Adaptation of Cross-Lingual Transfer Methods for the Building of Medical Terminology in Ukrainian

Thierry Hamon, Natalia Grabar

Adaptation of Cross-Lingual Transfer Methods for the Building of Medical Terminology in Ukrainian

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIMSI-CNRS, Orsay, Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr,
https://perso.limsi.fr/hamon/
[2] CNRS, UMR 8163, F-59000 Lille, France;
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille3.fr,
http://natalia.grabar.perso.sfr.fr/

Abstract. An increasing availability of parallel bilingual corpora and of automatic methods and tools makes it possible to build linguistic and terminological resources for low-resourced languages. We propose to exploit corpora available in several languages for building bilingual and trilingual terminologies. Typically, terminology information extracted in better resourced languages is associated with the corresponding units in lower-resourced languages thanks to the multilingual transfer. The method is applied on corpora involving Ukrainian language. According to the experiments, precision of 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.

Keywords: Cross-Lingual Transfer, Parallel Corpora, Terminology, Ukrainian

1 Introduction

Automatic acquisition of terminological resources has gone through a very active period and provides nowadays several automatic tools and methods [10, 5, 21] for several European languages and for Japanese. Nevertheless, other languages remain low-resourced and may require specific Natural Language Processing (NLP) developments, which must take into account morphological specificities of such languages [13, 22], for instance. In the existing studies, statistical methods for extracting collocations and repeated segments are often exploited [26, 6, 9, 22] and allow to extract results with reliable recall but low precision.

We propose to take advantage of the advanced research work done in languages like English or French, and to transpose it on low-resourced languages. For such objectives, we propose to exploit the transfer methodology together with parallel and aligned corpora. We consider that the transfer methodology can be suitable for the objectives related to terminology extraction. The principle is the following. Suppose we have parallel and aligned corpora with two

languages $L1$ and $L2$, and we have several types of syntactic or semantic annotations and information associated to $L1$. The transfer approach permits to transpose these annotations or information from $L1$ to $L2$, and to obtain in this way the corresponding annotations and information in the $L2$ text. From this point of view, $L1$ is considered as the source language while $L2$ 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 [29, 15]. 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 [30, 16]. 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.

In the following of this paper, we start with the presentation of the motivation for this study (section 2). We then present the material used for the acquisition of terminology (section 3), and the methods designed for achieving this objective (section 4). We discuss the results we obtain (section 5), and conclude with directions for future work (section 6).

2 Motivation and Rationale

The motivation of our work is double. We want (1) to design specific methods for the acquisition of terminological resources for low-resourced languages such as Ukrainian, and (2) to automatically build medical terminology for Ukrainian.

If little digitalized resources are currently available for Ukrainian, terminological work is an active research area there, although it mainly is concerned with theoretical and linguistic issues. For instance, terminological descriptions are available for several specialized domains and languages: physics [35]; law [40]; computer science [34, 7]; religion [39]; literature [24]; Crimean tatar language [1, 17]. Then, following recent research orientations, work on construction of electronic corpora [11, 33], on their use for building of terminologies and dictionaries [38, 31, 32], and on transformation of traditional dictionaries in electronic format [37] appear. Still, little work is oriented on automatic building and utilization of terminologies. Since terminology extraction tools require Part-of-Speech (POS) tagging of texts and the two POS taggers developed for Ukrainian are not easily available or usable (the UGtag tagger [12] does not perform the syntactic and morphological disambiguation, and a module for the TNT POS tagger [3] is difficult to obtain), it remains difficult to use such tools for the pre-processing of corpora in order to prepare traditional terminology acquisition process. As for the utilization of terminologies, we can cite for instance localization of tools [25] and indexing of a language therapy terminology [36].

As we indicated above, we propose to take advantage of the advanced research work done in languages such as English or French, and to transpose it on lower-resourced languages using cross-lingual transfer methodology. We work with medical data and in three languages: Ukrainian, French, and English.

The general rationale of our approach is the following. Our work is based on exploitation of two kinds of corpora: Wikipedia in Ukrainian which provides several useful kinds of information (such as term labels and their codes) with a high level of quality, and parallel corpus *MedlinePlus*. Each corpus is exploited 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.

