

Creation of a multilingual aligned corpus with Ukrainian as the target language and its exploitation

Natalia Grabar¹ and Thierry Hamon²

¹ CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

² LIMSI-CNRS, Orsay, Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr

Abstract. The question on creation of linguistic resources (such as corpora, lexica or terminologies) occupies an important place in the research areas related to linguistics, Natural Language Processing, Computer Sciences, psycholinguistics, etc. In this paper, we propose the description of a multilingual corpus in which Ukrainian is the target language, while source languages are Polish, French and English. The corpus contains literary texts and a small subset built with texts provided by medical area. On the whole, the corpus is composed of 62 literary texts and 129 medical texts. The corpus counts over 1 million words in the target Ukrainian language, and at least as much in the source languages taken all together. This is a directional corpus aligned at the level of sentences. After the description of this corpus, we introduce some possible exploitations and first results. We then conclude and indicate some directions for future work. The corpus presented in this work is available for the research purposes: <http://natalia.grabar.free.fr/resources.php>

1 Introduction

The question on creation of linguistic resources (such as corpora, lexica or terminologies) occupies an important place in the research areas related to linguistics, Natural Language Processing, Computer Sciences, psycholinguistics, etc. Indeed, the availability of such resources provides the possibility to design, develop and evaluate methods and tools specific to several contexts and applications (information retrieval, acquisition of lexica, machine translation, question/answering, categorization of documents...). As a matter of fact, different applications may require the availability of different kinds of resources.

The purpose of this work is to introduce and describe multilingual parallel and aligned corpus, in which the target language is Ukrainian, while the current source languages are Polish, French and English.

In what follows, we describe first the existing resources and NLP tools developed for the Ukrainian language (section 2), and then present our method for collection (section 3) and building (section 4) of the parallel and aligned corpus. We then present some possible exploitations of this aligned corpus and the currently obtained results (section 5). We conclude with directions for future research (section 6).

2 Existing resources and methods for Ukrainian

Ukrainian language is part of the Slavic family of languages. Currently, little resources are freely available for Ukrainian, especially when looking at NLP tools and resources. We propose here a short review of some existing resources and tools: corpora, morphological resources, dictionaries and terminologies, and NLP tools.

2.1 Corpora

We have found several corpora dedicated to the description of the modern Ukrainian language: national corpus of the Ukrainian language [33] which is available online³, literary corpus with the work by Ivan Franko [29] built for the research and educational purposes, and corpus with dialectal texts [38].

Besides, several parallel corpora involving Ukrainian have been proposed, such as Polish-Ukrainian [16] and Bulgarian-Ukrainian [23] corpora. Let's also notice a platform for the development and repository of comparable corpora in several languages including Ukrainian [4].

Although it started recently, there is an ongoing research on building of the electronic corpora [14, 13, 34], and on the related research questions such as representativeness of corpora [35], general methodological basis for the creation of corpora [26], creation of signed corpora [39], morphological annotation of corpora [32], methods for frequency studies [30].

2.2 Morphological resources

Two sets of morphological resources dedicated to Ukrainian can be mentioned: Multex-East Ukrainian lexicon for the general language⁴ with morphological features [8, 17], and a corpus-based lexicon with pairs of morphologically related words from general and medical area languages [9]. Let's also notice the Mondilex infrastructure with digital resources in Slavic lexicography [7], that gathers resources for several Slavic languages.

2.3 Dictionaries and terminologies

Several dictionaries exist and describe the general language and specialized areas in Ukrainian. Yet, such dictionaries are mostly available in traditional paper format. Nevertheless, we can notice the frequency dictionary of texts written by Ivan Franko [28], and an electronic dictionary of fire security [40]. Besides, some of the existing dictionaries can be queried online⁵.

Notice that the current research in Ukraine increasingly addresses the use of electronic corpora for the building of dictionaries and terminologies [37, 27, 31], and the transformation of traditional dictionaries in electronic format [36]. As for the terminology-related research, a short review has been proposed [10].

³ <http://www.mova.info/corpus.aspx?11=209>

⁴ <https://www.clarin.si/repository/xmlui/handle/11356/1041>

⁵ <http://lcorp.ulif.org.ua/dictua/>

2.4 NLP tools

Among the existing NLP tools, we can mainly mention two Part-of-Speech (POS) taggers: the UGtag POS tagger [18] which does not perform the syntactic and morphological disambiguation, and a TNT model for Ukrainian [1].

