

## Acquisition of medical terminology for Ukrainian from parallel corpora and Wikipedia

**Thierry Hamon**  
LIMSI-CNRS, Orsay  
U Paris 13  
Sorbonne Paris Cité  
France  
hamon@limsi.fr

**Natalia Grabar**  
UMR8163 STL  
CNRS, U Lille 3  
Villeneuve d'Ascq  
France  
natalia.grabar@univ-lille3.fr

### Abstract

The increasing availability of parallel bilingual corpora and of automatic methods and tools for their processing makes it possible to build linguistic and terminological resources for low-resourced languages. We propose to exploit various corpora available in several languages in order to build bilingual and trilingual terminologies. Typically, terminology information extracted in French and English is associated with the corresponding units in the Ukrainian corpus thanks to the multilingual transfer. According to the used approaches, precision of the term extraction varies between 0.454 and 0.966, while the quality of the interlingual relations varies between 0.309 and 0.965. The resource built contains 4,588 medical terms in Ukrainian and their 34,267 relations with French and English terms.

### 1 Introduction

The acquisition of terminology has gone through a very active period and provides nowadays several automatic tools and methods (Kageura and Umino, 1996; Cabré et al., 2001; Pazienza et al., 2005) for several European languages and Japanese. Nevertheless, other languages remain low-resourced and require specific Natural Language Processing (NLP) developments.

Our main objective is to create terminological resources for Ukrainian, for which very little digitized or electronic resources are available. Yet, the terminology extraction tools usually require the morpho-syntactic tagging of texts, which can be problematic if the corresponding automatic tools are not available for a given language. For instance, the UGtag Part-of-Speech (POS) tagger

(Kotsyba et al., 2009) developed for Ukrainian does not perform the syntactic and morphological disambiguation of the tags. Hence, it becomes impossible to use it for the pre-processing of corpora before the traditional terminology acquisition process.

In this situation, we propose first to compile terminological resources for Ukrainian in order to build the basis for the observation of the specificities of terminological units in this language. Such observations will allow to develop and parameter the terminology extraction tool for Ukrainian.

The motivation of our work is double. We want to

1. automatically build terminologies for Ukrainian,
2. design specific methods for the acquisition of such terminological resources.

The work is carried out with medical data, and in three languages (Ukrainian, French, and English).

The work we present starts from the exploitation of two kinds of corpora (Section 2.1): Wikipedia in Ukrainian which provides several useful kinds of information (such as term labels and their codes) with a high level of quality, and the parallel corpus MedlinePlus. The term detection and extraction can be either manual or automatic. Since, there is no appropriate POS-tagging and term extraction tools for Ukrainian, we propose to use such tools in French and English, and to take advantage of these to transfer English and French extracted terms on the Ukrainian corpus.

Indeed, the transfer methodology can be considered as suitable for such objectives. Suppose we

Створити обліковий запис Увійти

Стаття | Обговорення

Читати | Редагувати | Редагувати код | Перегляд історії

Пошук

**Вікіпедія**  
Вільна енциклопедія

Головна сторінка  
Поточні події  
Нові редагування  
Нові сторінки  
Випадкова стаття

Участь  
Портал спільноти  
Кнайпа  
Довідка  
Пожертування

Панель інструментів  
Посилання сюди  
Пов'язані редагування  
Спеціальні сторінки  
Постійне посилання  
Інформація про сторінку  
Елемент Вікіданих  
Цитувати сторінку  
Статистика  
Відслідкувати

<https://uk.wikipedia.org/wiki/Нанізм#>

**Нанізм** [ред. • ред. код]

Матеріал з Вікіпедії — вільної енциклопедії.

**Нанізм**, або **карликовість** (грец. *ναλός* — карлик) — ненормально низький, карликовий зріст (нижче за 130 см для чоловіків і 121 см для жінок), що є наслідком порушення функції залоз внутрішньої секреції.