3 Material

3.1 Corpora

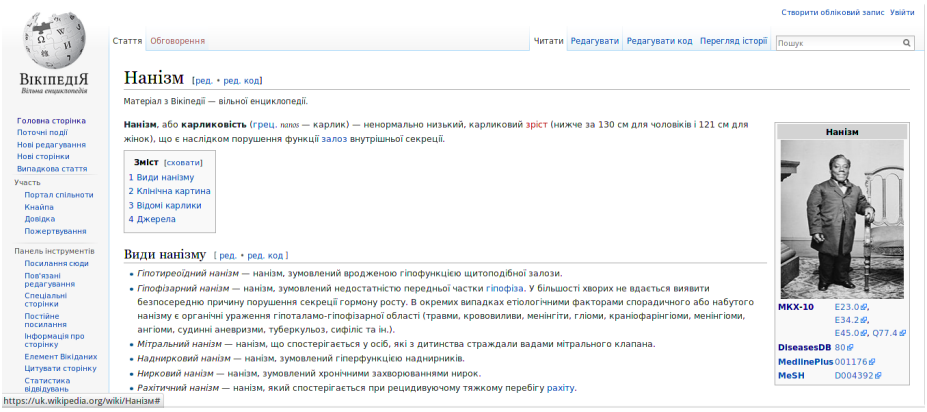

Fig. 1. Example of Ukrainian Wikipedia source page (for *Dwarfism*). The infobox with the coding is on the right.

We use two kinds of corpora:

– *MedlinePlus:* parallel medical corpus from MedlinePlus[3]. The source data are built by MedlinePlus from the National Library of Medicine. They contain patient-oriented 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. In Figure 2, we present an excerpt from this corpus;

[3] *www.nlm.nih.gov/medlineplus/healthtopics.html*

– *Wikipedia:* medicine-related articles from Wikipedia. This corpus is extracted from the Ukrainian part of 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 source page composed of three parts: the navigation frame on the left, the text and explanations in the center, and the infobox with illustrations and codings 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 explain it in Section 4.

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

Table 1. Size of the exploited corpora (Ukrainian=UK, French=FR, English=EN).

3.2 UMLS: Unified Medical Language System

The UMLS (Unified Medical Language System) [14] merges over 100 biomedical terminologies, such as international terminologies MeSH [19] and ICD [4]. Such 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 identifier, which allows finding the corresponding terms in other terminologies or languages.

3.3 Stopwords in Ukrainian

We use a list with 385 stopwords in Ukrainian (*на (on)*, *або (or)*, etc.) issued from an existing resource dedicated to the localisation of graphical interfaces[4].

4 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 4.1) and the *Wikipedia* corpus (section 4.2). We then present their cross-fertilization (Section 4.3), and evaluation of the results (Section 4.4).

[4] https://github.com/fluxbb/langs/blob/master/Ukrainian/stopwords.txt

4.1 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: (1) the source pdf documents are converted in text format; (2) in each language, documents are segmented in paragraphs; (3) French/Ukrainian and English/Ukrainian alignments are generated, in which n_{th} paragraph from one language is put in front of the n_{th} paragraph from the other language; (4) alignments within pairs of languages are then verified manually. In Figure 2, we present an excerpt from the English/Ukrainian aligned corpus.

English	*Ukrainian*
Cancer cells grow and divide more quickly than **healthy cells**. **Cancer treatments** are made to work on these **fast growing cells**.	**Ракові клітини** ростуть і діляться швидше, ніж **здорові клітини**. 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 paragraph-aligned MedlinePlus corpus (English/Ukrainian).

In French and English, we can use terminology extraction tools for bootstraping the acquisition of terminology. Hence, we use the YaTeAterm extractor [2], that is applied to POS-tagged documents (Treetagger [23], accompanied by Flemm morphological analyzer [18] in French). On the left of Figure 2, we show in bold the extracted candidate terms in English. Then, exploitation of the *MedlinePlus* parallel and aligned corpus is performed in several ways (Figure 3).

