ACTA ALBARUTHENICA 20: 2020
DOI: 10.32612/uw.18988091.2020.20.pp.185–198

Радаслаў Калета / Radosław Kaleta
Uniwersytet Warszawski / University of Warsaw
ORCID: https://orcid.org/0000-0001-6892-9332
e-mail: rkaleta@uw.edu.pl

Ігар Капылоў / Ihar Kapylou
Нацыянальная акадэмія навук Беларусі / National Academy of Sciences
ORCID: https://orcid.org/0000-0002-1862-9005
e-mail: iharkap@gmail.com

Уладзімір Кошчанка / Uladzimir Koshchanka
Нацыянальная акадэмія навук Беларусі / National Academy of Sciences
ORCID: https://orcid.org/0000-0002-4942-0896
e-mail: koshul@gmail.com

Корпусная лінгвістыка як перспектыўны напрамак развіцця беларуска-польскіх моўных сувязей

Corpus linguistics as a perspective direction for the development of Belarusian-Polish linguistic relations

Lingwistyka korpusowa jako perspektywiczny kierunek rozwoju białorusko-polskich związków językowych

Корпусы як інфармацыйна-даведачныя сістэмы, створаныя на аснове прадстаўнічых электронных збораў тэкстаў, на сучасным этапе з'яўляюцца добра вядомымі і шырока запатрабаванымі рэсурсамі для рознабаковага вывучэння любой мовы. Актуальнасць стварэння корпусаў прадыктаваная неабходнасцю замены інтуітыўнага падыходу дакументаваным падыходам, што відавочна вядзе да значнага павышэння ўзроўню лінгвістычных даследаванняў і аб'ектыўнасці атрыманых вынікаў. У Беларусі, як і ў Польшчы, была праведзена пэўная праца ў гэтым кірунку.

У Інстытуце моваznаўства імя Якуба Коласа Нацыянальнай акадэміі навук Беларусі праца па аўтаматызацыі лінгвістычных даследаванняў пачалася ў канцы 1970-х па ініцыятыве прафесара Віктара Мартынава, адным з кірункаў якой было стварэнне канкардансаў беларускай мовы. З таго часу і па сённяшні дзень быў створаны шэраг канкардансаў: *Канкарданс беларускай мовы XIX стагоддзя*, *Канкарданс твораў Кандрата Крапівы*, *Канкарданс твораў Кузьмы Чорнага* (быў страчаны). Пазней праца ў гэтым напрамку была фактычна згорнута і толькі

ў апошнія гады аднавілася: быў пашыраны *Канкарданс беларускай мовы XIX стагоддзя*, дадаліся канкардансы твораў асобных аўтараў (Вінцэнта Дуніна-Марцінкевіча, Аляксандра Ельскага, Кастуся Каліноўскага і інш.) [Сянкевіч 2017].

Першыя спробы стварэння ўласна корпусаў адносяцца да пачатку 2000-х гадоў. Так, у выніку выканання аднаго з падзаданняў дзяржаўнай праграмы *Электронная Беларусь* быў створаны *Машынны фонд беларускай мовы* (2005 г.), які ўтрымліваў і паралельны руска-беларускі корпус юрыдычных тэкстаў аб'ёмам каля 1 млн. словаўжыванняў [Рубашко 2006]. У 2001 г. у Гродзенскім дзяржаўным універсітэце быў створаны корпус твораў Янкі Купалы [Рычкова 1997].

У 2001 г. у Інстытуце мовазнаўства імя Якуба Коласа Нацыянальнай акадэміі навук Беларусі выконвалася заданне *Праблема моўнай рэпрэзентатыўнасці і прынцыпы пабудовы корпусу беларускай мовы* дзяржаўнай праграмы фундаментальных навуковых даследаванняў (навуковы кіраўнік – прафесар Генадзь Цыхун). У выніку даследавання былі вызначаны: *жанравая рэпрэзентатыўнасць, храналагічна-гістарычная рэпрэзентатыўнасць, геаграфічная рэпрэзентатыўнасць, варыянтна-правапісная рэпрэзентатыўнасць.* Дадаткова вялася праца па распрацоўцы стандарту і прынцыпаў структурнага анатавання (разметкі) тэкстаў. У цэлым даследаванне мела тэарэтычны характар.

