Mariana González · Silvia Susana Reyes ·
Andrea Rodrigo · Max Silberztein (Eds.)

Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities

16th International Conference, NooJ 2022
Rosario, Argentina, June 14–16, 2022
Revised Selected Papers

Springer

Editors
Mariana González
Universidad Nacional de Rosario
Rosario, Argentina

Andrea Rodrigo
Universidad Nacional de Rosario
Rosario, Argentina

Silvia Susana Reyes
Universidad Nacional de Rosario
Rosario, Argentina

Max Silberztein ⓘ
Université de Franche-Comté
Besancon, France

Natural Language Processing Applications

Author Index

Creation of Parallel Medical and Social Domains Corpora for the Machine Translation and Speech Synthesis Systems

Mikita Suprunchuk[1]([⊠]), Nastassia Yarash[2], Yuras Hetsevich[2], Valery Varanovich[3], Siarhey Gaidurau[2], Yauheniya Zianouka[2], and Palina Sakava[2]

[1] Minsk State Linguistic University, Zakharov Str. 21, Minsk, Belarus
ssrlab221@gmail.com
[2] United Institute of Informatics Problems, Surhanava Str. 6, Minsk, Belarus
[3] Belarusian State University, Nezavisimosti Av. 4, Minsk, Belarus

Abstract. This paper represents the procedure of creating medical and social domains corpora in NooJ. It illustrates a primary analysis of the corpora. Based on the research, the Belarusian NooJ module has been supplemented and thematic dictionaries of medical and social areas in NooJ format have been prepared. Automatic text processing emphasized words' linguistic peculiarities (mainly morphological) that the main module of Belarusian language for NooJ did not recognize. In addition, the article shows how the corpora help to improve the quality of translation, to identify heterogeneous translation options.

Keywords: Automatic language processing · Corpus of a medical domain · Corpus of a social domain · Dictionary · NooJ · Concordance · Belarusian · Russian · English

1 First Section Introduction

In the Republic of Belarus, Belarusian and Russian have the status of official languages. Since there is bilingualism at the legislative level, it is necessary to provide the availability of texts for different purposes in both official languages. O7ne of the important purposes is public education in medical and social fields.

Medical and social texts contain essential information. The first are intended to convey information between health professionals and scientists (for example, in articles, research papers) or to educate and inform the population about public and personal health. Meanwhile, the second group represents the historical and cultural heritage of Belarus and is aimed at acquainting country visitors with it.

Therefore, the need for a system of machine translation and speech synthesis in Belarusian and Russian is still relevant. Based on the above, there is a data layer of mentioned domains poorly covered with such systems, which, in turn, require large parallel corpora for their training [3, 7]. At the moment, there are no large parallel corpora, which include Belarusian, on open access. As a result, the first step in the machine translation and speech synthesis development may be the creation of

corresponding corpora (cf., for example, [5, 8]. Using the existing proofreading and word processing services and NooJ corpora processor [6], it is possible to work with big data to further use it to train machine translation systems and the speech synthesis systems as well. Thus, the access to texts of medical and social domains in both official languages in Belarus and English, a language of global communication, will be obtained by a significant number of people.

Currently, Speech Synthesis and Recognition Laboratory of the United Institute of Informatics Problems (further on—Laboratory; https://ssrlab.by/en/) works at creating corpora of medical and social domains. At this stage, a medical corpus consisting of 848 parallel texts in Belarusian (303,469 word forms), English (373,709 word forms) and Russian (330,996 word forms) has been compiled. In addition, a corresponding parallel corpus of social domain based on parallel texts in Belarusian, English and Russian from the KrokApp travel audio guide [4] was created (cf. Table 1).