3 Collection of texts

For the purpose of our objectives, we use two kinds of texts, one covering general and one covering specialized languages:

- *Literature*. The literary corpus in Ukrainian is collected from the *UkrLit*⁶ and *UkrLib*⁷ websites which purpose is to promote literature in Ukrainian, with both original and translated works. According to the policy of these websites, these works are publicly available and can be used as far as they are cited. For some translated works, we could collect publicly available originals from websites like *Project Gutenberg*⁸. Three source languages are thus covered: Polish, French and English. This set of data contains the literary work written in Polish, French or English, and then translated in Ukrainian. These data provide a good basis for the creation of parallel corpora;
- *MedlinePlus*. Medical documents are obtained from the MedlinePlus of the National Library of Medicine⁹. These documents contain patient-oriented brochures on several medical topics, such as body systems, disorders and conditions, diagnosis and therapy, demographic groups, health and wellness. These brochures have been created in English and then translated in several languages, among which Ukrainian. These works, produced by the MedlinePlus team, are not copyrighted under U.S. law and can be freely used. Here again, Ukrainian is the target language.

<i>Corpus</i>	<i>Occ_{words}</i>	<i>Nb_{text}</i>
<i>Literature/UK</i>	3,111,656	110
<i>Literature/FR</i>	1,310,732	29
<i>Literature/EN</i>	2,203,350	51
<i>Literature/PL</i>	260,536	30
<i>MedlinePlus/UK</i>	43,184	129
<i>MedlinePlus/EN</i>	46,544	129

Table 1. Size of the collected parallel texts per language, in terms of number of texts and word occurrences.

In Table 1, we indicate the size of the collected corpora for each language: Ukrainian *UK*, French *FR*, Polish *PL*, and English *EN*. This dataset contains parallel texts, while

⁶ <http://ukrlit.org>

⁷ <http://www.ukrlib.com.ua/>

⁸ <http://www.gutenberg.org/>

⁹ www.nlm.nih.gov/medlineplus/healthtopics.html

in each pair of languages Ukrainian is the target language. The other three languages (French, Polish and English) are the source languages. Among the English-language authors we can find Charlotte Bronte, Lewis Carroll, Izak Azimov, Raymond Chandler, Agatha Christie, James Joyce, Jack London, George Orwell and JRR Tolkien. Among the French-language authors we can find Honoré de Balzac, Albert Camus, Alexandre Dumas, Charles Perrault, Guy de Maupassant, Antoine de Saint-Exupéry and Jules Verne. The Polish-language texts have all been written by Stanislaw Lem.

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 [24, 21], provided that there are suitable parallel and aligned corpora, and resources.

As indicated in Table 1, the Ukrainian part of the corpus is the most extensive because it covers the works in the three source languages. We can also observe that specialized subset of texts contains greater number of documents but smaller number of word occurrences. This subset is much smaller than the literary work subset.

4 Building of corpus

The documents indicated in Table 1 are all converted in the text format and the UTF-8 encoding. The original documents can be in different formats (text, word, pdf, html...). We use Linux tools for converting them into text, such as `pdftotext`, `antiword` or home-made `perl` program `html2txt`. For managing the encoding, we use the Linux tool `recode`. Once these two aspects are homogeneous, these text files are segmented in sentences in each language, for which we use strong punctuation and upper-cased characters. Specific `perl` scripts have been created for each of the processed languages.

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. For instance, in Charlotte Bronte's *Jane Eyre*, the source sentence in Example (1) is segmented in two sentences during its translation in Ukrainian (by Петро Соколовський), as indicated in Example (2).

- (1) *I was glad of it: I never liked long walks, especially on chilly afternoons: dreadful to me was the coming home in the raw twilight, with nipped fingers and toes, and a heart saddened by the chidings of Bessie, the nurse, and humbled by the consciousness of my physical inferiority to Eliza, John, and Georgiana Reed.*

- (2) *Щодо мене, то я була рада: я страх не любила довгих пообідніх прогулянок, а надто взимку. Жахливо було вертатися додому в холодному присмерку, коли заходять зашпори в руки і в ноги, а серце ниє від сердитого бурчання Бесі, нашої няні, та від принизливого усвідомлення фізичної переваги наді мною Елізи, Джона та Джорджіани Рід.*

Hence, the manual control and correction during the alignment at the sentence level is necessary. This is a very long and thorough process, which guarantees the quality of the aligned corpora. Notice that the human annotator must understand the source and target languages involved in order to be able to control the correct alignment of sentences.