Далейшая праца над корпусам была перанесеная ў Мінскі дзяржаўны лінгвістычны ўніверсітэт і вялася пад кіраўніцтвам прафесара Аляксандра Зубава. На гэтым этапе адпрацоўвалася сістэма граматычнай і структурнай разметкі, пашыралася граматычная база. У выніку выканання задання быў створаны аднамоўны корпус аб'ёмам 1 млн. словаўжыванняў, паралельны англа-беларускі корпус тэкстаў, а таксама нямецка-беларускі корпус тэкстаў, кожны аб'ёмам 300 тыс. словаўжыванняў. Корпусы, створаныя ў межах задання, не з'яўляюцца публічнымі [Зубов, Кощенко 2006] .

З 2008 года ў межах 7-ай рамкавай праграмы Еўрапейскага саюзу ў Беларусі ажыццяўляўся міжнародны праект *BalticGrid-*⍰ Адным з кірункаў у праекце была распрацоўка лінгвістычных рэсурсаў для краін Балтыі і Беларусі. У прыватнасці, Літва (Вільнюскі ўніверсітэт) і Беларусь (Беларускі нацыянальны тэхнічны ўніверсітэт – навукова-даследчая лабараторыя дынамікі сістэм і механікі матэрыялаў пры ўдзеле спецыялістаў з Інстытута мовазнаўства імя Якуба Коласа) распрацоўвалі тэкставыя корпусы навуковай мовы, адпаведна літоўскай і беларускай. У выніку быў створаны першы публічны аднамоўны корпус навуковых тэкстаў беларускай мовы *Corpus Albaruthenicum* аб'ёмам 350 тыс. словаўжыванняў (са знятай аманіміяй), які даступны па адрасе: http://grid.bntu.by/corpus/ [Кошчанка, Капылоў, Міклашэвіч 2009].

У сектары камп'ютарнай лінгвістыкі Інстытута мовазнаўства імя Якуба Коласа створаны *Беларускі Біблійны корпус* (http://biblija.bnkorpus.info). Ён уяўляе сабой збор перакладаў Бібліі на беларускую мову

розных гадоў. Корпус змяшчае 16 перакладаў Бібліі на беларускую мову. Для параўнання былі ўключаны таксама асобныя небеларускія пераклады [Булойчык, Кошчанка 2019].

З 2015 года ў Інстытуце мовазнаўства імя Якуба Коласа вядзецца праца па стварэнні *Нацыянальнага корпусу беларускай мовы*, прататып якога даступны анлайн (http://bnkorpus.info).

У Польшчы корпусная лінгвістыка развіваецца з 70-х гг. XX ст. Падрабязна пра тагачаснае супрацоўніцтва інфарматыкаў факультэта матэматыкі і лінгвістаў факультэта паланістыкі Варшаўскага ўніверсітэта піша ў сваім артыкуле праф. Марэк Свідзіньскі (польск. ⍰arek Świdziński) – былы кіраўнік аддзела камп'ютарнага мовазнаўства [Świdziński 2006]. Паводле яго слоў, адна з першых у свеце задач камп'ютарнай лінгвістыкі была выканана ў Польшчы і датычыла польскай мовы. У 1967–1971 гг. у Варшаўскім універсітэце быў створаны першы (500 тыс. словаўжыванняў) корпус, які быў асновай для частотнага слоўніка польскай мовы *Słownik frekwencyjny polszczyzny współczesnej* (*SFPW*). У 2001–2004 гг. у рамках навуковага праекта (навуковы кіраўнік – Адам Пшэп'юркоўскі, польск. Adam ⍰rzepiórkowski) у Інстытуце падстаў інфарматыкі Польскай акадэміі навук быў створаны 100-мільённы корпус польскіх тэкстаў з разметкай (http://nlp.ipipan.waw.pl/~adamp/⍰apers/2004-corpus/) [⍰rzepiórkowski 2004; Świdziński 2006: 27–28]. Існуе таксама вельмі папулярны і даступны ў Сеціве (https://sjp.pwn.pl/korpus) корпус польскай мовы Польскага навуковага выдавецтва *Korpus Języka Polskiego PWN* (100 мільёнаў словаўжыванняў). У корпусе знаходзяцца фрагменты тэкстаў з часопісаў, рэклам, мастацкай літаратуры, палажэнняў, інструкцый, інтэрнэт-выданняў і запісаў вусных тэкстаў. З 2007 г. назіраецца дынамічнае развіццё польскай камп'ютарнай лінгвістыкі ў сувязі з навуковым грантам польскага Міністэрства навукі і вышэйшай адукацыі, які дазволіў стварыць *Нацыянальны корпус польскай мовы* – Narodowy Korpus Języka ⍰olskiego (больш за паўтара мільярда словаўжыванняў) – інтэрнэт-скарбніцу, якая з'яўляецца публічнай (http://nkjp.pl/) [Рэд. ⍰rzepiórkowski, Bańko, Górski, Lewandowska-⍰omaszczyk 2012; Bańko, Górski 2014]. Згаданы праект – гэта супольная ініцыятыва Інстытута падстаў інфарматыкі ПАН (кардынатар праекта), Інстытута польскай мовы ПАН, Навуковага выдавецтва ⍰WN, а таксама аддзела камп'ютарнага і корпуснага мовазнаўства Лодзьскага ўніверсітэта. Корпус дазваляе аналізаваць, сярод іншага, скланенне і спражэнне часцін мовы і нават структуру польскіх сказаў. Корпус змяшчае не толькі тэксты класікаў польскай літаратуры, але таксама тэксты прэсы, інтэрнэт-выданняў, адмысловыя тэксты і запісы вусных размоў людзей рознага ўзросту і полу з розных рэгіёнаў Польшчы.