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

Transfer 2. When paragraphs contain complex expressions or sentences, the processing is done as follows (Figure 3):

1. paragraph-aligned corpora are aligned at the word level using GIZA++ [20],
2. in each paragraph pair (French/Ukrainian and English/Ukrainian) of the word-aligned corpora, terms recognized in French and English are transfered on Ukrainian paragraph;
3. extracted alignments are recorded as candidates for building the 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*

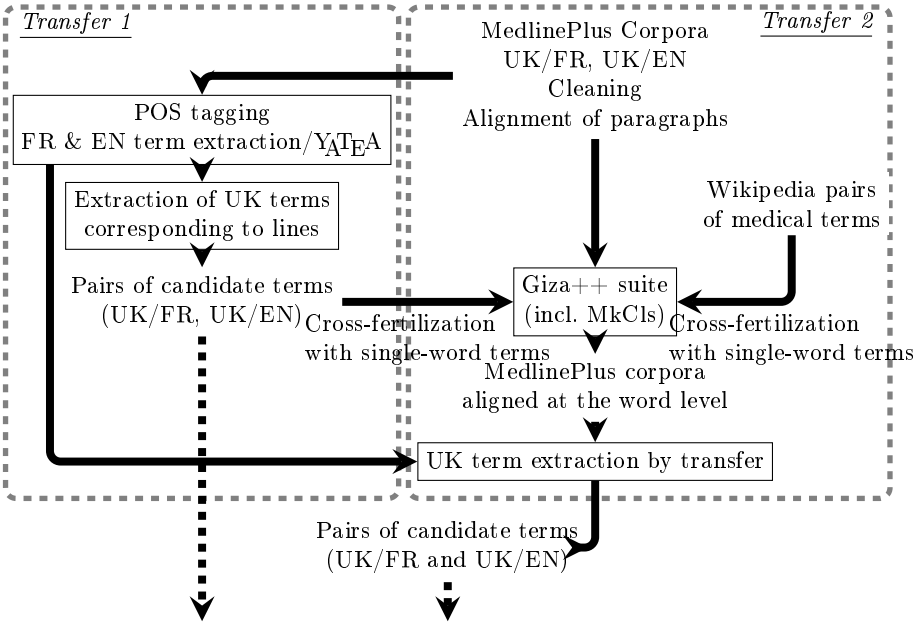

Fig. 3. Extraction of medical terms from the *MedlinePlus* corpora (Ukrainian=UK, French=FR, English=EN).

is the translation of *Ракові клітини*. This processing is performed on two pairs of languages (French/Ukrainian and English/Ukrainian).

As indicated in Table 1, our corpora are rather small for statistical alignment performed by GIZA++. For this reason, we provide GIZA++ with a bilingual dictionary in order to help alignment at the word level. Besides, we remove term pairs in which Ukrainian unit, corresponding to stopwords, is aligned as candidate term with French or English terms.

4.2 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 exploited content is extracted from infoboxes (on the right in Figure 1) and is reachable through the MediaWiki source code of Wikipedia. This provides labels of medical terms in Ukrainian and their MeSH codes. The process is the following (Figure 4):

1. the infobox content is parsed[5] which provides label and MeSH code of terms,
2. the MeSH code is used to query the UMLS, and to get the corresponding French and English terms,

[5] `Text::MediawikiFormat` (http://search.cpan.org/szabgab/Text-MediawikiFormat)

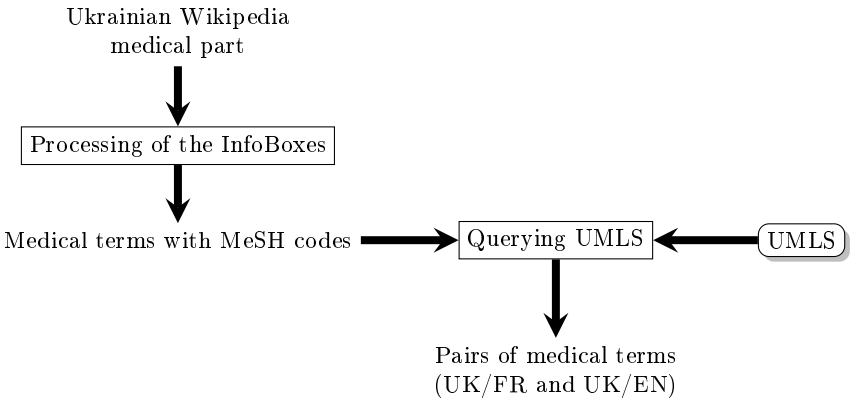

Fig. 4. Extraction of medical terms from Wikipedia (Ukrainian=UK, French=FR, English=EN).