**Зніст** [скловати]

- 1 Види нанізму
- 2 Клінічна картина
- 3 Відомі карлики
- 4 Джерела

**Види нанізму** [ред. • ред. код]

- **Гіпотиреоїдний нанізм** — нанізм, зумовлений вродженою гіпофункцією щитоподібної залози.
- **Гіпофізарний нанізм** — нанізм, зумовлений недостатністю передньої частки гіпофіза. У більшості хворих не вдається виявити безпосередню причину порушення секреції гормону росту. В окремих випадках етіологічними факторами спорадичного або набутого нанізму є органічні ураження гіпоталамо-гіпофізарної області (травми, крововиливи, менінгіти, гліоми, краніофарінгіоми, менінгіоми, ангиоми, судинні аневризми, туберкульоз, сифіліс та ін.).
- **Мітральний нанізм** — нанізм, що спостерігається у осіб, які з дитинства страждали вадами мітрального клапана.
- **Надчирковий нанізм** — нанізм, зумовлений гіперфункцією надчирків.
- **Нирковий нанізм** — нанізм, зумовлений хронічними захворюваннями нирок.
- **Рахитичний нанізм** — нанізм, який спостерігається при рецидивуючому тяжкому перебігу рахіту.

**Нанізм**

**MKX-10** E23.0 E34.2 E45.0 Q77.4

**DiseasesDB** 80

**MedlinePlus** 001176

**MeSH** D004392

Figure 1: Example of the Ukrainian Wikipedia source pages (Dwarfism). The infobox with the coding is on the right.

have parallel and aligned corpora with two languages  $L_1$  and  $L_2$ , and we have several types of syntactic or semantic annotations and information associated to  $L_1$ . The transfer approach permits to transpose these annotations or information from  $L_1$  to  $L_2$ , and to obtain in this way the corresponding annotations and information in the  $L_2$  text. From this point of view,  $L_1$  is considered as the source language while  $L_2$  is considered as the target language. This kind of approach is particularly interesting when working with low-resourced languages for which less tools and semantic resources are available. An increasing availability of parallel bilingual corpora, and of automatic methods and tools for their processing makes it possible to build linguistic and terminological resources using the transfer methodology (Yarowsky et al., 2001; Lopez et al., 2002). Very few works have been done in this direction, and we assume they open novel and efficient ways for the processing of multilingual texts in particular from low-resourced languages (Zeman and Resnik, 2008; McDonald et al., 2011). Notice that the modeling of cross-language features aims at using language-independent features to create various types of annotations. Among such features, we can mention part-of-speech, semantic categories or even acoustic and prosodic features.

We propose to apply this method for the acquisition of bilingual or trilingual terminologies involving Ukrainian. In our work, each corpus is ex-

ploited through dedicated methods. The MedlinePlus corpus provides the basis for the building of the terminology, while the Wikipedia corpus permits to enrich this information and helps the word-level alignment of the MedlinePlus corpus.

Terminology-related research on Ukrainian is an active area, although the main terminological work shows mainly theoretical and linguistic orientation (Косцак, 2000; Dmytruk, 2009; Рожанківський and Кузан, 2000; Ivashchenko, 2013; Oliinyk, 2013). Very few works are oriented on the use of terminologies and their automatic processing, such as the software localization (Shyshkina et al., 2010).

In the following of this paper, we first present the material used for the acquisition of bilingual terminology (section 2), and the methods designed for achieving this objective (section 3). We then discuss the results we obtain (section 4), and conclude with directions for the future work (section 5).

## 2 Material

### 2.1 Corpora

We use two kinds of corpora:

- *MedlinePlus*: parallel medical corpus from MedlinePlus. These data are built by MedlinePlus from the National Library of Medicine<sup>1</sup>. They contain patient-oriented

<sup>1</sup>[www.nlm.nih.gov/medlineplus/healthtopics.html](http://www.nlm.nih.gov/medlineplus/healthtopics.html)

| <i>Corpus</i>          | <i>Size (occ of words)</i> |
|------------------------|----------------------------|
| <i>Wikipedia/UKmed</i> | 246,368,411                |
| <i>MedlinePlus/UK</i>  | 43,184                     |
| <i>MedlinePlus/FR</i>  | 53,067                     |
| <i>MedlinePlus/EN</i>  | 46,544                     |

