МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ СОЦИОКУЛЬТУРНЫХ КОММУНИКАЦИЙ

ЛИНГВИСТИКА, ЛИНГВОДИДАКТИКА, ЛИНГВОКУЛЬТУРОЛОГИЯ: АКТУАЛЬНЫЕ ВОПРОСЫ И ПЕРСПЕКТИВЫ РАЗВИТИЯ

Материалы V Международной
научно-практической конференции

Минск, 18–19 марта 2021 г.

МИНСК
БГУ
2021

УДК 81(06)+811.1/.8(072)(06)+81:008(06)
ББК 81я431+81.40/59р30-2я431+81.006.3я431
Л59

Л59 **Лингвистика,** лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития : материалы V Междунар. науч.-практ. конф., Минск, 18–19 марта 2021 г. / Белорус. гос. ун-т ; редкол.: О. Г. Прохоренко (гл. ред.) [и др.]. – Минск : БГУ, 2021. – 338 с.

ISBN 978-985-811-085-6.

Рассматриваются современные направления лингвистических исследований, традиции и инновации в обучении иностранным языкам, проблемы языковой картины мира и взаимодействия культур, освещаются вопросы литературоведения, теории перевода и интерпретации текста.

УДК 81(06)+811.1/.8(072)(06)+81:008(06)
ББК 81я431+81.40/59р30-2я431+81.006.3я431

ISBN 978-985-811-085-6

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Нифагин В.А., Дубровина О.В. Инновационные аспекты руководства дипломным проектированием студентов специальности «Прикладная информатика» // Актуальные проблемы гуманитарного образования: Материалы V международной научно-практической конференции, Минск 18-19 октября 2018 г., Минск, БГУ, 2018.
2. PostgreSQL: The world's most advanced open source database [электронный ресурс]. URL: https://www.postgresql.org.
3. Spring Boot – Краткое руководство [электронный ресурс]. URL: https://coderlessons.com/tutorials/java-tekhnologii/learn-spring-boot/spring-boot-kratkoe-rukovodstvo.
4. React – JavaScript-библиотека для создания пользовательских интерфейсов [электронный ресурс]. URL: https://ru.reactjs.org.

ПРАКТЫЧНЫЯ АСПЕКТЫ СТВАРЭННЯ ПАРАЛЕЛЬНАГА БЕЛАРУСКА-РУСКАГА КОРПУСА ДАНЫХ

PRACTICAL ASPECTS OF CREATING A PARALLEL BELARUSIAN-RUSSIAN CORPUS

Ц. Пракапенка

T. Prakapenka

Аб`яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі
Мінск, Беларусь

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus
Minsk, Belarus

e-mail: tsimafei.prakapenka@gmail.com

У дадзеным артыкуле прыведзены падрабязны аналіз існуючых карпусоў тэкстаў для беларускай мовы. Таксама прадстаўлены распрацаваны аўтарам паралельны беларуска-рускі корпус даных, апісаны пакрокавы алгарытм яго стварэння і абгрунтавана яго важнасць для задачы машыннага перакладу.

Ключавыя словы: беларуская мова; паралельны корпус; руская мова; камп'ютарная лінгвістыка; машынны пераклад.

This article provides a detailed analysis of the existing text corpora for the Belarusian language. The parallel Belarusian-Russian corpus developed by the author is also presented, the step-by-step algorithm of its creation is described and its importance for the machine translation problem is substantiated.

Keywords: the Belarusian language; parallel corpus; the Russian language; computer linguistics; machine translation.

Немагчыма адмаўляць, што цяпер машынныя перакладчыкі адыгрываюць важную ролю у жыцці людзей. Амаль у любой краіне свету можна паразумецца з мясцовымі жыхарамі дзякуючы якасным аўтаматычным сэрвісам (*‘Google Translate’, ‘Яндекс.Переводчик’*). Аднак, калі звярнуцца да больш спецыфічных задач (пераклад медыцынскіх ці юрыдычных тэкстаў) для не самых распаўсюджаных моў (беларуская ў тым ліку), можна зразумець, што крыніц з адпаведнымі паралельнымі данымі не хапае. Задача машыннага перакладу патрабуе вялікай колькасці сказаў на дзвюх мовах, прычым чым іх больш, тым якаснейшай будзе мадэль. На жаль, для беларускай мовы дагэтуль няма дастаткова вялікага корпуса у свабодным доступе. Такім чынам, магчымасць правядзення даследаванняў мадэляў перакладу з беларускай на рускую моцна абмежаваная, і стварэнне новага корпусу дапаможа развіць галіну аўтаматычнага перакладу для беларускай мовы.