3. the term pairs French/Ukrainian and English/Ukrainian are associated and provide good candidates for the bilingual terminology building.

For instance, in Figure 1, the term *нанізм* is extracted, as well as its MeSH code D004392. In 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 building of medical terminology in the Arabic language [28]. This part of the method exploits specific and intentionally created content and provides reliable information.

4.3 Cross-fertilization

Two kinds of cross-fertilization between the two methods (Sections 4.1 and 4.2) are performed:

- Wikipedia terms are used to enrich the extracted terminology,
- single-word terms extracted from *Wikipedia* and/or the Transfer 1 method in *MedlinePlus* are provided to GIZA++, as additional bilingual dictionary, in order to help the alignment of *MedlinePlus* at the word level. In our best setting, both resources of single-word term pairs are used.

4.4 Evaluation

Evaluation is handled manually in order to check whether candidate terms extracted for building bilingual terminologies are correct. The evaluation is performed by a native Ukrainian speaker, non expert in medicine but with several years experience in medical informatics. Terms are validated independently in each language. Besides, bilingual and trilingual relations between the Ukrainian, English and French terms are also evaluated. Thus, precision of results can be computed, *i.e.* the ratio between correct answers and all the answers.

5 Results and Discussion

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

Source	*UK*		*FR*		*EN*	
	#terms	Prec.	#terms	Prec.	#terms	Prec.
Wikipedia	357	1	1,428	1	3,625	1
$MedlinePlus_{Transfer1}$	436	0.966	316	0.971	354	0.989
inexact match		0.998		0.987		0.997
$MedlinePlus_{Transfer2}$	9,040	0.454	3,671	0.674	3,597	0.761
inexact match		0.840		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. Evaluation 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.
Wikipedia	1,515	1	3,789	1	28,840	1	28,840	1
$MedlinePlus_{Transfer1}$	63	0.937	115	0.965	282	0.954	460	0.954
inexact match		0.984		1		0.982		0.987
$MedlinePlus_{Transfer2}$	3,724	0.309	4,745	0.401	4,724	0.419	13,218	0.381
inexact match		0.751		0.840		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. Evaluation of pairs and triples (Ukrainian=UK, French=FR, English=EN).

5.1 Extraction of bilingual terminology from the *Wikipedia* corpus

Exploitation of Wikipedia infoboxes allows to collect 357 Ukrainian medical terms among which 177 are single-word terms. By querying the UMLS with MeSH codes, those terms are associated with 1,428 French terms (including 339 single-word terms) and 3,625 English terms (including 448 single-word terms). Higher number of French and English terms, compared to the number of Ukrainian terms, is due to synonyms proposed by MeSH. As for the pairs of terms, we obtain 1,515 Ukrainian/French and 3,789 Ukrainian/English term pairs, including, 270 and 405 pairs between single-word terms, respectively. Since each Ukrainian

term is associated with at least one French and English term, it allows to build 28,840 triples. We assume that precision of this terminology is 1 because the source information is created manually and is highly reliable.

5.2 Extraction of bilingual terminology from the *MedlinePlus* corpus

Use of the Transfer 1 method allows to extract 436 Ukrainian terms with a high precision (0.966). These terms are associated with 316 French and 354 English terms, within 282 triples between Ukrainian, French and English terms, with 0.954, 0.937 and 0.965 precision, respectively. Within this set, 63 pairs exist only between Ukrainian and French terms, and 115 pairs only between Ukrainian and English terms. Then, the Transfer 2 method allows to collect 334 Ukrainian/French term pairs (including 108 pairs with single-word terms) and 380 Ukrainian/English term pairs (including 135 pairs with single-word terms). We observe that these relations can involve synonyms in either language: {*фаллопієва труба*, *trompes de fallope/trompe utŕine*} (*fallopian tube*), {*втрата слуху/втрачається слух*, *hearing loss*}, {*втома*, *fatigue/tiredness*}. Besides, in Ukrainian, several inflectional forms can be associated to a given English or French form: {*вагітність*, *pregnancy*} and {*вагітності*, *pregnancy*}.