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

Table 2. Currently aligned corpora, size indicated in word occurrences in each language.

In Table 2, we indicate 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. In Table 2, we can also observe that the Ukrainian texts translated from English and French are usually shorter in number of words than the original texts, while the translation from Polish contains slightly higher number of words.

5 Exploitation of aligned corpus

In Figures 1 and 2, we present two excerpts from the English/Ukrainian sentence-aligned corpora: literary corpus from Charlotte Bronte's *Jane Eyre* and medical corpus, respectively.

These aligned corpora can be used for instance for the acquisition of bilingual lexica for the general and medical languages, for the acquisition of paraphrases [3, 5, 15], for the stylistic analysis of the source and target languages, for the contrastive studies, and for the machine translation. For instance, we have started to use the *Medline* aligned corpus for the acquisition of bilingual medical terminology in Ukrainian thanks to the use of the multilingual transfer [11]. Hence, in Figure 3, we underline the terms extracted in the English text and then transferred on the Ukrainian text thanks to their further alignment at the word level with the GIZA++ algorithm [22].

<i>English</i>	<i>Ukrainian</i>
"What does Bessie say I have done?" I asked.	— Що вам Бесі наговорила на мене?— спитала я.
"Jane, I don't like cavillers or questioners; besides, there is something truly forbidding in a child taking up her elders in that manner. Be seated somewhere; and until you can speak pleasantly, remain silent."	— Джейн, я не люблю, коли чіпляються до слів і допитуються. Дитина не сміє так розмовляти зі старшими! Іди сядь собі десь і, поки не навчишся бути чемною, мовчи.
A breakfast-room adjoined the drawing-room, I slipped in there.	З вітальні був хід у невеличку їдальню; отож я й шмигнула туди.
It contained a bookcase:	Там стояла шафа з книжками.
I soon possessed myself of a volume, taking care that it should be one stored with pictures.	Я вибрала собі одну з них, спершу подивившись, чи вона з малюнками.

Fig. 1. Example of the sentence-aligned literary corpus (English/Ukrainian), from Charlotte Bronte's *Jane Eyre*.

<i>English</i>	<i>Ukrainian</i>
Cancer cells grow and divide more quickly than healthy cells.	Ракові клітини ростуть і діляться швидше, ніж здорові клітини.
Cancer treatments are made to work on these fast growing cells.	При лікуванні раку здійснюється вплив на ці клітини, що швидко ростуть.
- Tiredness	- Втома
- Nausea or vomiting	- Нудота або блювота
- Pain	- Біль
- Hair loss called alopecia	- Втрата волосся, що називається алопецією

Fig. 2. Example of the sentence-aligned MedlinePlus corpus (English/Ukrainian), file *CANCERTREATMENTSIDEFFECTS.TXT*.

<i>English</i>	<i>Ukrainian</i>
<u>Cancer cells</u> grow and divide more quickly than <u>healthy cells</u> .	<u>Ракові клітини</u> ростуть і діляться швидше, ніж <u>здорові клітини</u> .
<u>Cancer treatments</u> are made to work on these <u>fast growing cells</u> .	При <u>лікуванні раку</u> здійснюється вплив на ці <u>клітини, що швидко ростуть</u> .
- <u>Tiredness</u>	- <u>Втома</u>
- <u>Nausea or vomiting</u>	- <u>Нудота</u> або <u>блювота</u>
- <u>Pain</u>	- <u>Біль</u>
- <u>Hair loss called alopecia</u>	- <u>Втрата волосся, що називається алопецією</u>

Fig. 3. Example of the transferred terminological units using sentence-aligned MedlinePlus corpus (English/Ukrainian), file *CANCERTREATMENTSIDEFFECTS.TXT*.