2 Compilation of a Medical and Social Domain Corpus

For the corpora of medical and social domains, original texts from Internet resources and online programs were collected. First of all, materials provided on the websites of polyclinics and hospitals in Minsk. They are:

- 1st Central district clinical polyclinic of the Central district of Minsk, http://www.1crp.by/en/home (Fig. 1);
- 4th City clinical hospital named after Mikalaj Saučanka, https://4gkb.by;
- Health Department of the Minsk City Executive Committee, https://komzdrav-minsk.gov.by/en/about-committee.
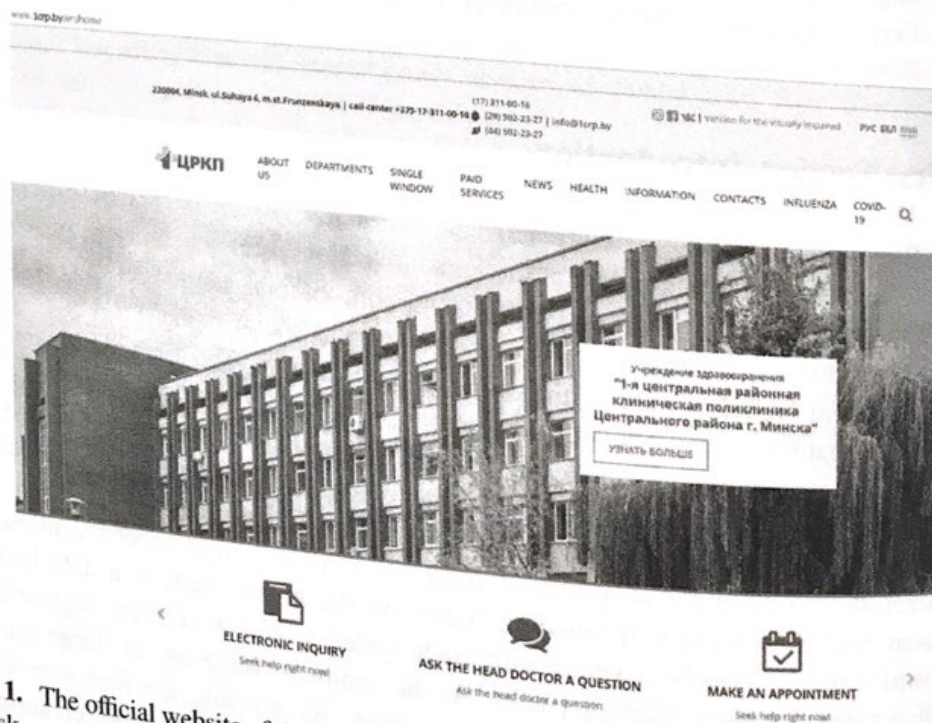

Fig. 1. The official website of the 1st Central District clinical polyclinic of the Central District of Minsk.

Also the texts from websites and mobile applications KrokApp (https://krokapp. com) and Krokam (https://krokam.com) for tourists and local historians were added [4]. The Laboratory staff work on the content of enumerated medical and travel web-sites, translate articles, news and maintain them. We received and used these materials for the corpora: current news, instructions, rules, and other information from medical institutions, and essays on the history and interesting objects in different settlements of Belarus from museums and the Institute of History of the Academy of Sciences. Most of the texts were in Russian, some in Belarusian. The texts were translated into Belarusian and English and uploaded to the appropriate platforms. Naturally, the translations were checked by the editor and the person responsible for the project. The medical corpus of 303,469 word forms and history and social corpus of 61,827 word forms were translated.

Figure 2 shows a fragment of parallel texts in Belarusian, English, Russian from the corpus of social texts.

[BE]У горадзе Лунінцы ў 1911–1912 гадах жыў і працаваў Якуб Колас – знакаміты беларускі пісьменнік, паэт і перакладчык. Адзін з класікаў і заснавальнікаў новай беларускай літаратуры. Пра гэта нагадвае турыстам і падарожнікам мемарыяльная дошка, усталяваная на будынку мясцовага краязнаўчага музея. Якуб Колас(Канстанцін Міхайлавіч Міцкевіч) нарадзіўся ў 1882 годзе ў вёсцы Акінчыцы на Стаўбцоўшчыне ў сямі лесніка. У 1902 годзе ён скончыў Нясвіжскую настаўніцкую семінарыю. У 19021906 гадах працаваў настаўнікам на Палессі і на Міншчыне. За ўдзел у нелегальным настаўніцкім зездзе быў звольнены з працы, а ў 1908 годзе асуджаны на тры гады турэмнага зняволення.