У Польшчы развіваюцца і меншыя праекты, напр., корпус польскіх тэкстаў XVII і XIX стст. [Bronikowska, ⍰rzyborska-Szulc 2018], корпус выказванняў польскіх палітыкаў [Graszewicz 2014], корпус размоўнай

мовы (маўлення) [Demenko 2015] і нават корпус развітальных лістоў самазабойцаў (http://www.pcsn.uni.wroc.pl/). Падрабязна польскамоўныя паралельныя корпусы, у тым ліку польска-рускі (http://pol-ros.polon.uw.edu.pl, 30 млн. словаўжыванняў) і польска-ўкраінскі (www.domeczek.pl, 6,5 млн. словаўжыванняў), разглядаюцца ў манаграфіі, выдадзенай Інстытутам прыкладной лінгвістыкі Варшаўскага ўніверсітэта [Рэд. Gruszczyńska, Leńko-Szymańska 2016]. Вельмі цікавы праект вядзе Ангеліка Пэльяк-Лапінская (польск. Angelika ?eljak-Łapińska), выпускніца кафедры беларусістыкі Варшаўскага ўніверсітэта, дактарантка ўніверсітэта Суонсі (анг. Swansea University, Вялікабрытанія) і супрацоўніца Лодзьскага ўніверсітэта. Яна напісала на польскай мове бакалаўрскую працу *Моўны вобраз жанчыны і мужчыны: Польска-беларускі корпусны аналіз* (Варшава 2015, навуковы кіраўнік – Радаслаў Калета), на беларускай мове – магістарскую працу *Негацыя ў беларускай мове: Марфалагічна-сінтаксічны і прагматычны аналіз на аснове корпусу беларускага маўлення* (Варшава 2017, навуковы кіраўнік – Ніна Баршчэўская), і на англійскай мове – кандыдацкую дысертацыю *Design, compilation and applications of an English-Polish-Belarusian Parallel Literary Corpus* (Суонсі 2020, навуковы кіраўнік – Том Чысман, анг. ?om Cheesman), у якой аўтар даследуе пераклады англійскай літаратуры XX і XXI стст. на беларускую і польскую мовы, стварыўшы перакладны паралельны англійска-польска-беларускі корпус (https://epbcorpus.wordpress.com/, 10 млн. словаўжыванняў). А. Пэльяк – гэта першая вядомая нам польская даследчыца, якая ў навуковым артыкуле (даступным у Сеціве) аналізавала таксама Беларускі N-корус [?eljak-Łapińska 2016]. Гэта другі артыкул пра беларускія корпусы на польскай мове (першы належыць беларусу Аляксею Яскевічу – рэдактару і тэхнічнаму адміністратару ?hilology.BY [гл. Yaskevich 2014]).

Асобнае месца ў сістэме корпусных даследаванняў займае стварэнне паралельных корпусаў. Інстытутам мовазнаўства імя Якуба Коласа сумесна з Мінскім дзяржаўным лінгвістычным універсітэтам і Інстытутам рускай мовы імя В.У. Вінаградава РАН у 2011–2013 гг. быў створаны паралельны беларуска-рускі і руска-беларускі корпус тэкстаў [Сичинава 2012; Кошчанка 2013]. Аб'ём корпуса сёння складае каля 11 млн. словаўжыванняў, ён даступны на сайце Нацыянальнага корпусу рускай мовы (http://ruscorpora.ru/search-para-be.html).