Table 1: Size of the exploited corpora.

brochures on several medical topics (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 other languages, among which French and Ukrainian;

- *Wikipedia*: medicine-related articles from Wikipedia. This corpus is extracted from the Ukrainian part of the Wikipedia using medicine-related categories, such as *Медицина (medicine)* or *Захворювання (disorders)*. The corpus potentially covers a wide range of medical notions. In Figure 1, we indicate an example of the source pages which propose the navigation frame on the left, the text with explanations and the infobox with illustration and coding on the right.

In Table 1, we indicate the size of the corpora. Not surprisingly, the Wikipedia corpus is much larger although only part of its information is exploited, as we will see in the next section.

## 2.2 UMLS: Unified Medical Language System

The UMLS (Unified Medical Language System) (Lindberg et al., 1993) merges several (over 100) biomedical terminologies, such as international terminologies MeSH (NLM, 2001) and ICD (Brämer, 1988). Such international terminologies may exist in several languages. For instance, French and English versions of MeSH are included in the UMLS. No terminologies in Ukrainian are part of the UMLS. Each UMLS term is provided with unique identifiers, which allows to find the corresponding terms in other terminologies or languages.

## 3 Methods

The methods we propose for the extraction of bilingual terminology are adapted to each kind of

corpora and of data they contain: the MedlinePlus corpus (section 3.1) and the Wikipedia corpus (section 3.2). We then present their cross-fertilization (Section 3.3), and the evaluation of the results (Section 3.4).

### 3.1 Extraction of bilingual terminology from the MedlinePlus corpus

Prior to the exploitation of the MedlinePlus data, the documents are first transformed in a suitable format:

- the source PDF documents are converted in the text format;
- in each language, the documents are segmented in paragraphs;
- the alignments French/Ukrainian and English/Ukrainian are generated, in which  $n_{th}$  paragraph from one language is associated with the  $n_{th}$  paragraph from the other language;
- the alignment between the two pairs of languages is then verified manually.

In Figure 2, we present an excerpt from the English/Ukrainian aligned corpus.

Then in French and English, we can use the existing terminology extraction tools which results bootstrap the acquisition of bilingual terminology. Hence, we use the  $\text{\AA}T\text{\E}A\text{term}$  extractor (Aubin and Hamon, 2006), that is applied to documents POS-tagged. The extracted terms are then projected on the French and English corpora. In Figure 2, candidate terms are marked in bold.

The exploitation of the MedlinePlus parallel and aligned corpus is performed in several ways (Figure 3).

**Transfer 1** First, the simplest situation is when the two aligned lines contain term candidates in either language: these terms are recorded as candidates for the alignments. For instance, in Figure 2, the pairs *{Tiredness, Втома}* and *{Pain, Біль}* are issued from this kind of alignment.

**Transfer 2** Secondly, when the paragraphs contain complex expressions or sentences, the processing is done as follows (Figure 4):

1. the paragraph-aligned corpora are aligned at the word level using GIZA++ (Och and Ney, 2000),

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

Figure 2: Example of the paragraph-aligned MedlinePlus corpus (English/Ukrainian).