[EN]Jakub Kolas, a famous Belarusian writer, poet and translator, lived and worked in Luniniec in 1911–1912. Jakub Kolas is one of the classics and founders of the new Belarusian literature. A memorial plaque installed on the building of the museum of local lore reminds tourists and travelers about this. Jakub Kolas (Kanstancin Michajlavič Mickievič) was born in 1882 in the village of Akinčycy in the Stoŭbcy Region in a family of a forester. In 1902, he graduated from the Niasviž Teacher Seminary. In 19021906 he worked as a teacher in Paliessie and in the Minsk Region. He was dismissed from work for participation in an illegal teachers congress, and in 1908 convicted to three years in prison.

[RU]В городе Лунинец в 1911–1912 годах жил и работал Якуб Колас – знаменитый белорусский писатель, поэт и переводчик. Один из классиков и основателей новой белорусской литературы. Про это напоминает туристам и путешественникам мемориальная доска, установленная на здании местного краеведческого музея. Якуб Колас (Константин Михайлович Мицкевич) родился в 1882 году в деревне Акинчицы на Столбцовщине в семье лесника. В 1902 году он закончил Несвижскую учительскую семинарию. В 19021906 годах работал учителем на Полесье и на Минщине. За участие в нелегальном учительском съезде был уволен с работы, а в 1908 году осуждён на три года тюремного заключения.

Fig. 2. Parallel texts in Belarusian, English, Russian from the corpus of social texts.

The texts described above have become the basis for the creation of the main corpora. Manual text proofreading and correction of errors was accomplished sequentially by some editors, qualified linguists and historians are among them. Before uploading them to NooJ, the following items were additionally checked by means of automatic services provided on the portal *corpus.by* [1]:

– spelling checking,
– analysis of statistical and supplemental information about all symbols in texts,
– analysis of statistics on the use of arbitrary symbolic sequences in an electronic text (counting the frequency of word forms),
– search and correction of errors in the spelling of the Belarusian letters "у" and "ў",
– recognition and selection of homographs in the text.

Statistical data on the created corpora in NooJ is represented in Table 1.

Table 1. The statistics of medical and social corpora.

Unit	Medical corpus			Social corpus		
	BY	RU	EN	BY	RU	EN
Texts	848	848	848	386	386	386
Tokens	419,165	457,756	481,089	81,481	80,390	101,036
Word forms	303,469	330,996	373,709	61,827	61,085	81,651
Text units delimited by "\n"	32,969	37,398	32,952	4,168	4,172	4,082

The main linguistic peculiarities of the medical and social corpora are:

– There are more complex sentences than simple sentences.
– The noun is the most frequent part of speech.
– Some words are specific for this field, especially in medical texts. They should be added to the dictionary of NooJ. So, it was decided to compose and process the list of medical and social terms.

Slavic languages are flectional synthetic ones (except Bulgarian). Therefore, morphological features require great attention in the development of computer analyzers.

While creating the dictionary of corpora, a list of unknown words was defined. These are words that were not identified by the general_be dictionary of the Belarusian module [2] in NooJ). The list of unknown words was processed manually (Fig. 3). It includes these groups:

– Errors in the original texts. They are misprints, incorrect translations, missing spaces, Latin characters in the Belarusian or Russian texts, etc.;
– 2277 specific medical terms: *дыясталічны* 'diastolic', *колератэрапія* 'color therapy';
– 383 social terms: *лютаранскі* 'Lutheran', *плябанія* 'parsonage';
– Words and word forms that are not included in the Belarusian dictionary general_be.nod: *квартал* 'quarter', *заплечнік* 'backpack';
– English words and symbols: *Sunday, view, html*;
– Roman numerals: *IV, XI, XV*;
– Simple mistakes, punctuation or orthography: *магістарата (магістрата), адзаначыць (адзначыць), вядзяляецца (выдзяляецца)*;