Паралельныя корпусы маюць важнае значэнне для вывучэння блізкароднасных моў і адыгрываюць асаблівую ролю для выяўлення тонкіх адрозненняў у іх структуры, лексічнай семантыцы і граматыцы. Выкарыстанне лінгвістычных корпусаў з'яўляецца агульнапрызнаным у мовазнаўстве спосабам кантрастыўных даследаванняў моўнай структуры. Таксама, паралельныя корпусы – добры інструмент для перакладчыка, часта больш эфектыўны за перакладныя слоўнікі, паколькі дае большую колькасць эквівалентаў для перакладу і дазваляе верыфі-

каваць значэнні лексічных і фразеалагічных адзінак, зафіксаваных у слоўніках.

Паралельныя корпусы з'яўляюцца высокаэфектыўным інавацыйным дадаткам да традыцыйных адукацыйных тэхналогій, яны выконваюць важныя дыдактычныя і метадычныя функцыі, што знайшло сваё адлюстраванне ў сусветнай практыцы навучання замежным мовам.

Стварэнне паралельнага польска-беларускага і беларуска-польскага корпуса з'яўляецца адным з перспектыўных напрамкаў развіцця польска-беларускага супрацоўніцтва ў галіне мовазнаўства. З аднаго боку, актуальнасць стварэння падобнага корпуса прадыктавана неабходнасцю павышэння ўзроўню аб'ектыўнасці пры кантрастыўных даследаваннях нашых нацыянальных моў, а з іншага боку, паралельны корпус можа стаць незаменным інструментам пры навучанні як польскай, так і беларускай моў (вывучаюцца ў ВНУ абедзвюх краін). Народы Польшчы і Беларусі маюць багатую і працяглую гісторыю суседства і адпаведна напрацавалі істотны корпус перакладных тэкстаў. Так, на беларускую мову перакладаліся творы Адама Міцкевіча, Уладзіслава Сыракомлі, Элізы Ажэшкі, Марыі Канапніцкай, Генрыка Сянкевіча, Мацея Стрыйкоўскага, Дамініка Рудніцкага, Чэслава Мілаша, Стэфана Жаромскага, Цыпрыяна Норвіда, Віславы Шымборскай, Анджэя Сапкоўскага, Януша Вішнеўскага, Вольгі Такарчук і інш., на польскую – Якуба Коласа, Янкі Купалы, Максіма Танка, Уладзіміра Караткевіча, Васіля Быкава, Альгерда Бахарэвіча, Андрэя Адамовіча і інш.

За час незалежнасці Беларусі аб'ём перакладзеных на беларускую мову тэкстаў з польскай мовы істотна вырас, польская мова таксама ўсё шырэй вывучаецца як замежная і ў тым ліку праз пасрэдніцтва беларускай мовы. Створана некалькі перакладных слоўнікаў. Таксама павялічылася колькасць навуковых даследаванняў, звязаных з польскай мовай. У гэтай сітуацыі наспела вострая неабходнасць у стварэнні паралельнага польска-беларускага корпусу.

Для ацэнкі мэтазгоднасці дадзенага праекта трэба ў першую чаргу вывучыць нарматыўна-прававую базу Польшчы і Беларусі адносна аўтарскіх правоў і выкарыстання тэкстаў для навуковага цытавання. Акрамя таго, неабходна высветліць аб'ём тэкстаў, перакладзеных з польскай мовы на беларускую і з беларускай на польскую.

