this paper.

The parallel word alignment editor is a WEB application, designed within Django framework, the Python programming language. The application consists of 3 general modules: the parallel text editing and parallel corpus formation module, the word processing module, and the word weighted graph creation module.

The text editing module is the gateway in the alignment application. It offers the possibility to view and edit several parallel texts simultaneously, and the changes are automatically propagated in the corpus. At the same time, in this module we can add texts directly to the corpus.

Some functionalities already defined in the word processing module are: computing of 1-, 2-, 3-, 4-grams, counting and visualization of n-grams [Figure 2], division into tokens (words and punctuation), associating tokens with numeric identifiers (each word is assigned an ID and the ID is an integer), and calculating the BLEU score between sentences or texts.

The text graph creation module involves mapping the word graphs in the parallel text. This module does not have fully defined functionalities, but we focus on a powerful visualization and interactivity device.

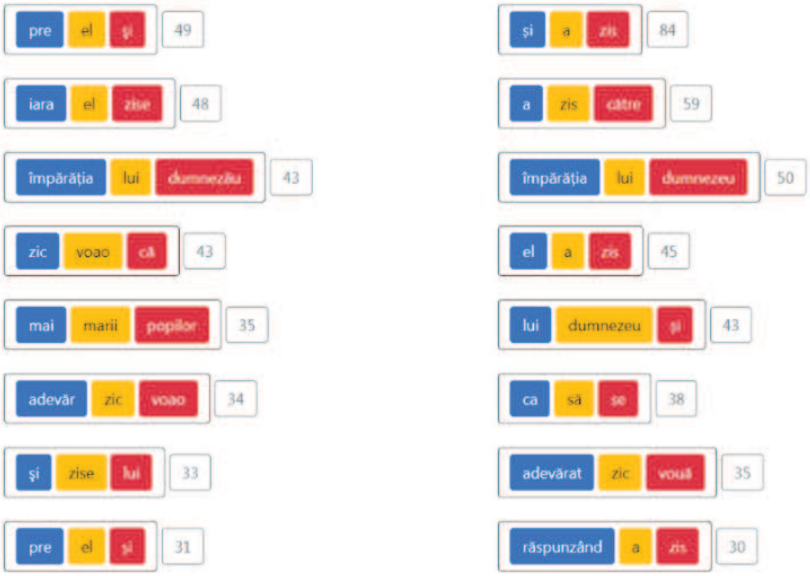


Figure 2. Visualization of 3-grams of 2 parallel texts, ordered descending by frequency

One of the basic functionalities we are going to use to improve the similarity between parallel texts is the BLEU score (Bilingual Evaluation Understanding), a metric used to evaluate a sentence generated in machine translation compared to a reference sentence (translated by an expert). The score was invented by Kishore Papineni et al., 2002 [6] to evaluate the predictions made by machine translation systems. This method has many advantages: it is fast and easy to calculate; easy to understand; language independent; it correlates very much with human evaluation of the translation and it can be used in diachronic analysis. The approach works by counting n-grams in the translated text and the reference text that match character by character, where 1-gram or uni-gram would be a word and bi-gram would be each pair of words. The comparison is made regardless of the order of the words and can reach up to 4 grams.

We assume that our translation is the source text or NTV in our corpus and the reference text is the target text or NTM in our corpus.

4 Conclusions

This paper presents 3 tools for aligning words in a parallel text. One of these tools is developed by us to frame special functionalities needed to align words in a parallel diachronic corpus. A specific functionality that is integrated by default in our application is the calculation of the BLEU score between sentences. We have included it as an approach to improve the similarity between the texts in a parallel corpus.

Acknowledgement. This work was supported by research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

[1] T. Bumbu, “Building a Diachronic Parallel Corpus for the Alignment of the Old Romanian Texts,” in *Proceedings of the Conference on Mathematical Foundations of Informatics MFOI-2019, July 3-6, 2019, Iasi, Romania*, pp. 263–269.

[2] *Noulu Testamentu sau Înpacarea, au Leagea noao a lui Is. Hs. Domnului nostru*, CR XVII V 13, Gheorghie Rakoti, Alba-Iulia, Romania, 1648, 676 p.

[3] *Biblia. Noul Testament*, Bartolomeu Valeriu Anania, Cluj-Napoca, Romania, 2001.

[4] A. Colesnicov, L. Malahov, and T. Bumbu, "Digitization of Romanian Printed Texts of the 17th Century," in *Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language. Alexandru Ioan Cuza University Press, 2016*, pp. 1–11.

[5] S. Cojocaru, A. Colesnicov, L. Malahov, T. Bumbu, and S. Ungur, "On Digitization of Romanian Cyrillic Printings of the 17th-18th Centuries," *Computer Science Journal of Moldova*, vol. 25, no. 2(74), pp. 217–225, 2017.

[6] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318.

[7] P. Liang, B. Taskar, and D. Klein, "Alignment by Agreement," in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006*, pp. 104–111.

[8] R. L. Stratonovich, "Conditional Markov Processes," *Theory of Probability and Its Applications*, vol. 5, no. 2, pp. 156–178, 1960.

[9] Franz J. Och, "Giza++: Training of statistical translation models," 1999-2000. Available: http://www.fjoch.com/GIZA++.html.

[10] R. Rapp, "Identifying word translations in non-parallel texts," in *Proceedings of the 33rd Meeting of the Association for Computational Linguistics, Cambridge, MA, 1995*, pp. 320–322.

Tudor Bumbu

Received October 28, 2020

State University "Dimitrie Cantemir"
E–mail: bumbutudor10@gmail.com

Vladimir Andrunachievici Institute of Mathematics and Computer Science
E–mail: tudor.bumbu@math.md

Technical University of Moldova
E–mail: tudor.bumbu@iis.utm.md