– Words with an apostrophe (the apostrophe divides words into parts): *сур'ёзна*, *аб'ект*. Sometimes the apostrophe is a functional symbol, but in the Belarusian language it has the role of a letter, it does not separate words, it is a part of words;
– Words for which it was not possible to find an appropriate paradigm. For individual words and their categories, it was necessary to supplement the general dictionary of the Belarusian module and add new samples. These are words such as *адзёр* 'measles', *алеін* 'olein', *біскуп* 'bishop', *ваксігрып* 'Vaxigripp', *месціч* 'citizen', *пробашч* 'catholic priest', etc.

Fig. 3. The list of unknown words in NooJ corpora.

English words, symbols and Roman numerals were omitted as these words should be processed separately in another research. Almost all words with an apostrophe were included in the list of general vocabulary.

3 Creating Medical and Social Trilingual Dictionaries in NooJ

Text corpora are a valuable knowledge source about language peculiarities. They allow monitoring of facts and tendencies that are difficult to notice by common speaking or reading. Lexical and grammatical elements of the Belarusian module in NooJ were used for the lexical and grammar marking of the medical and social corpus which was created by the specialists of the laboratory.

Further work on the dictionaries of medical and social terms consisted in the morphological markup of words selected from the medical corpus. The Paradigm Generator service on the corpus.by platform was used for this purpose [1]. In the processing window you need to enter a word (or several forms) with a part of the language. The Word Paradigm Generator outputs several variants of annotations in NooJ format and a user can choose the right word form with its paradigm. The resulting paradigms have an accent, indicated by the "+" symbol, and special tags separated from the word by an underscore "_". This is semi-automatic processing of unknown words. Thus, each initial form was assigned a morphological class that shows the word change of the lexeme.

The tool is free and available on Computational Platform for Electronic Text and Speech Processing corpus.by (Fig. 4) [1].

```
герыятр,NOUN
герыятра,NOUN
герыятрам,NOUN
```

○ Processing according to wordforms dictionary
◉ Processing according to dictionary of inflections in NooJ format

Generate probable paradigms!

Result

Paradigms are based on 3 forms (Total 72):

```
герыятр,NOUN+FLX=IHЖЫР
герыятр/Accusative+Common+Inanimate+Masculine
герыятр/Common+Inanimate+Masculine+Nominative
герыятра/Common+Genitive+Inanimate+Masculine
герыятрам/Common+Inanimate+Instrumental+Masculine
герыятру/Common+Dative+Inanimate+Masculine
герыятры/Common+Inanimate+Masculine+Prepositional;
```

Fig. 4. An interface of The Word Paradigm Generator service.

Due to unknown reasons, there is no sample in the general_be dictionary for these word groups:

- Feminine words with the suffix *-асц-*: *самотнасць* 'loneliness', *забудаванасць* 'built - up area', *двухграннасць* 'dihedral', *радасць* 'joy', etc.;
- Masculine singularia tantum common ending with *- інг*: *дайвінг* 'diving', *ліфтынг* 'lifting', *маркетынг* 'marketing'.

These shortcomings are to be corrected.