Besides, parallel and aligned corpora can provide other interesting insights on language and grammar, typically issued from contrastive linguistics studies and Natural Language Processing. Let's mention some existing works:

- study of grammatical verbal constructions in English and Norwegian [12];
- cross-lingual disambiguation [2], which shows that, depending on its context of occurrence, the English noun *plant* can be translated as French *plante* ("living thing in soil") or *usine* ("factory"). Such disambiguation of the source text improves the overall results of word sense disambiguation by up to 25%;
- improving the quality of lexicon bootstrapping in one language using translations in other languages [25], which shows that the results with German and English data are improved by 25%;
- semantic study of morphological units [6], in which the semantics of agentive suffixes in French *-iste* and Italian *-ista* rely on translation data obtained from an Italian-French bilingual dictionary and corpora;
- study of translations for out-of-dictionary words and expressions, such as translation of evaluative prefixes [19] or argumentative and discourse-organizing sequences [20]. Hence, in the study on translation of evaluative prefixes [19], the authors found out that several situations are possible: (1) translation with a derivative containing an evaluative prefix {*sous-estimer*, *underestimate*}; (2) translation with a derivative containing a non-evaluative prefix {*sous-utilisé*, *unused*}; (3) translation with a non-prefixed word (which can be a simplex word, a suffixed word or a compound) {*sous-alimenté*, *starving*}, {*sous-équipé*, *ill-equipped*}, {*surpoids*, *obesity*}; (4) translation with a periphrasis {*ultra-concurrence*, *competition taken to extremes*}, {*hyper-fédéraliste*, *extremely federalist*}; (5) zero translation, when the prefixed word is not translated in the target text.

Hence, the availability of parallel and aligned corpora provides several research possibilities for creating and enriching resources for Ukrainian.

6 Conclusion and Future Work

In this work, we propose parallel and aligned corpus involving Ukrainian language. The corpus is aligned at sentence level. This is a directional corpus because the source and target languages, as well as the translation direction are identified: Ukrainian is the target language, while Polish, French and English are source languages. The corpus contains texts from the general language (literary texts) and medical area. On the whole, the aligned corpus contains over 1 million words in the target Ukrainian language, and at least as much in the source languages.

In the future, we plan to extend the currently available aligned corpus with new sentence-aligned texts. The current three source languages (Polish, French and English) will be given advantage. This will allow to efficiently design and exploit transfer methodologies [24, 21] and statistical approaches such as those used in word alignment and machine translation [22]. Besides, several experiments, such as those cited in Section 5, can be performed and open the way to creation and enrichment of terminologies, lexica and contrastive studies involving Ukrainian language.

Another direction for future work consists of creation of parallel corpora, in which Ukrainian is the source language.

In order to make the alignment process and verification easier, we will test and exploit automatic sentence alignment tools. Currently, only one human annotator (NLP researcher) is involved in the building of corpus. If several human annotators are involved in the manual alignment, we will be able to compute the inter-annotator agreement, which will be indicative of the sophistication and difficulty of this task.

This sentence-aligned corpus is freely available for the research purposes:
<http://natalia.grabar.free.fr/resources.php>

References

1. Babych, B.: Representation and interpretation of ambiguous deep syntactic structures. *Ukrainian Linguistics* 21, 89--100 (1997), in Ukrainian
2. Banea, C., Mihalcea, R.: Word sense disambiguation with multilingual features. In: *International Conference on Computational Semantics (ICCS 2011)*. pp. 25--34 (2011)
3. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: *ACL*. pp. 597--604 (2005)
4. Benko, V.: Aranea: Yet another family of (comparable) web corpora. In: *Text, Speech and Dialogue*. pp. 247--256 (2014)
5. Callison-Burch, C., Cohn, T., Lapata, M.: Parametric: An automatic evaluation metric for paraphrasing. In: *COLING*. pp. 97--104 (2008)
6. Cartoni, B., Namer, F.: Linguistique contrastive et morphologie : les noms en -iste dans une approche onomasiologique. In: *CMLF*. pp. 1245--1259 (2012)
7. Dimitrova, L., Koseska-Toszewa, V., Garabik, R., Erjavec, T., Iomdin, L., Shyrovkov, V.: *MONDILEX - Towards the Research Infrastructure for Digital Resources in Slavic Lexicography*, pp. 147--162 (2010)
8. Erjavec, T.: *MULTEXT-East: Morphosyntactic resources for central and eastern european languages*. *Language Resources and Evaluation* 46(1), 131--142 (2012)
9. Grabar, N., Hamon, T.: Acquisition non supervisée de ressources morphologiques en ukrainien. In: *Atelier Traitement Automatique des Langues Slaves (TASLA)*. pp. 1--10 (2015)
10. Grabar, N., Shyshkina, N., Zorko, H., Hamon, T.: Terminological research in ukraine. In: *Terminologie et Intelligence Artificielle (TIA)* (2015)
11. Hamon, T., Grabar, N.: Acquisition of medical terminology for Ukrainian from parallel corpora and Wikipedia. In: *Terminologie et Intelligence Artificielle (TIA)* (2015)
12. Hantson, A.: English gerund clauses and norwegian det + infinitive / at clause constructions. In: Granger, S., Lerot, J., Petch-Tyson, S. (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, pp. 75--90. Rodopi, New-York, Amsterdam (2003)
13. Kelih, E.: Preliminary analysis of a slavic parallel corpus. In: *Corpus based Grammar research*. pp. 173--183 (2009)
14. Kelih, E., Buk, S., Grzybek, P., Rovenchak, A.: Project description: designing and constructing a typologically balanced ukrainian text database. In: *Методи аналізу тексту*. pp. 125--132 (2009)
15. Kok, S., Brockett, C.: Hitting the right paraphrases in good time. In: *NAACL*. pp. 145--153 (2010)
16. Kotsyba, N.: Polukr (a polish-ukrainian parallel corpus) as a testbed for a parallel corpora toolbox. *Philological Studie LXIII*, 181--196 (2012)