Спецыялісты ў галіне машыннага перакладу рэкамендуюць выкарыстоўваць як мага больш даных (звычайна каля мільёна паралельных сказаў з пераважна агульнаўжывальнай лексікай) для пабудовы якаснага машыннага перакладчыка. Аднак, з улікам спецыфікі абранай пары (руская і беларуская належаць да адной моўнай групы) можна паспрабаваць дасягнуць карэктных вынікаў на меншым наборы. Паставім за мэту каля чатырохсот тысяч сказаў.

Правядзём агляд існуючых корпусаў для беларускай мовы, каб упэўніцца ў актуальнасці пастаўленай задачы:

1. Беларускі N-корпус [1] – самы буйны на дадзены момант публічны агульны корпус беларускай мовы са структурнай і граматычнай разметкай.

2. Эксперыментальны корпус беларускай мовы [2, с. 231–238] уяўляе сабой алічбаваныя і размечаныя мастацкія творы і тэксты беларускамоўных часопісаў XX стагоддзя.

3. Corpus Albaruthenicum – корпус навуковых тэкстаў беларускай мовы, падрыхтаваны спецыялістамі БНТУ разам з навукоўцамі Інстытута мовы і літаратуры імя Якуба Коласа і Янкі Купалы НАН Беларусі.

4. Руска-беларускі корпус на старонцы Нацыянальнага корпуса рускай мовы (НКРМ) [3] – паралельны корпус, распрацаваны сумесна рускімі і беларускімі навукоўцамі. Даступны для анлайн-пошуку.

5. Статрыццаціпяцітысячны паралельны публічны корпус А. Ф. Брыля, створаны на аснове тэкстаў Еўрарадыё за 2016 год [4].

6. Праект OPUS [5], які ўключае вялікую калекцыю перакладных тэкстаў з Інтэрнэту, апрацаваных выключна аўтаматычна. Для

беларускай мовы самы вялікі корпус сабраны з Вікіпедыі (каля ста шасцідзесяці тысяч сказаў).

Ні адзін з вышэйпералічаных карпусоў не падыходзіць для задачы пабудовы перакладчыка з нуля па наступных прычынах:

1. Моналінгвістычныя даныя (Беларускі N-корпус, Эксперыментальны корпус беларускай мовы). Відавочна, што гэтыя корпусы могуць быць выкарыстаны толькі для паляпшэння ўжо натрэніраванай мадэлі перакладу [6]. Для пабудовы новага аўтаматычнага перакладчыка трэба мець паралельныя даныя.

2. Спецыфічная крыніца даных (Corpus Albaruthenicum). Дадзены корпус не можа быць выкарыстаны, паколькі навуковыя тэрміны не адносяцца да агульнаўжывальнай лексікі. На базе дадзенай крыніцы можна ствараць вузканакіраваныя слоўнікі.

3. Недаступныя даныя (корпус у рамках НКРМ). Нацыянальны корпус рускай мовы мае мільён паралельных сказаў для моўнай пары руская-беларуская, але гэтыя дадзеныя недаступныя для спампавання. На ліст з просьбай даць доступ да іх для даследвання адказ атрыманы не быў.

4. Недастатковая колькасць паралельных сказаў (праца А.Ф. Брыля, праект OPUS). Гэтыя крыніцы нам падыходзяць, але для пабудовы якаснай сістэмы перакладу неабходна ў два разы больш даных.

Абагулім высновы з дапамогай наступнай табліцы:

Табліца 1

Агляд існуючых карпусоў

Назва	Паралельнасць	Агульнаўжывальнасць	Даступнасць	Аб'ем
Беларускі N-корпус	-	+	+	+
Эксперыментальны корпус беларускай мовы	-	+	+	+
Corpus Albaruthenicum	-	-	-	-
Корпус на старонцы НКРМ	+	+	-	+
Распрацоўка А.Ф. Брыля	+	+	+	-
Праект OPUS	+	+	+	-

Такім чынам, задача пабудовы корпуса паралельных беларуска-рускіх сказаў застаецца актуальнай. У якасці крыніцы даных быў абраны сайт 'БелаПАН' дзякуючы досыць вялікаму архіву навін (даныя знаходзяцца ў адкрытым доступе з кастрычніка 2007 года) і зручна структураванаму адрасу старонак (*'https://belapan.by/archive/2007/11/30/200521/'*). Тэксты адрозніваюцца шырокай тэматыкай: палітыка, грамадства, культура, спорт, эканоміка.