So, specialized dictionaries (additional for general_be) were created. They are designed for medical (Fig. 5) (*анальгетык, ангіёграф, глюкометр, дыстрэс* 'analgesic, angiography, glucose meter, distress', etc.) or social (*арханёл, біягрупа, бруcaванка, дамініканец* 'archangel, biogroup, paving stone, Dominican', etc.) domains. A number of words with paradigms have also been prepared for inclusion in the general dictionary of the Belarusian NooJ module, for example: *агульнасусветны, адсканаваць, бескантактны, аэрагрыль, ірвота* 'worldwide, scan up, contactless, aerogrill, vomiting', etc.

```
# NooJ V7
# Dictionary
#
# Language is: be
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' description files (.nof), e.g.:
# Special Command: #use paradigms.nof
#
# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB (unambi
#                   +FLX= (inflectional paradigm) +DRV= (derivational paradigm)
#
#
# Special Characters: '\' '"' ' ' ',' '+' '-' '#'
#
герыятр,NOUN+FLX=АВАР
адрэнаблакатар,NOUN+FLX=ІНЖЫР
адыктыўны,ADJECTIVE+FLX=ААЗІСНЫ
акарыцыдны,ADJECTIVE+FLX=ААЗІСНЫ
акрацыяноз,NOUN+FLX=ААГЕНЕЗ
актываваны ,ADJECTIVE+FLX=ААЗІСНЫ
акустыкафобія,NOUN+FLX=ААГАМІЯ
алігаартрыт,NOUN+FLX=ІЎРЫТ
атачаць,VERB+FLX=ЗАВЯШЧАЦЬ
атрафічны,ADJECTIVE+FLX=ААЗІСНЫ
атэрагеннасць,NOUN+FLX=ЗНАКАМІТАСЦЬ
аўтадонарства,NOUN+FLX=АВАВЯЗАЦЕЛЬСТВА
аўтыст,NOUN+FLX=АВАЛІЦЫЯНІСТ
аўтыстычны,ADJECTIVE+FLX=ААЗІСНЫ
бедаквілін,NOUN+FLX=АЛЕІН
біятэрапія NOUN+FLX=ААГАМІЯ
```

Fig. 5. Belarusian medical dictionary in NooJ format.

A trilingual dictionary is a perspective phenomenon in the field of Belarusian diplomacy, modern foreign policy, and international relations. The presence of a trilingual dictionary in everyday practice makes it possible to constantly replenish and adjust the lexical stock in three languages. The knowledge of Belarusian and Russian languages, and the spread of English as the "language of international integration" contributes to the creation of extensive interaction between the systems of the three languages in communication. This makes it necessary to compile trilingual dictionaries, especially of medical and social terminology in terms of this research.

All terms of the Belarusian medical and social dictionaries were correlated with their Russian equivalents from the parallel corpora of medical and social domains. The next step was their translation into English which was done by linguists manually. The specialists faced the following difficulties and noticed the following features of trilingual translation into English:

– some of the Belarusian-Russian terms can have more than one terminological equivalent with different meanings in the English language:

абязбольвальны,   ADJECTIVE+TRANS+RU =   "обезболивающий"+TRANS +EN = "anesthetic";

абязбольвальны,   ADJECTIVE+TRANS+RU =   "обезболивающий"+TRANS +EN = "analgesic";

– there is no exact correspondence to a certain term in English, the translation can only be descriptive (in a few words):

вірусаносьбіцтва,   NOUN+TRANS+RU =   "вирусоносительство"+TRANS +EN = "carriage_of_viruses";

маламабільны, ADJECTIVE+TRANS+RU = "маломобильный"+TRANS+EN = "with_limited_mobility".

Trilingual dictionaries can be composed in NooJ format by special command "TRANS+RU" for the Russian equivalent and "TRANS+EN" for English. The whole line of one translated word looks like "шпіталізацыя, NOUN+TRANS+RU = "госпитализация"+TRANS+EN = "hospitalization". In this way, all the terms of the medical and social domain were processed manually by translators and then compiled into NooJ format. The medical dictionary contains 642 words, the social dictionary contains 675 words.

4 NooJ for Proofreading and Translating

4.1 Common Names in the Corpora

Concordances of articles were prepared in the NooJ platform. They made it possible to check uniformity of translation of the same word, to find out the use of synonyms, homonyms, uniformity of names of persons, cities, events. We were able to detect some errors in translations, unify texts.

Here are some examples of how to check compliance during such work. Some interesting cases are connected with the choice of synonyms. For example, the usage of words *city* and *town* and related (Table 2).