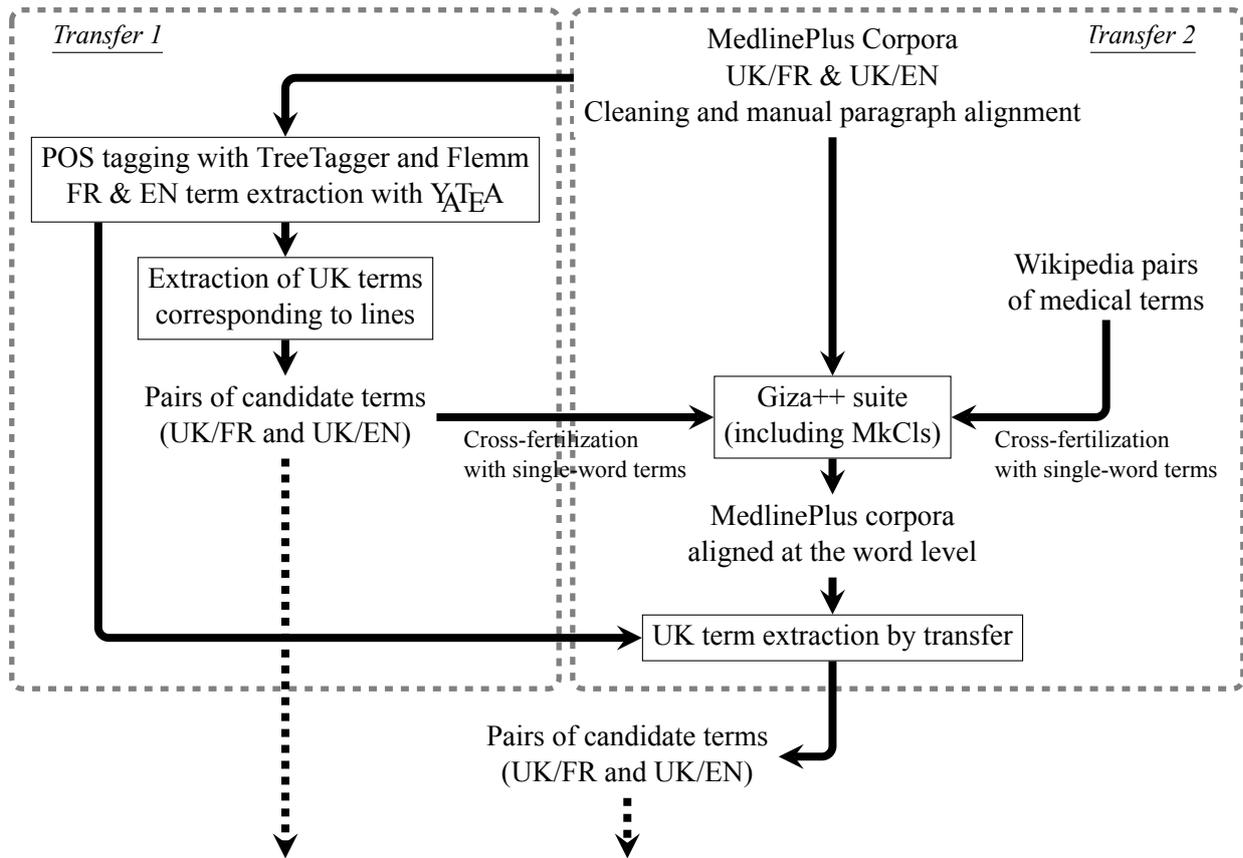


Figure 3: Extraction of medical terms from MedlinePlus corpora (Ukrainian = UK, French = FR, English = EN).

2. using the word-aligned corpora, in each paragraph pair (French/Ukrainian and English/Ukrainian), the terms recognized in French and English are transferred on the Ukrainian paragraph (conceived as the target language);
3. the alignments extracted are recorded as candidates for building the bilingual terminology.

For instance, in Figure 2, the term *Cancer cells* is automatically extracted from the English corpus. GIZA++ proposes that *Cancer cells* is aligned with *Ракові клітини*. Thus, through the word-aligned text, we can propose that *Cancer cells* is the translation of *Ракові клітини*. This processing is performed on the two pairs of languages (French/Ukrainian and English/Ukrainian).

As indicated in Table 1, the size of our corpora is rather small for the statistical alignment performed

by GIZA++. For this reason, we provide GIZA++ with a bilingual dictionary in order to help the alignment at the word level (see Section 3.3). Besides, in preliminary experiments, we also observe that word level alignment errors lead to the extraction of Ukrainian stopwords as term candidates (*на* (*on*), *або* (*or*), etc.). To remove such obvious errors, we filter out such candidates if they occur in a list of 385 stop-word forms issued from an existing resource dedicated to the localization of graphical interfaces<sup>2</sup>.