Для ўкладання корпусаў выкарыстоўваюцца электронныя версіі, атрыманыя ад аўтараў, з выдавецтваў, з агульнадаступных у Сеціве інтэрнэт-бібліятэк, а таксама ў неабходным аб'ёме праводзіцца першасная алічбоўка (сканаванне) друкаваных крыніц. У сувязі з гэтым паўстае пытанне аўтарскіх правоў і адпаведна ўзнікае праблема атрымання тэкстаў. Беларускае заканадаўства дазваляе выкарыстоўваць тэксты, выдадзеныя ў Беларусі, у мэтах навуковага цытавання. У Польшчы сітуацыя больш складаная, аб чым сведчыць досвед працы над польска-рускім паралельным корпусам. Некаторыя юрысты сцвярджаюць, што выкарыстоўванне паасобных сказаў або абзацаў не патрабуе

дазволу аўтараў, але з іншага боку, уключэнне канкрэтнага тэксту ў корпус (нават з мэтай выкарыстаць толькі яго фрагменты-цытаты) патрабуе апрацоўкі ўсяго твора [Лазинский, Куратчик 2015: 87; Łaziński, Kuratczyk 2016: 86]. Звычайна тэксты мастацкай літаратуры, якія не абараняюцца аўтарскімі правамі, можна выкарыстоўваць без дазволу, але ёсць нюансы, напр., калі перакладчык беларускага твора на польскую мову памёр у 60–70 гг. XX ст., то абарона яго правоў (паводле польскага закона) будзе працягвацца яшчэ 70 гадоў пасля яго смерці. Такія тэксты (пераклады) не могуць знаходзіцца ў адкрытым доступе, а толькі ў закрытым, дзе могуць быць даступнымі толькі для некаторых асоб, які хочуць карыстацца корпусам для дыдактычных або навуковых мэт. У выпадку мастацкай літаратуры, якая абараняецца аўтарскімі правамі, трэба старацца атрымаць дазвол на выкарыстанне тэкстаў для корпусных мэт (або хаця б адказ з боку ўладальнікаў аўтарскіх правоў на просьбы супрацоўнікаў праекта). Гэта трэба браць пад увагу пры ўкладанні польска-беларускага паралельнага корпуса, хаця сёння ёсць новыя магчымасці атрымліваць і тэксты, і дазвол на іх выкарыстанне, дзякуючы ўсё больш папулярным інтэрнэт-бібліятэкам. З другога боку, паводле польскага заканадаўства, ёсць таксама тэксты, якія не абараняюцца аўтарскімі правамі, напр., кароткія тэксты навін, інструкцый, законаў, дамоў і г. д. Сёння на практыцы стваральнікі малых корпусаў звычайна не звяртаюцца да ўладальнікаў аўтарскіх правоў і робяць доступ да сваіх корпусаў закрытым – аднак любы даследчык можа да іх звярнуцца з просьбай атрымаць пароль для карыстання корпусам. Некаторыя менавіта так разумеюць арт. 27 *Закона аб аўтарскіх правах* (ад 4 лютага 1994 г.), дзе сцвярджаецца, што навуковыя і дыдактычныя адзінкі для навуковых і дыдактычных мэт могуць карыстацца распаўсюджанымі творамі ў арыгінале і перакладзе. Аднак доступ да такіх выкарыстаных твораў можа мець толькі вузкае кола асоб, якія вучацца, выкладаюць або вядуць навуковыя даследаванні. Выкарыстаныя цытаты павінны паказваць звесткі пра аўтара арыгінала (і аўтара перакладу), загаловак арыгінала (і перакладу).

Вырашэння патрабуе і тэхнічны аспект: месцазнаходжанне пляцоўкі, на якой будзе размешчаны паралельны корпус.

Таксама трэба вывучыць пытанне жанравай разнастайнасці перакладных тэкстаў і вызначыцца з крытэрыямі адбору. У сітуацыі з польска-беларускімі перакладамі можна меркаваць, што большасць даступных тэкстаў будуць адносіцца да канфесійнага, публіцыстычнага, мастацкага, і ў невялікай ступені, афіцыйна-справавога стыляў.

Асобная праблема – гэта тэксты, перакладзеныя пры дапамозе машынных перакладчыкаў. Верагодна, што гэта праблема менш актуальная, чым пры руска-беларускіх перакладах, але калі будуць трапляцца такія тэксты, ці павінны яны ўключацца, нават калі ў іх няма памылак, у корпус?

Пасля вызначэння з крытэрыямі і структурай корпусаў можна прыступаць да адбору тэкстаў. Гэты этап будзе ўключаць як сканаванне, так і атрыманне электронных версій. Працэс сканавання і вычыткі беларускімі спецыялістамі ў галіне корпуснай лінгвістыкі адпрацаваны падчас працы над Корпусам для новага *Тлумачальнага слоўніка беларускай мовы*. Распрацаванае праграмнае забеспячэнне будзе пасаваць і для апрацоўкі тэкстаў для польска-беларускага і беларуска-польскага корпусаў.