Thanks to the work on the corpora and concordance, errors were identified in the translation of this group of words. For example:

The most beautiful manor house in a classicism style with a landscape park became a true decoration of the small city. – Найпрыгажэйшы сядзібны дом у стылі кла- сіцызму з пейзажным паркам стаў сапраўдным упрыгажэннем невялікага гарадка. (Najpryhažejšy siadzibny dom u styli klasicyzmu z piejzažnym parkam stаŭ sapraйdnym upryhaženniem nievialikaha haradka).

Here are a few examples that have been found thanks to the NooJ analysis. Thus, the collocation *small city* seems out of place. The word *гарадка* should be translated as *small town*, not *city*.

Or consider the following specific context in the description of *Narowlia*:

У 1840 годзе у сувязі з частым разлівам ракі царква была перанесена на край ракі **мястэчка** *(цяпер гэта цэнтр* **горада***). (U 1840 hodzie u suviazi z častym razlivam raki carkva byla pieraniesiena na kraj* **miastečka** *(ciapier heta centr* **horada***).*

In eighteen forty the church was transferred to border of the **town** *(now its the centre of the* **city***).*

Table 2. Translations of *town, city,* etc.

Part of speech	Be	Ru	En
Noun	горад 266 мястэчка 11 (old type of a small urban settlement) гарадок 2 (diminutive)	город 258 местечко 5 (old type of a small urban settlement) городок 9 (diminutive)	town 43 city 384
Adjective	гарадскі 111	городской 108	
Total	390	380	423

The names of the inhabitants of these localities are connected with the words *horad, miastečka, vioska, town, city.* To translate those who live in the city, the words *citizen* and *resident* were used. There was also the word *people.* We analyzed theirs' distribution: *haradžanin* 18, *miescič* 5, *žychar* 18, *citizens* 19, *residents* 17, *people* 5. Here are some examples of variegated distribution of these words:

- *абагульненых вобразах* **жыхароў** *горада* (abahulnienych vobrazach žycharoŭ horada) – 'generalized images of the **citizens** of the city';
- *Тут праводзіў свае пасяджэнні выбраны* **гараджанамі** *орган самакіравання - магістрат.* (Tut pravodziŭ svaje pasiadženni vybrany haradžanami orhan samakiravannia - mahistrat.) – 'City Council, which was elected by the **people**, held meetings there'.
- *Парк імя Першага мая ў горадзе Брэсце з'яўляецца не толькі самым папулярным сярод* **гараджан** *і турыстаў...(Park imia Pieršaha maja ŭ horadzie Brescie zjaŭliajecca nie toĺki samym papuliarnym siarod haradžan i turystaŭ...)* – 'The First of May Park in the city of Brest is not only the most popular park among the **residents** and tourists...'
- *На гэтай вуліцы пражывалі заможныя* **месцічы**-*яўрэі.* (Na hetaj vulicy pražyvali zamožnyja miesciču-jaŭrei.) – 'Wealthy **citizens**-Jews lived on this street'.

4.2 Personal Names in the Corpora

Belarusian toponyms have many variants in original and translated texts. Consider the example of the city of Hrodna (Table 3). It is a regional center in the west of Belarus. About 370 thousand people live there.

It appears under the names: *Гродна, Гародня, Горадня.* The following cognate adjectives are possible in the Belarusian language: *гродзенскі, гродненскі, гарадзенскі.* Using NooJ's ability to create concordances, we analyze their applicability. Belarusian and Russian adjectives are translated into English as one word – *Hrodna.*

Deviations from the main variant are associated with the name of Prince David: *Davyd Haradzienski* (1283–1326) (6 cases). In addition, one erroneous translation was

detected: *The author of the... monument to Ciotka was the famous* **Grodno** *sculptor Alieś Lipień.*

Different final figures make it possible to identify other deviations, features in translations.

Table 3. Translation variants of the city name *Гродна* (*Hrodna*).

Part of speech	Be	Ru	En
Noun	Гродна 64 Горадня 1 Гародня 0	Гродно 68	Hrodna 154 * Grodno 1 * Grodna 0
Adjective	гродзенскі 75 гарадзенскі 10 Гарадзенскі (Давід) 6	гродненский 89 Давид Гродненский 6	Davyd Haradzienski 6
Total	156	163	161

Another Belarusian city – *Mahilioŭ* – has a similar problem of variability. It is a regional center in the east of Belarus, with 380,000 inhabitants. The following results were obtained (Table 4).

