# Multilingual aligned corpus with Ukrainian as the target language

Natalia Grabar[1], Olga Kanishcheva[2], Thierry Hamon[3,4]

[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*
[4] *Université Paris 13, Sorbonne Paris Cité, F-93430 Villetaneuse, France*
*natalia.grabar@univ-lille3.fr, hamon@limsi.fr*

# Motivation

**Context**:

- Importance of linguistic resources for NLP:
  - lexicons, corpora, terminologies...
- Relevant for various NLP applications:
  - information retrieval, acquisition of lexicons, machine translation, question/answering, categorization of documents...

**Objective**:

- Introducing and describing the Ukrainian target corpus
  - multilingual parallel and aligned corpus
  - source languages: English, French, Polish
  - target language: Ukrainian
- Presenting usages of the corpus

# Existing corpora for Ukrainian

- Ukrainian language is currently provided with little freely available resources
- Available corpora:
  - National corpus of the Ukrainian language [Дарчук 2010a]
  - Literary corpus with the work by Ivan Franko built for the research and educational purposes [Бук 2010b]
  - Corpus with dialectal texts [Сірук 2012]
- Parallel and comparable corpora:
  - Polish-Ukrainian [Kotsyba 2012]
  - Bulgarian-Ukrainian [Siruk and Derzhanski 2013]

# Overview

1. Building of the parallel corpus with Ukrainian as target language
2. Current use of the corpus
   - Unsupervised acquisition of morphological resources
   - Terminology acquisition

# Collection of texts

Two kinds of texts:

- Literary texts translated in Ukrainian
- Medical documents originally written in English

Publicly available sources from several websites

# Literary corpus

- Ukrainian texts from two websites: УкрЛіт *http://ukrlit.org*, UkrLib *http://www.ukrlib.com.ua/*
  Promotion of the literature in Ukrainian
- Original texts: *http://www.gutenberg.org/*
- Source languages: Polish, French, English
- Target language: Ukrainian

Authors: Charlotte Bronte, Lewis Carroll, Raymond Chandler, Agatha Christie, James Joyce, Jack London, George Orwell, JRR Tolkien, Honoré de Balzac, Albert Camus, Alexandre Dumas, Charles Perrault, Guy de Maupassant, Antoine de Saint-Exupéry, Jules Verne, Stanislaw Lem...

# Medical corpus

- Texts from NLM MedlinePlus website: *www.nlm.nih.gov/medlineplus/healthtopics.html*
- Patient-oriented brochures: body systems, disorders and conditions, diagnosis and therapy, demographic groups, health and wellness...
- Source language: English
- Target language: Ukrainian (among others)

# Collection of texts

| *Corpus* | *Number of texts* | *Occurrence of words* |
|---|---|---|
| *Literature/UK* | 110 | 3,111,656 |
| *Literature/FR* | 29 | 1,310,732 |
| *Literature/EN* | 51 | 2,203,350 |
| *Literature/PL* | 30 | 260,536 |
| *MedlinePlus/UK* | 129 | 43,184 |
| *MedlinePlus/FR* | 129 | 53,067 |
| *MedlinePlus/EN* | 129 | 46,544 |

# Building of corpus

- Texts collected in PDF, Word, Text, HTML
- Conversion in text, UTF-8
- Automatic segmentation in sentences, based on strong punctuation and upper-cased characters
- Alignment: difference in source and target texts due to the translation
  - Changes in the sentence organization
  - Omitted sentences

⇒ necessary to manually verify the alignment

| *English* | *Ukrainian* |
|---|---|
| "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; | — Джейн, я не люблю, коли чіпляються до слів і до-питуються. |
| besides, there is something truly forbidding in a child taking up her elders in that manner. | Дитина не сміє так розмовляти зі старшими! |
| Be seated somewhere; and until you can speak pleasantly, remain silent." | 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 їдаленьку; 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. |

# Currently aligned texts

| *Corpus* | *Source* | *Target(UK)* | Verified alignment |
|---|---|---|---|
| *Literature/FR* | 507,063 | 419,479 | (✓) |
| *Literature/EN* | 502,393 | 424,730 | (✓) |
| *Literature/PL* | 260,536 | 264,200 | ✓ |
| *Medline/EN* | 46,544 | 43,184 | ✓ |
| *Medline/FR* | 53,067 | 43,184 | ✓ |