Стыль напісання у асноўным афіцыйна-справавы, часам публіцыстычны. Той факт, што на гэтым сайце артыкулы перакладаюцца фактычна слова ў слова, значна спрашчае наступную задачу выраўноўвання паралельнага корпуса. Перад напісаннем праграмы для спампавання даных быў атрыманы афіцыйны дазвол ад ‘БелаПАН’.

Для камп’ютарнай рэалізацыі была абрана мова праграмавання Python і распрацаваны наступны алгарытм:

1. Атрымліваем усе лінкі на беларускамоўныя артыкулы з кастрычніка 2007 па снежань 2018. Прыклад старонкі з спасылкамі: *‘https://belapan.by/archive/?filterby=month&startdate=2007-11-01’*. Шмат архіўных артыкулаў даступныя толькі платна, але адкрытыя даныя адрозніваюцца спецыяльным тэгам разметкі.

2. Кожнаму артыкулу на беларускай мы шукаем аналаг на рускай мове, выкарыстоўваючы рэгулярныя выразы для змены url (*‘https://by.belapan.by/archive/2007/11/30/200521_200545/’* – беларуская версія, *‘https://belapan.by/archive/2007/11/30/200521/’* – руская версія). Асобна правяраем, ці ёсць доступ да паралельнага артыкула на рускай мове. Да таго ж, старыя матэрыялы не падыходзяць нашаму шаблону *(‘https://by.belapan.by/archive/2007/08/23/pragnoz_nadvorya_na_sennya/’* – немагчыма знайсці рускі адпаведнік па агульных правілах).

3. Для разбіцця тэкста артыкулаў на сказы выкарыстоўваем бібліятэку *‘nltk’* і метад *‘sent_tokenize’*. Вынікі запісваем у тэкставы файл.

Пасля першай спробы запуску праграмы былі высветлены наступныя недахопы:

1. У пачатку кожнага артыкула ёсць дата і месца дзеяння (напрыклад, *‘Мінск, 31 кастрычніка’*). Падобныя сказы не ўтрымліваюць у сабе карысную лінгвістычную інфармацыю. Рашэнне: падчас запісу ў файл заўсёды прапускаем першы сказ артыкула.

2. Стандартны такенізатар працуе некарэктна для рускай і беларускай мовы. Мяжой сказаў лічацца словы *‘тыс.’, ‘млрд.’, ‘ул.’*. Рашэнне: існуе спецыяльны такенізатар для рускай мовы [7]. Ён добра падыходзіць і для беларускай, але неабходна уручную пашырыць базу беларускамоўных абрэвіятур: *‘тэл.’, ‘вул.’*.

3. Зашумленасць даных. Сярод разнастайных артыкулаў штодня публікуюцца курсы валют, а таксама рэкламныя звесткі, якія не маюць лінгвістычнай каштоўнасці. Рашэнне: знайсці пэўны падрадок, па якім можна аддзяліць лішнія даныя. Потым трэба запісаць у файл толькі тыя сказы, якія не змяшчаюць ніводнага з раней вызначаных падрадкоў.

Прыклады падрадкоў: *‘Нацыянальны банк Беларусі ўстанавіў’, ‘Даведка БелаПАН.’, ‘можна замовіць у службе падпіскі’*.

4. Часам колькасць сказаў у паралельных матэрыялах не супадае. Напрыклад, у адным з артыкулаў прапушчаны прабел і такенізатар ужо не дзеліць сказы, якія ідуць адзін за адным. Альбо мае месца неадпаведнасць даных у артыкулах на рускай і беларускай мовах. (*‘https://belapan.by/archive/2007/10/26/194788’* – у гэтым артыкуле загаловак вынесены ў тэкст, а ў беларускай версіі такога няма). Рашэнне: правяраць даныя на супадзенне колькасці сказаў для кожнай асобнай пары паралельных артыкулаў і пісаць у файл толькі прыдатныя варыянты. Апрацоўка асобных выпадкаў патрабуе шмат часу, а неабходную колькасць даных можна атрымаць і з такой умовай.

Пасля спампавання ўсіх даступных матэрыялаў было атрымана каля сямісот тысяч паралельных сказаў, што дастаткова для заяўленай мэты. Прыклад пары: *‘Анализ рынков сбыта.’ – ‘Аналіз рынкаў збыту’*. Аднак неабходна яшчэ правесці фінальнае выраўнаванне даных. Асноўныя праблемы:

1. Лішнія або прапушчаныя прабелы.