Table 4. Translation variants of the city name *Магілёў* (*Mahilioŭ*).

Part of speech	Be	Ru	En
Noun	Магілёў 36	Могилёв 24 Могилев 13	Mahilioŭ 100 Mahiliou 2
Adjective	магілёўскі 66	могилевский 19 могилёвский 48	
Total	102	104	102

There is also a regular feature of the English language – nouns and adjectives homonyms. And in two cases the translator did not use a special character for the «ў» sound.

The peculiarity of the Russian language was also manifested here: the variance of the spelling "*Е/Ё*". According to actual rules, the usage of the letter *ё* is not obligatory. So, when automating processes, you need to take into account this detail.

So, while translating and proofreading, NooJ is used:

– to find synonyms;

- to achieve uniformity in the spelling of proper names;
- to accurately and uniformly translate terms and set expressions;
- to detect missing text snippets.

5   Conclusion

The creation of a specialized corpora in the NooJ processor is described in the article. New medical and social corpora contain about 1,6 million tokens. It includes medical and social texts in the Belarusian, English and Russian languages. The texts were taken from specialized websites of medical institutions and from audio guides on historical sights of Belarus. The original texts in Belarusian have been translated into Russian and English. The translation quality was improved due to the combined machine-manual approach: texts were initially translated automatically, after that they were manually proofread and edited, then automatically proofread.

A general dictionary of the Belarusian NooJ module was complemented by new words in the process of corpus preparation; specialized dictionaries of words were compiled: a Belarusian-Russian-English list of medical and social terms. The dictionary is planned to be supplemented.

So, NooJ allows us to achieve uniformity in writing, detect lexical and spelling discrepancies between the original and the translation, facilitates the correction of errors in language or content. There were also several cases when the sentence was omitted during the translation, probably due to the mistake of the translators. Using corpora helped to unify the translation of proper names, too.

References

1. Computational platform for electronic text & speech processing. http://www.corpus.by. Accessed 01 Sept 2022
2. Hetsevich, Y., Varanovich, V., Kachan, E., Reentovich, I., Lysy, S.: Semi-automatic part-of-speech annotating for belarusian dictionaries enrichment in NooJ. In: Barone, L., Monteleone, M., Silberztein, M. (eds.) NooJ 2016. CCIS, vol. 667, pp. 101–111. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55002-2_9
3. Hetsevich, Y., Kirdun, A.A.: Vykarystannie sistem mashynnaga pierakladu i sintezu mawliennia dlia zabiespiachennia dastupnastsi zakanadawchykh tekstaw na roznykh movakh u Respublitsy Bielarus. Minsk (2018). Гецэвіч, Ю.С., Кірдун, А.А.: Выкарыстанне сістэм машыннага перакладу і сінтэзу маўлення для забеспячэння даступнасці заканадаўчых тэкстаў на розных мовах у Рэспубліцы Беларусь. In: Коваленко, Е. И. (ред.). Информационные технологии и право (Правовая информатизация – 2018): матер. VI Междунар. науч.-практ. конф., 17 мая 2018 г.; с. 123–128. Минск (2018)
4. KrokApp – personal audio guide in Belarus. https://krokapp.by/about/. Accessed 01 Sept 2022
5. Michigan Corpus of Academic Spoken English. https://quod.lib.umich.edu/m/micase. Accessed 01 Sept 2022
6. NooJ: A Linguistic Development Environment. http://www.nooj4nlp.org/. Accessed 01 Sept 2022

7. Reentovich, I., et al.: The first one-million corpus for the Belarusian NooJ module. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.) NooJ 2015. CCIS, vol. 607, pp. 3–15. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42471-2_1

8. The Colorado Richly Annotated Full Text Corpus. http://bionlp-corpora.sourceforge.net/CRAFT/. Accessed 01 Sept 2021