### 3.2 Extraction of bilingual terminology from the Wikipedia corpus

The Wikipedia corpus is used to complete and to help the method applied to the MedlinePlus corpus. The content we propose to exploit is included in infoboxes (on the right in Figure 1) and is reachable through the MediaWiki source code of the Wikipedia. This provides the label of the medical terms in Ukrainian and their MeSH codes. The process is the following (Figure 4):

1. the infobox content is extracted and parsed<sup>3</sup> in order to obtain the term label and its MeSH code,
2. the MeSH code is used to query the UMLS, and to get the corresponding French and English terms,
3. the term pairs French/Ukrainian and English/Ukrainian are then built and provide good candidates for the bilingual terminology.

This part of the method exploits specific and intentionally created content for a given medical notion in Ukrainian: term for a given medical notion and its MeSH code. This information is reliable. For instance, in Figure 1, the term *нанізм* is extracted, as well as its MeSH code D004392. Through the UMLS, the corresponding English terms are *dwarfism* and *nanism*, while the corresponding French term is *nanisme*. Notice that similar method has been used for the building of medical terminology in the Arabic language (Vivaldi and Rodríguez, 2014).

<sup>2</sup><https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt>

<sup>3</sup>We use the Perl module `Text::MediawikiFormat` (<http://search.cpan.org/~szabgab/Text-MediawikiFormat>)

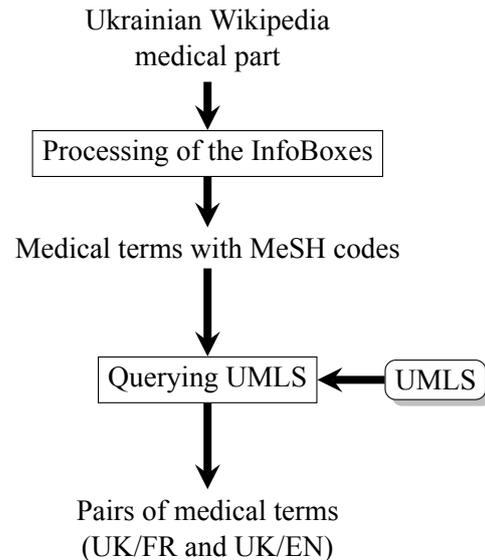


Figure 4: Extraction of medical terms from Wikipedia (Ukrainian = UK, French = FR, English = EN).

### 3.3 Cross-fertilization and Experiments

The cross-fertilization of the two methods (Sections 3.1 and 3.2) is done in two ways:

- the Wikipedia terms are used to enrich the extracted terminology,
- the single-word terms extracted by other approaches can be provided to GIZA++, as an additional bilingual dictionary, in order to help the alignment of MedlinePlus at the word level.

During preliminary experiments, we test several combinations of parameters for the pre-processing and the alignments. While pre-processing the French corpus, the Part-of-Speech is performed by TreeTagger (Schmid, 1994) and can be improved by the morphological analyzer Flemm (Namer, 2000). We also experiment with the use of GeniaTagger (Tsuruoka et al., 2005) on the English corpus. We also experiment with the use of the terms extracted from Wikipedia, or by the MedlinePlus method Transfer 1, or both, for guiding the Giza++ alignment.

Thus, based on the results of the preliminary experiments, we choose to pre-process the English corpus with TreeTagger and the French corpus with TreeTagger and Flemm. Single-word terms extracted from Wikipedia and by the method Transfer 1 are used as bilingual dictionary to help

the Giza++ word level alignment. We only present the results obtained with this configuration in the following.

### 3.4 Evaluation

The evaluation is performed manually in order to check whether the candidates extracted for building the bilingual terminologies are correct. It has been performed by an Ukrainian native speaker having knowledge in medical informatics. Terms are validated independently in each language, but we also evaluation the bilingual and trilingual relations between the Ukrainian, English and French terms. With this kind of evaluation, precision of the results can be computed, i.e. the ratio between the correct answers and all the answers.

## 4 Results and Discussion

Table 2 presents the results and the precision for the extracted terms by the three methods. Table 3 presents the results and the precision concerning the pairs and triples of terms.