У прыватнасці, у Інстытуце мовазнаўства распрацаваная праграма *Параўнанне сканаў*. Тэкст апрацоўваецца ў праграме распазнавання тэкстаў у два этапы: першы этап – без слоўніка, а другі этап – са слоўнікам. Праграма *Параўнанне сканаў* параўноўвае файл, распазнаны без слоўніка, і файл, распазнаны са слоўнікам, сінхранізуе гэтыя файлы, прымяняе шаблон і вынікі захоўвае ў асобны файл, які мае наступны выгляд:

I

Восень заўсёды падкрадваецца неўзаметку. Янічэ {Яшчэ }прыгравае сонца, усё вакол зялёнае, яшчэ цвітуць гуркі, на доўгай гарбузовай касе смела ўсміхаецца вялізная пяцікутная жоўта-залатая кветка. Лета не хоча здавацца. І {І }раптам прачнешся раніцаю {ранідаю }— на вуліцы туман, густы, белы, як малако. А калі ён развеецца, на бярозе ярчэй засвецяцца жоўтыя лісцікі, бульбоўнік пачарнее, бы абвараны, і раса на траве халодная, аж коле ў босыя пяткі.

Звычайна на трэцім перапынку хлапчукі выбягалі на ўзгор'е і скакалі з абрыву ў пясчаную ямку. На дне яе жвірысты пясок быў халаднаваты, хоць зверху трохі прагравaўся, ды неўзабаве парэпаныя ад цыпак ногі перамешвалі яго, ператаўкалі, як проса ў ступе.

Скакалі хлапчукі не проста так сабе. Тут ішло зацятае спаборніцтва — хто далей скокне. Не раз быў чэмпіёнам Андрэйка Сахута. Але нечакана для ўсіх уперад вырваўся Парасчын Данілка. Андрэйка гэтага не мог перажыць. Калі б Паўлік, ягоны даўні сяброк, было б не так крыўдна, а то смаркаты Данька раптам сігануў далей за ўсіх. Мабыць, і Паўліка гэта раздражніла, ён шмаргануў носам і са здзіўленнем у голасе сказаў:

— Смелы ёты {еты }байструк. У матку ўдаўся. Яна ваўка забіла...

— Хто байструк? А то ў нос як дам зараз! — замахнуўся брудным кулачком Данька.

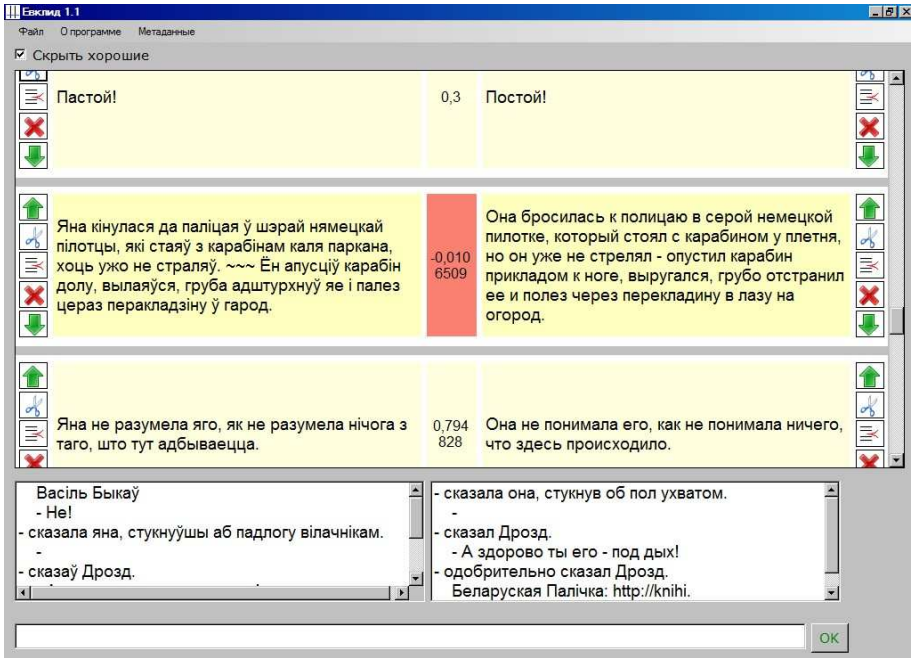
— Ну, паспрабуй! — натапырыўся Паўлік.— Як дам выспятка, дык і пакоцішся прама ў Бесядзь.