Available for the research purposes:
http://natalia.grabar.free.fr/resources.php

# Usage of the corpus

- Unsupervised acquisition of morphological resources
- Taking advantage of resourced-languages
  - Using English and French NLP tools with transfer methodology [Yarowsky et al. 2001, Lopez et al. 2002]
    Provided that there are suitable parallel and aligned corpora
  - Tunnig methods and tools developed for the Polish language
    Provided that there are suitable corpora and resources

# Unsupervised acquisition of morphological resources

Extraction of word pairs related morphologically

[Hamon and Grabar 2017]

- Identification of words from the same morphologically family
  - Inflected forms of unknown words: {гематома; гематоми} (hematoma)
  - Derivations: {вакцинацію; вакцина} ({*vaccination*, *vaccine*})
  - Compounds: {ангіопластика; ангіограми} ({*angioplasy*, *angiogram*})
- Method [Zweigenbaum et al. 2003]
  - Sliding window with *M* words ($M = 10$ on left and on right)
  - Same initial sub-string with at least *c* characters ($c = 3$) without taking into account prefixes
  - Cooccurrence in the same window
  - Association strength: Likelyhood ratio [Manning and Schutze 1999]
- Results on Ukrainian medical corpus: 1,757 pairs (prec.: 76.7%)

# Terminology acquisition

Use of the medical part of the corpus: acquisition of medical terminology in three languages (English, French and Ukrainian) using the transfer methodology [Hamon and Grabar 2016]

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

Hypothesis:

- parallel and word aligned corpora with two languages $L1$ and $L2$
- syntactic or semantic annotations from $L1$

Method:

- transpose these annotations from $L1$ to $L2$,
- obtain the corresponding annotations in $L2$

The transfer methodology depends on the quality of the annotation on $L_1$ texts, the quality of word alignment and the corpus size

# Terminology acquisition: Results

- Use of terminology acquisition methods and tools in English and French
- Extraction of 4,588 terms in Ukrainian (prec.: 0.454) and the corresponding terms in French (prec.: 0.674) and English (prec.: 0.761)
- Detection of 34,267 relations between these terms (prec.: between 0.309 and 0.419)

# Conclusion

- Description of parallel corpus with Ukrainian as target language, and Polish, French and English as source languages
- Texts issued from literature and medical domain
- Corpus partly aligned at the sentence level
- Current use of the medical texts for
  - acquisition of morphological resources
  - terminological extraction

# Future work

- Pursue the sentence level alignment
- Use of automatic sentence aligners
- Use of the corpus:
  - Acquisition of cross-lingual paraphrases and disambiguation
  - Creation of tools for the linguistic processing of texts in Ukrainian, like POS-tagging and syntactic parsing

Сірук (Олена). --

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

Дарчук (Н). --

Дослідницький корпус української мови: основні засади і перспективи. *Вісник Київського національного університету імені Тараса Шевченка*, vol. 21, 2010, pp. 45--49.

Бук (Соломія). --

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.

Lopez (Adam), Nossal (Mike), Hwa (Rebecca) et Resnik (Philip). --

Word-Level Alignment for Multilingual Resource Acquisition. *In : LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data*. --
Las Palmas, Spain, 2002.

Hamon (T) et Grabar (N). --

Adaptation of cross-lingual transfer methods for the building of medical terminology in Ukrainian. *In : Computational Linguistics and Intelligent Text Processing*, pp. 1--12.

Hamon (Thierry) et Grabar (Natalia). --

Unsupervised acquisition of morphological resources for Ukrainian. *In : Proceedings of Computational Linguistics and Intelligent Systems (COLINS 2017)*, pp. 20--30. --
Kharkiv, Ukraine, April 2017.

Siruk (Olena) et Derzhanski (Ivan). --

Linguistic Corpora as International Cultural Heritage: The Corpus of Bulgarian and Ukrainian Parallel Texts. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 3, 2013, pp. 91--98.

Manning (C. D.) et Schütze (H.). --

*Foundations of statistical natural language processing*. --
Cambridge, MA, MIT Press, 1999.

Yarowsky (David), Ngai (Grace) et Wicentowski (Richard). --

Inducing multilingual text analysis tools via robust projection across aligned corpora. *In : HLT*.

Kotsyba (Natalia). --

PolUKR (A Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel Corpora Toolbox. *Philological Studie*, vol. LXIII, 2012, pp. 181--196.

Zweigenbaum (Pierre), Hadouche (Fadila) et Grabar (Natalia). --

Apprentissage de relations morphologiques en corpus. *In : Traitement Automatique des Langues Naturelles (TALN)*.