### 4.1 Extraction of bilingual terminology from the Wikipedia corpus

The exploitation of the Wikipedia infobox allow to collect 357 Ukrainian medical terms among which 177 are single-word terms. By querying UMLS with the MeSH codes, those terms are associated with 1428 French terms (among them, 339 single-word terms) and 3625 English terms (among them, 448 single-word terms). The number of French and English terms compared to the number of Ukrainian terms are due to the synonyms proposed by MeSH. As for the bilingual pairs of terms, we obtain 1,515 Ukrainian/French term pairs and 3,789 Ukrainian/English term pairs, including, respectively, 270 and 405 pairs between single-word terms. Since each Ukrainian term is associated with at least one French and English terms, this allows to build 28,840 triples. We consider that the precision of this terminology is 1 because the collecting manner.

### 4.2 Extraction of bilingual terminology from the MedlinePlus corpus

The use of the first method of transfer (Transfer 1) allows to extract 436 Ukrainian terms with a high precision unsurprisingly (0.966). These terms are associated with 316 French terms and 354

English terms in 282 triples between Ukrainian, French and English terms, 63 pairs only between Ukrainian and French terms and 115 pairs only between Ukrainian and English terms, with 0.954, 0.937 and 0.965 precision, respectively. Thus, the Transfer 1 method allows to collect 334 Ukrainian/French term pairs (among them 108 pairs between single-word terms) and 380 Ukrainian/English term pairs (among them 135 pairs between single-word terms). We observe that these relations can involve synonyms in either language: {*фаллопієва труба, trompes de fallope/trompe utérine*} (*fallopian tube*), {*втрата слуху/втрачається слух, hearing loss*}, {*втома, fatigue/tiredness*}. Besides, in Ukrainian, several case forms can be associated to a same English of French form: {*вагітність, pregnancy*} and {*вагітності, pregnancy*}.

As the precision values suggest, this first transfer method leads to few errors. Their analysis shows that they mainly concern partial match between one language and another involved by the translation: {*наяву виразок у роті, mouth sores*} -- lit. (*appearance of*) *mouth sores*, {*ви можете спати, dormir/sleep*} -- lit. *you can sleep*. The silence of the method can be explained by two reasons. First, again the variation due to the translation prevents the transfer 1 method to extract term in French or English. For instance, since the title *Soins* in the French corpus is the English translation of *Your care*, the French term matches with the line, contrary to the English term. The Transfer 2 method will solve this problem. However, the main reason of the silence is the incapacity of the term extractor to identify French or English terms because its extraction strategy or errors in the POS tagging.

As for the second transfer method (transfer 2), we present the results obtained when the pairs of single-words terms issued from the MedlinePlus corpus and from Wikipedia are used to help the GIZA++ alignment. In that context, the transfer 2 method allows to extract 9,040 Ukrainian terms with 0.454 precision (exact match). Precision of the French and English terms is higher: 0.674 and 0.761 respectively (exact match). Moreover, the number of French and English terms is dramatically lower (about -45% and -40%) than in Ukrainian: the rich morphology of the Ukrainian language provides several inflected

| Source                      | UK     |       | FR     |       | EN     |       |
|-----------------------------|--------|-------|--------|-------|--------|-------|
|                             | #terms | Prec. | #terms | Prec. | #terms | Prec. |
| <i>Wikipedia</i>            | 357    | 1     | 1,428  | 1     | 3,625  | 1     |
| <i>MedlinePlusTransfer1</i> | 436    | 0.966 | 316    | 0.971 | 354    | 0.989 |
| <i>inexact match</i>        |        | 0.998 |        | 0.987 |        | 0.997 |
| <i>MedlinePlusTransfer2</i> | 9,040  | 0.454 | 3,671  | 0.674 | 3,597  | 0.761 |
| <i>inexact match</i>        |        | 0.84  |        | 0.726 |        | 0.799 |
| Total                       | 9,529  | 0.481 | 5,200  | 0.769 | 7,335  | 0.883 |
| Total of correct terms      | 4,588  |       | 3,998  |       | 6,476  |       |