Колерам у тэксце вылучаны спрэчныя месцы, на якія трэба асабліва звяртаць увагу. У аснову кладзецца тэкст, расчытаны пры дапамозе слоўніка, але распазнаванне са слоўнікам мае адзін мінус: праграма падбірае найбліжэйшае слова, якое, паводле алгарытму, найбольш пасуе ў гэтым месцы і здараюцца выпадкі, калі слова падабранае, але яно памылковае, таму побач для праверкі ў фігурных дужках пакідаецца слова, распазнанае без слоўніка. У нашым выпадку відаць, што формы *Янічэ* (*Яшчэ*) і *ёты* (*еты*) былі падабраны няўдала, на што ўказвае варыянт у фігурных дужках, а слова раніцаю – удала. Такі падыход дазваляе істотна зэканоміць час пры вычытцы тэкстаў. Адначасова з вычыткай тэксту правяраецца і пазначаецца яго структура, закладзеная ў шаблоне: іерархія загалоўкаў, эпіграфы, подпісы, паэтычныя ўстаўкі ў празаічных тэкстах і г. д.

Для выраўноўвання тэкстаў можна выкарыстоўваць, атрымаўшы дазвол, праграму *Евклид*, распрацаваную спецыялістамі Інстытута рускай мовы імя В.У. Вінаградава РАН. Праграма з'яўляецца графічным

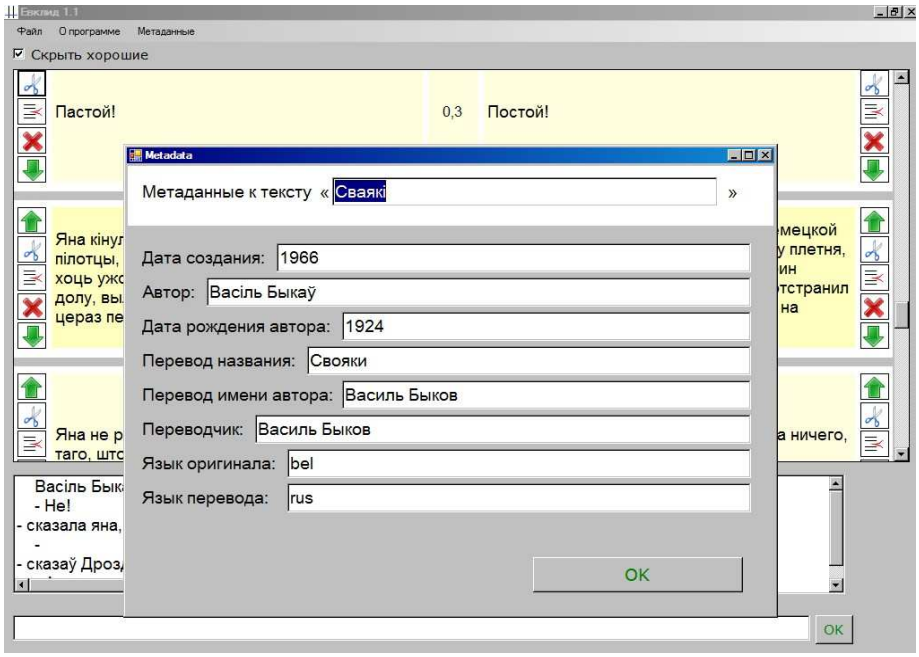
інтэрфейсам карыстальніка да даступнай у адкрытым рэжыме праграмы выраўноўвання тэкстаў *unAlign*. У праграму загружаюцца два тэксты – арыгінал і пераклад, пасля гэтага яны праходзяць этап папярэдняй апрацоўкі: выдаляюцца пустыя радкі, збыткоўныя прагалы, кожны сказ пераносіцца на новы радок, а потым два тэксты выраўноўваюцца і загружаецца графічны інтэрфейс для ручнога пастрэдагавання.

Выраўнаваны тэкст можна захаваць у фармаце XL (кадаванне Юнікод), як у канчатковым корпусным, так і ў прамежкавым фармаце, даступным для далейшага рэдагавання. Пры захаванні выраўнаванага тэксту трэба вызначыць мову арыгінала і перакладу. Яны будуць захаваны ў XL у складзе адпаведных тэгаў. У захаваным XL-файле пары сказаў пранумараваны (ім нададзены ўнікальныя нумары) для абляг-чэння спасылак. У далейшым словам будзе прыпісвацца граматычная інфармацыя.