17. Kotsyba, N.: Overview of the ukrainian language resources within the multilingual european MULTEXT-East project. Інформаційні системи та мережі 770, 122--129 (2013)
18. Kotsyba, N., Mykulyak, A., Shevchenko, I.V.: UGTag: morphological analyzer and tagger for the Ukrainian language. In: Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009) (2009)
19. Lefer, M., Grabar, N.: Evaluative prefixes in translation: From automatic alignment to semantic categorization. Linguistic Issues in Language Technology journal 11(6), 169--187 (2014)
20. Lefer, M.A., Grabar, N.: N-grams in multilingual corpora: extracting and analyzing lexical bundles in contrastive studies. In: EUROPHRAS 2015 (2015)
21. 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)
22. Och, F., Ney, H.: Improved statistical alignment models. In: ACL. pp. 440--447 (2000)
23. 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)
24. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: HLT (2001)
25. Ziering, P., van der Plas, L., Schütze, H.: Multilingual lexicon bootstrapping. Improving a lexicon induction system using a parallel corpus. In: International Joint Conference on Natural Language Processing. pp. 844--848 (2013)
26. Бобкова, Історичні та концептуальні передумови корпусної лінгвістики. Філологічні науки 2, 13--17 (2014)
27. Бугаков, О.: Создание семантического словаря предложных конструкций на основе украинского национального лингвистического корпуса. Tech. ger., Украинский языково-информационный фонд НАН Украины, Киев, Украина (2006)
28. Бук, Ровенчак, Частотний словник роману Івана Франка "Перехресні стежки", pp. 138--369 (2007)
29. 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)
30. Бук, Сучасні методи дослідження мови письменника у слов'янознавстві. Проблеми слов'янознавства 61, 86--95 (2012)
31. Глибовец, А., Решетнев, І.: Метод ітеративного побудови термінології в колекціях наукових текстів на українському мові. Кибернетика и системний аналіз 50(6), 53--62 (2014)
32. Дарчук, Н.: Морфологічне анотування Корпусу української мови. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 16--18 (2012)
33. Дарчук, Дослідницький корпус української мови: основні засади і перспективи. ВІСНИК Київського національного університету імені Тараса Шевченка 21, 45--49 (2010)
34. Демська, О.: Текстовий корпус: ідея іншої форми. ВПЦ НаУКМА, Київ, Україна (2011)
35. Демська-Кульчицька, О.: Репрезентативність як ознака текстового корпусу. Українська мова 3, 100--107 (2005)
36. Левченко, О., Кульчицький, І.: Технологія перетворення п'ятимовного словника порівнянь в електронну форму. In: Інформаційні системи та мережі. pp. 129--138 (2013)
37. Монахова, Т.: Застосування прийомів корпусної лінгвістики в лексикографії. Наукові праці 98(85), 55--60 (2009)
38. Сірук, Підготовка діалектних текстів для корпусного опрацювання. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 43--45 (2012)

39. Тищенко, Засади створення корпусу української жестової мови глухих. Лексикографічний бюлетень 13, 47--52 (2006)
40. Шуневич, Українсько-англійський комп'ютерний словник пожежно-технічних термінів: лексичні матеріали, програмне забезпечення. In: Комп'ютерна лінгвістика: сучасне та майбутнє. pp. 46--48 (2012)