Table 2: Number of terms extracted (Ukrainian = UK, French = FR, English = EN).

| Source                      | UK/FR |       | UK/EN |       | UK/FR/EN |       | Total  |       |
|-----------------------------|-------|-------|-------|-------|----------|-------|--------|-------|
|                             | #rel. | Prec. | #rel. | Prec. | #trpl.   | Prec. | #trpl. | Prec. |
| <i>Wikipedia</i>            | 1,515 | 1     | 3,789 | 1     | 28,840   | 1     | 28,840 | 1     |
| <i>MedlinePlusTransfer1</i> | 63    | 0.937 | 115   | 0.965 | 282      | 0.954 | 460    | 0.954 |
| <i>inexact match</i>        |       | 0.984 |       | 1     |          | 0.982 |        | 0.987 |
| <i>MedlinePlusTransfer2</i> | 3,724 | 0.309 | 4,745 | 0.401 | 4,724    | 0.419 | 13,218 | 0.381 |
| <i>inexact match</i>        |       | 0.751 |       | 0.84  |          | 0.586 |        | 0.724 |
| Total                       | 3,798 | 0.318 | 4,819 | 0.41  | 33,845   | 0.918 | 42,462 | 0.807 |
| Total of correct relations  | 1,207 |       | 1,974 |       | 31,086   |       | 34,267 |       |

Table 3: Number of term pairs and triples (Ukrainian = UK, French = FR, English = EN).

forms for a given term (*{напад, нападу}* -- *attack*, *{припадків, припадки}* -- *seizure*, *{кісткeй, кістки}* -- *bones*). Besides, the method allows also to extract synonymous terms (*{приступам, припадків}* -- *attacks/seizures*, *{биття, удару}* -- *beats*). The precision values with the inexact match (the correct term is included or includes the term candidates) are much higher and gain 0.40 points for the Ukrainian terms and 0.05 for the French and English terms. We assume this difference on Ukrainian candidate terms is mainly due to the alignment quality. As for the interlingual relations, the Transfer 2 method collects 3,724 pairs of Ukrainian/French terms with 0.309 precision, 4,745 pairs of Ukrainian/English terms with 0.401 precision and 4,724 triples with 0.419 precision.

An analysis of the results shows that most of the errors are due to the alignment problems. Indeed, we observe that when the alignment is correct, the Ukrainian terms are correctly extracted by the transfer. Otherwise, the errors occur.

Moreover, even if the documents (patient-oriented brochures) are not highly specialized, most of the extracted terms are specific to the medical domain (*{трахеотомією, tracheostomy}*), *{фактори ризику, risk factors}*, *{унрица, sy-*

*ringe}*, *{холестерину, cholesterol}*). Other terms also refer to close and approximating notions which reflects this type of documents: *{діти, children}*, *{здорову їжу, healthy diet}*, *{серцевий напад, heart attack}*, *{склянок рідини, glasses of liquid}*. An interesting observation is that some French and English terms correspond to propositions in Ukrainian: *{не до кінця приготовлену їжу, undercooked foods}* (lit. food which is not fully cooked), *{При цьому обстеженні Ви не відчуєте жодного болю, indolore (rainless)}* (lit. With this exam you will feel no pain).

Finally, all the methods combined allow to build a terminological resource containing 4,588 Ukrainian medical terms and their 34,267 relations with French and English terms.

## 5 Conclusion and Future Work

In this work, we propose to exploit two kinds of freely available multilingual corpora in French, English and Ukrainian. Each corpus is exploited with appropriate methods which allows to extract the term candidates and to create term pairs Ukrainian/French and Ukrainian/English. In particular, French and English corpora are processed with NLP and term extraction tools. Then,

thanks to the transfer methods these terms are transposed on the Ukrainian language. We also propose to use existing terminologies and to exploit simple terms for improving the alignment performed at the word level with GIZA++.