2. Лінгвістычная розніца некаторых моўных канструкцый. Дзеепрыметнікавы зварот у рускай мове замяняецца складаным сказам у беларускай: *‘Лучшим из белорусов в ней стал Евгений Абраменко, занявший 52-е место.’ – ‘Лепшым з беларусаў у ёй стаў Яўген Абраменка, які заняў 52-е месца.’*.

3. Тэкст не заўсёды абсалютна паралельны. Напрыклад, некаторыя артыкулы ў рускім варыянце распісаны больш дэталёва, чым у беларускім *(‘Той уцёк.’ – ‘Тот кинулся убегать, догнать его, к сожалению, прохожий не смог.’*).

Спроба выраўняць сказы ўручную, што займае шмат часу, ці выкарыстаць гатовыя праграмы, сярод якіх няма адпаведных, была непаспяховай, таму быў абраны іншы шлях. З дапамогай мовы праграмавання Python быў распрацаваны алгарытм пасімвальнага параўнання радкоў. Калі адносіны меншай даўжыні сказа да большай апускаюцца ніжэй за эмпірычна вызначаны парог, то праграма выводзіць нумар радка, тэкст сказаў і спыняецца.

```
Coef: 0.6923076923076923
Border: 0.7341772
Index: 930
Пресс-конференция, посвященная развитию сферы бытовых услуг в Беларуси.

Прэс-канферэнцыя, прысвечаная развіццю сферы бытавых паслуг у Беларусі (пачатак у 11.00, Белпрэсцэнтр).
```

Мал. 1. Прыклад працы праграмы па параўнанні радкоў

Потым адбываецца ручная карэкціроўка, што дазваляе беспамылкова вырашаць спрэчныя лінгвістычныя моманты, якія былі б

даволі складаныя для аўтаматычныга выраўнования. Такі падыход патрабуе шмат часу, аднак выніковы корпус атрымліваецца якасным. У рамках фінальнай апрацоўкі даных пасля выраўнования былі выдаленыя пары аднолькавых сказаў. Увесь код і апрацаваныя даныя змешчаны ў адкрытым доступе [8].

Заключэнне. Такім чынам, у рамках дадзенай працы быў створаны новы паралельны корпус. Ягоны памер – 430 тысячаў сказаў, складаецца з навінавых тэкстаў шырокай тэматыкі. Асабліва карысны дадзены корпус для даследванняў у галіне машыннага перакладу для беларускай мовы. У якасці падзадач быў рэалізаваны скрыпт, які дазваляе спампоўваць навінавыя артыкулы сайта ‘БелаПАН’, праведзена ручное выраўнованне і чыстка даных. У «Аб'яднаным інстытуце праблем інфарматыкі» Нацыянальнай акадэміі навук Беларусі студэнты вырашаюць шмат задач, звязаных з беларуска-рускай моўнай парай. У прыватнасці, атрыманыя практычныя вынікі (паралельны корпус) будуць карысныя для будучай навуковай дзейнасці ўсіх даследчыкаў праблемы машыннага беларуска-рускага і руска-беларускага перакладу.

БІБЛІЯГРАФІЧНЫЯ СПАСЫЛКІ

1. Беларускі N-корпус [Электронны рэсурс]. URL: https://bnkorpus.info.
2. Волчек О. А., Порицкий В. В. Экспериментальный корпус белорусского языка: текущее состояние и перспективы развития // Труды международной конференции «Корпусная лингвистика–2013». СПб.: СПбГУ, 2013.
3. Национальный корпус русского языка [Электронны рэсурс]. URL: https://ruscorpora.ru/new/search-para-be.html.
4. Euroradio [Электронны рэсурс]. URL: https://euroradio.fm/cyaper-z-nulya-mozhna-lyogka-zrabic-autamatychny-perakladchyk-z-ruskay-na-belaruskuyu.
5. Open Parallel Corpus [Электронны рэсурс]. URL: http://opus.nlpl.eu/.
6. Sennrich R., Haddow B., Birch A. Improving Neural Machine Translation Models with Monolingual Data / Rico Sennrich, // [Electronic resource]. URL: https://www.aclweb.org/anthology/P16-1009.pdf.
7. Russian Tokenizer [Электронны рэсурс]. URL: https://github.com/Mottl/ru_punkt/.
8. Translator [Электронны рэсурс]. URL: https://github.com/tsimafeip/Translator.