As the precision values suggest, the Transfer 1 method generates few errors. Their analysis indicates that they are mainly due to partial or non-literal translations: {*появу 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: (1) Again translation can prevent the Transfer 1 method from extracting terms in French or English. For instance, *Soins* is the French translation of *Your care*, which gives simple term in one language and complex term in another language and makes impossible their extraction by the Transfer 1 method. The Transfer 2 method will solve this problem; (2) However, the main reason of the silence is the incapacity of the term extractor to identify French or English terms because of errors in the POS tagging.

As for the Transfer 2 method, we present the results obtained when the pairs of single-word terms from the *MedlinePlus* corpus and *Wikipedia* are used to help the GIZA++ alignment. Then, the Transfer 2 method allows to extract 9,040 Ukrainian terms with 0.454 precision (exact match). Precision of French and English terms is higher: 0.674 and 0.761 respectively (exact match). Moreover, number of French and English terms is dramatically lower (about -45% and -40%) than in Ukrainian: the rich morphology of Ukrainian provides indeed several inflected forms for a given term ({*напад*, *нападу*} – *attack*, {*припадків*, *припадки*} – *seizure*, {*костей*, *кістки*} – *bones*). Besides, we can also extract synonymous terms ({*приступам*, *припадків*} – *attacks/seizures*, {*биття*, *удару*} – *beats*). Precision of inexact match, when the correct answer is included or includes the candidate term, is much higher and gains up to 0.40 points in Ukrainian and 0.05 in French and English. We assume this difference on Ukrainian is mainly due to the improvement of 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 methods. Otherwise, the errors occur.

As we indicated, the *MedlinePlus* corpus contains patient-oriented brochures, which are not highly specialized. Yet, most of the extracted terms are specific to the medical domain ({*трахеотомією*, *tracheostomy*}), {*фактори ризику*, *risk factors*}, {*шприца*, *syringe*}, {*холестерину*, *cholesterol*}). The extracted terms can also refer to close and approximating notions which reflect this type of documents: {*діти*, *children*}, {*здорову їжу*, *healthy diet*}, {*серцевий напад*, *heart attack*}, {*склянок рідини*, *glasses of liquid*}. Another 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 відчуете жодного болю*, *indolore (painless)*} (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.

6 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 allow extracting candidate terms and creating Ukrainian/French and Ukrainian/English term pairs. In particularly, French and English corpora are processed with NLP and term extraction tools. Then, thanks to the transfer methods these terms are transposed on Ukrainian. 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 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, which is currently a major source of errors. In that respect, we plan to investigate the use of other alignment algorithms, such as FastAlign [8] or the Lingua::Align toolbox [27].

Acknowledgments

projet TH

References

1. Alieva, V.: Onomatopoeic words in the Crimean Tatar language. Uchenye zapiski 18(57), 8–11 (2005)
2. Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: FinTAL 2006. pp. 380–387. No. 4139 in LNAI, Springer (2006)
3. Babych, S., Eberle, K., Babych, B.: Development of hybrid machine translation systems for under-resourced languages: Automated creation of lexical and morphological resources for mt. In: Applied and Literary Translation and Interpreting: Theory, Methodology, Practice. p. 5pp. Kyiv, Ukraine (April 2013)
4. Brämer, G.: International statistical classification of diseases and related health problems. tenth revision. World Health Stat Q 41(1), 32–6 (1988)
5. Cabré, M., Estopà, R., Vivaldi, J.: Automatic term detection: a review of current systems, pp. 53–88. John Benjamins (2001)
6. Delač, D., Krleža, Z., Dalbelo Bašić, B., Šnajder, J., Šarić, F.: TermeX: A tool for collocation extraction. In: CICLing, LNCS 5449. pp. 149–157 (2009)
7. Dmytruk, V.: Typological features of word-formation in computing, the internet and programming in the first decade of the XXI century. In: УДК. pp. 1–11 (2009)
8. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of ibm model 2. In: NAACL/HLT. pp. 644–648 (2013)
9. Grigonyte, G., Rimkute, E., Utka, A., Boizou, L.: Experiments on Lithuanian term extraction. In: NODALIDA 2011. pp. 82–89 (2011)
10. Kageura, K., Umino, B.: Methods of automatic term recognition. In: National Center for Science Information Systems. pp. 1–22 (1996)
11. 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)
12. 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)
13. Kruglevskis, V., Vancane, I.: Term extraction from legal texts in latvian. In: Second Baltic Conference on Human Language Technologies (2005)
14. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods Inf Med 32(4), 281–291 (1993)
15. 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)
16. McDonald, R., Petrov, S., Hall, K.: Multi-source transfer of delexicalized dependency parsers. In: EMNLP (2011)
17. Memetova, E.: Lexicophraseological expressive means of the Crimean Tatar language. Uchenye zapiski 18(57), 37–39 (2007)
18. Namer, F.: FLEMM : un analyseur flexionnel du français à base de règles. Traitement automatique des langues (TAL) 41(2), 523–547 (2000)
19. National Library of Medicine, Bethesda, Maryland: Medical Subject Headings (2001), *www.nlm.nih.gov/mesh/meshhome.html*
20. Och, F., Ney, H.: Improved statistical alignment models. In: ACL. pp. 440–447 (2000)
21. Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.: Terminology extraction: An analysis of linguistic and statistical approaches. In: Sirmakessis, S. (ed.) Knowledge Mining, Studies in Fuzziness and Soft Computing, vol. 185, pp. 255–279. Springer Berlin Heidelberg (2005)

22. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostay, T.: Term extraction, tagging, and mapping tools for under-resourced languages. In: TKE 2012. pp. 193–208 (2012)
23. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. pp. 44–49 (1994)
24. Shatalina, O.: Literature terminology of the Old Ukrainian literature of the 18th century. Uchenye zapiski 18(57), 5–7 (2005)
25. Shyshkina, N., Zorko, G., Lesko, L.: Terminology work and software localization in Ukraine. In: Problems of Cybernetics and Informatics. pp. 17–20 (2010)
26. Tadić, M., Šojat, K.: Finding multiword term candidates in Croatian. In: IESL Workshop, RANLP Conference. pp. 102–107 (2003)
27. Tiedemann, J., Kotzé, G.: A discriminative approach to tree alignment. In: Ilisei, I., Pekar, V., Bernardini, S. (eds.) International Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography and Language Learning. pp. 33–39 (2009)
28. Vivaldi, J., Rodríguez, H.: Arabic medical term compilation from Wikipedia. In: Proc of CIST 2014 (2014)
29. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: HLT (2001)
30. Zeman, D., Resnik, P.: Cross-language parser adaptation between related languages. In: NLP for Less Privileged Languages (2008)
31. Бугаков, О.: Создание семантического словаря предложных конструкций на основе украинского национального лингвистического корпуса. Tech. rep., Украинский языково-информационный фонд НАН Украины, Киев, Украина (2006)
32. Глибовец, А., Решетнев, І.: Метод итеративного построения терминологии в коллекциях научных текстов на украинском языке. Кибернетика и системний анализ 50(6), 53–62 (2014)
33. Демська, О.: Текстовий корпус: ідея іншої форми. ВПЦ НаУКМА, Київ, Україна (2011)
34. Коссак, О.: Українська комп'ютерна термінологія. In: Сучасні проблеми в комп'ютерних науках. pp. 39–42 (2000)
35. Кочерга, О., Мейнарович, Е.: Англійсько-Українсько-Англійський словник наукової мови. Фізика та споріднені науки. Нова книга, Вінниця, Україна (2010)
36. Лалаєва, Р., Сурованець, [U+FFFD], Тищенко, В.: Індексація польсько-, російсько- та українськомовної логопедичної термінології. Лексикографічний бюлетень 10, 29–36 (2004)
37. Левченко, О., Кульчицький, І.: Технологія перетворення п'ятимовного словника порівнянь в електронну форму. In: Інформаційні системи та мережі. pp. 129–138 (2013)
38. Монахова, Т.: Застосування прийомів корпусної лінгвістики в лексикографії. Наукові праці 98(85), 55–60 (2009)
39. Пуряєва, Н.: Дещо про мову богослужіння взагалі та про словник мови богослужіння зокрема. Лексикографічний бюлетень 10, 36–42 (2004)
40. Тименко, [U+FFFD]: Лексико-тематичні групи української юридичної термінології початку xx століття. Лексикографічний бюлетень 10, 65–70 (2004)