Our future work will address the enrichment of the created resource with terms from other corpora. Besides, in the Wikipedia corpus, we can use other codes, such as those from MKX-10 (ICD10) or MedlinePlus. This will also augment the coverage of the term pairs extracted in the current work. Another perspective of this work is the improvement of the bilingual alignment of documents at the word level. In that respect, we plan to investigate the use of other alignment algorithms, such as Fast-Align (Dyer et al., 2013) or the Lingua::Align toolbox (Tiedemann and Kotzé, 2009). Other curators will be involved. Further improvements of the proposed transfer method can be obtained with statistical and morphological cues.

## Acknowledgments

This work is funded by the LIMSI-CNRS AI project *Outiller l'Ukraine*. We are thankful to the reviewers for their useful comments which permitted to improve the quality of the paper.

## References

- S Aubin and T Hamon. 2006. Improving term extraction with terminological resources. In *FinTAL 2006*, number 4139 in LNAI, pages 380--387. Springer.
- GR Brämer. 1988. International statistical classification of diseases and related health problems. tenth revision. *World Health Stat Q*, 41(1):32--6.
- MT Cabré, R Estopà, and J Vivaldi, 2001. *Automatic term detection: a review of current systems*, pages 53--88. John Benjamins.
- Veronica Dmytruk. 2009. Typological features of word-formation in computing, the internet and programming in the first decade of the XXI century. In *VJK*, pages 1--11.
- Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL/HLT*, pages 644--648.
- VL Ivashchenko. 2013. Historiography of terminology: metalanguage and structural units. In *UDC*, pages 1--22.
- K Kageura and B Umino. 1996. Methods of automatic term recognition. In *National Center for Science Information Systems*, pages 1--22.
- Natalia Kotsyba, Andriy Mykulyak, and Ihor V. Shevchenko. 2009. Utag: morphological analyzer and tagger for the ukrainian language. In *Proceedings of the international conference Practical Applications in Language and Computers (PALC 2009)*.
- DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods Inf Med*, 32(4):281--291.
- Adam Lopez, Mike Nossal, Rebecca Hwa, and Philip Resnik. 2002. Word-level alignment for multilingual resource acquisition. In *LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data*, Las Palmas, Spain.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 62--72, Stroudsburg, PA, USA. Association for Computational Linguistics.
- F Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523--547.
- National Library of Medicine, Bethesda, Maryland, 2001. *Medical Subject Headings*. [www.nlm.nih.gov/mesh/meshhome.html](http://www.nlm.nih.gov/mesh/meshhome.html).
- FJ Och and H Ney. 2000. Improved statistical alignment models. In *ACL*, pages 440--447.
- OY Oliinyk. 2013. Terminology for description of linguistic landscape in native and foreign linguistics. *Terminolohichniy visnyk*, 2(1):1--7.
- Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: An analysis of linguistic and statistical approaches. In Spiros Sirmakessis, editor, *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pages 255--279. Springer Berlin Heidelberg.
- H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44--49.
- Nataliia Shyshkina, Galina Zorko, and Larisa Lesko. 2010. Terminology work and software localization in Ukraine. In *Problems of Cybernetics and Informatics*, pages 17--20.
- Jörg Tiedemann and Gideon Kotzé. 2009. A discriminative approach to tree alignment. In Iustina Ilisei, Viktor Pekar, and Silvia Bernardini, editors, *Proceedings of the International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning (in connection with RANLP'09)*, pages 33 -- 39.
- Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust

- part-of-speech tagger for biomedical text. *LNCS*, 3746:382--392.
- J Vivaldi and H Rodríguez. 2014. Arabic medical term compilation from Wikipedia. In *Proc of CIST 2014*.
- David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT*.
- D Zeman and P Resnik. 2008. Cross-language parser adaptation between related languages. In *NLP for Less Privileged Languages*.
- Орест Косак. 2000. Українська комп'ютерна термінологія. In *Сучасні проблеми в комп'ютерних науках*, pages 39--42.
- Р Рожанківський and М Кузан. 2000. Комп'ютерні проблеми стандартизації термінології. In *Сучасні проблеми в комп'ютерних науках (CCU'2000)*, pages 42--44.

