Multilingual aligned corpus with Ukrainian as the target language

Natalia Grabar (1), Olga Kanishcheva (2), Thierry Hamon (3)
1. CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
2. Intelligent Computer Systems Dept, National Technical University Kharkiv Polytechnical Institute, Kharkiv, Ukraine
3. LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France; Université Paris 13, Sorbonne Paris Cité, F-93430 Villetaneuse, France

1. Introduction

Creation of linguistic resources is a very important issue for linguistics and Natural Language Processing. Their availability provides the possibility to design, develop and evaluate methods and tools specific to several contexts and applications (information retrieval, acquisition of lexica, machine translation, categorization of documents...).

The purpose of this work is to describe multilingual parallel and aligned corpus, in which the target language is Ukrainian, while the current source languages are Polish, French and English. This means that the corpus is built with texts from these three languages (Polish, French and English) and their translations in Ukrainian.

Ukrainian language is currently provided with little freely available resources. We are mostly interested here by the available corpora, including parallel corpora. Among the existing work, let's notice the national corpus of the Ukrainian language (Дарчук, 2010) available online (http://www.mova.info/corpus.aspx?l1=209) and literary corpus with the work by Ivan Franko (Бук, 2010) built for the research and educational purposes, and a corpus with dialectal texts (Сірук, 2012). Besides, several parallel corpora involving Ukrainian have been proposed, such as Polish-Ukrainian (Kotsyba, 2012) and Bulgarian-Ukrainian (Siruk et al., 2013) corpora.

2. Collection of texts

For the building of our corpus, we use two kinds of texts. The first source is composed of literary corpus in Ukrainian collected from the УкрЛіт and UkrLib websites. The purpose of these websites is to promote literature in Ukrainian. According to the policy of these websites, these works are publicly available and can be used as far as they are cited. For the translated works, we collected publicly available originals from websites like Project Gutenberg (https://www.gutenberg.org/). Three source languages are covered: Polish, French and English.

The second source is composed of medical documents from the MedlinePlus website (https://medlineplus.gov/). These documents contain patient-oriented brochures on several medical topics. These brochures have been created in English and translated in several languages, among which Ukrainian. Here again, Ukrainian is the target language.

Corpus	Occurrence of words	Number of texts
Literature/UK	3,111,656	110
Literature/FR	1,310,732	29
Literature/EN	2,203,350	51
Literature/PL	260,536	30
MedlinePlus/UK	43,184	129

MedlinePlus/EN	46,544	129

In Table above, we indicate the size of the collected corpora (number of texts and number of word occurrences) for each language: Ukrainian (UK), French (FR), Polish (PL), and English (EN). Overall, we have over 3M word occurrences in Ukrainian.

This dataset contains parallel texts. These source languages have been chosen for their representativity and relation with the Ukrainian language. Polish is also a Slavic language, and is close to Ukrainian. Polish is now quite well researched within the NLP field. We assume that the methods and tools developed for the Polish language can be adapted to Ukrainian provided that there are suitable corpora and resources. English and French languages are well researched from the NLP point of view. We assume, it is possible to take advantage of this research using the transfer methodologies (Yarowsky et al., 2001; Lopez et al., 2002), provided that there are suitable parallel and aligned corpora, and resources.

3. Building of corpus

The original documents may be in different formats (pdf, word, text, html). They are all converted in the text format. Besides, the documents are also converted in the UTF-8 encoding. Then, the text files are automatically segmented in sentences in each language, for which we use strong punctuation and upper-cased characters. Ideally, such segmentation should provide corpus aligned at the sentence level. Yet, it is necessary to verify the correctness of the segmentation in sentences and the parallelism between the source and target versions of a given document. Indeed, during the translation process, the organization of the sentences and their segmentation can be modified by the translator in order to better convey the meaning. Besides, some sentences can also be omitted. Hence, the manual control and correction during the alignment at the sentence level is necessary. This is a very long and thorough yet necessary process, as it guarantees the quality of the aligned corpora. Only part of the whole set of texts available is aligned.

Corpus	Source	Target
Litterature/FR	507,063	419,479
Literature/EN	502,393	424,730
Literature/PL	260,536	264,200
Medline/EN	46,544	43,184

Table above indicates the size of the currently aligned texts, each of which has undergone manual verification. On the whole, the aligned corpus provides 1,151,593 word occurrences in the target Ukrainian language. As we can see, all medical texts and all literary texts in the Polish/Ukrainian pair has been aligned and verified, while only part of French and English source texts is operational up to now. The current version of this parallel and aligned corpus is intended to grow with new texts: other texts are being checked for the correct alignment.

4. Current usage of the corpus

Up to now, the medical part of the corpus has been used for the acquisition of medical terminology in three languages (English, French and Ukrainian) using the transfer methodology (Hamon & Grabar, 2016). This work relied on terminology acquisition methods and tools in English and French. It permitted to build a set with 4,588 terms in Ukrainian and the corresponding terms in French and English, for 34,267 relations in the acquired network.

5. Conclusion and Future Work

In this work, we proposed a description of parallel corpus in which Ukrainian is the target language, while the source languages are Polish, French and English. The corpus mainly contains fiction work but also some texts from the medical field. This corpus is partly aligned at the level of sentences. There are some current exploitations of the corpus for the acquisition of medical terminology.

Future exploitations of this corpus may be related to the machine translation, to the acquisition of cross-lingual paraphrases and disambiguation, to various contrastive studies (including stylistics and discourse). An important issue is the creation of tools for the linguistic processing of texts in Ukrainian, like POS-tagging and syntactic parsing.

A subset of the texts is being aligned by two annotators, so that the inter-annotator agreement can be computed. Besides, tools for the automatic alignment of sentences are being investigated, which may allow to enrich the set with the aligned sentences.

References

Дарчук, Дослідницький корпус української мови: основні засади і перспективи. ВІСНИК Київського національного університету імені Тараса Шевченка 21, 45-49 (2010)

Бук, Лінгводидактичний потенціал корпусу текстів Івана Франка у викладанні української мови як іноземної. In: Theory and Practice of Teaching Ukrainian as a Foreign Language. pp. 70-74 (2010)

Сірук, Підготовка діалектних текстів для корпусного опрацювання. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 43-45 (2012)

Hamon, T., Grabar, N.: Adaptation of cross-lingual transfer methods for the building of medical terminology in Ukrainian. Computational Linguistics and Intelligent Text Processing, pp. 1-12 (2016)

Kotsyba, N.: Polukr (a Polish-Ukrainian parallel corpus) as a testbed for a parallel corpora toolbox. Philological Studie LXIII, 181-196 (2012)

Siruk, O., Derzhanski, I.: Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. Digital Presentation and Preservation of Cultural and Scientific Heritage 3, 91-98 (2013)

Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: HLT (2001)

Lopez, A., Nossal, M., Hwa, R., Resnik, P.: Word-level alignment for multilingual resource acquisition. In: LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data. Las Palmas, Spain (2002)