Мал.1. Інтэрфейс праграмы Евклид

Праграма *Евклид* таксама дазваляе дадаць да кожнага захаванага тэксту метатэкставую інфармацыю (метаразметку), якая запісваецца ў якасці асобнага радка ў электронную табліцу (дакумент, даступны для рэдагавання праграмамі тыпу *Excel*). Карыстальнік запаўняе ў адмысловай форме назву тэксту і імя аўтара ў арыгінале і перакладзе, дату стварэння тэксту; мова арыгінала і перакладу захоўваюцца аўтаматычна, зыходзячы з раней захаванай інфармацыі.



Мал. 2. Форма для запаўнення метатэкставай інфармацыі

```
<para id="17">
  <se lang="be">Старэйшы прыняў яе ўдар з каменнай абыякавасцю на змрочным худым твары, нават не ўздрыгнуў,
  толькі мацней сцяў вусны, і яна зразумела, што ўсё гэта - дарма.</se>
  <se lang="ru">Старший принял ее удар с каменным безразличием на угрюмом худом лице, даже не вздрогнул, только
  плотнее сжал губы, и она поняла, что все это напрасно.</se>
</para>
<para id="18">
  <se lang="be">Дарма ўвесь яе гнеў, яе лаянка, яе запазнелая спроба вярнуць сваю ўладу над хлопцамі.</se>
  <se lang="ru">Напрасен весь ее гнев, ее брань, ее запальчивая попытка вернуть уходящую власть над сынами.</se>
</para>
<para id="19">
  <se lang="be">Роспач адразу падламіла яе і, кінуўшы вілачнік, яна выйшла ў сенцы.</se>
  <se lang="ru">Отчаяние враз сломило ее, и, бросив ухват, она вышла в сени.</se>
</para>
```

Мал. 3. Выніковы выраўнаваны тэкст у фармаце □□ L

Корпусы таксама павінны забяспечвацца граматычнай разметкай. Для граматычнай разметкі беларускіх тэкстаў можна выкарыстаць *Лексіка-граматычную базу беларускай мовы*, якая распаўсюджваецца на ўмовах ліцэнзіі *Creative Commons Attribution/Share-Alike 3.0*. На сённяшні дзень *Лексіка-граматычная база* беларускай мовы мае прыблізна 255 000 парадыгм і больш за 2 500 000 словаформаў.

База ўяўляе сабой збор слоў з марфалагічнымі і іншымі пазнакамі. У загалоўку парадыгмы падаецца ідэнтыфікацыйны нумар парадыгмы (*pdg□d*), пачатковая форма (*Lemma*), граматычная прыкмета лексемы (*□ag*). Пры патрэбе фіксуецца дадатковая інфармацыя: кіраванне для дзеясловаў (*Govern*), значэнне (□eaning), заўвагі. Кожная склонавая форма мае ўласцівыя толькі ёй прыкметы (*Form □ag*). Таксама пазна-

чаецца крыніца слова або формы слова, націск, правапіс, некананічныя формы:

```
<Paradigm pdgId="1114391" lemma="ме+сці" tag="VDMN1">¶
    <Variant id="a" lemma="ме+сці" pravapis="A1957,A2008">¶
        <Form tag="0" slouniki="dzsl2007,sbm2012,tsblm1996,tsbm1984,krapivabr2012">ме+сці</Form>
        <Form tag="R1S" slouniki="dzsl2007,sbm2012">мяту+</Form>¶
        <Form tag="R2S" slouniki="dzsl2007,sbm2012">мяце+ш</Form>¶
        <Form tag="R3S" slouniki="dzsl2007,sbm2012">мяце+</Form>¶
        <Form tag="R1P" type="nonstandard">мяце+м</Form>¶
        <Form tag="R1P" slouniki="dzsl2007,sbm2012">мяцё+м</Form>¶
        <Form tag="R2P" type="nonstandard">мяце+це</Form>¶
        <Form tag="R2P" slouniki="dzsl2007,sbm2012">мецяце+</Form>¶
        <Form tag="R3P" slouniki="dzsl2007,sbm2012">мяту+ць</Form>¶
        <Form tag="PMS" slouniki="dzsl2007">мёў</Form>¶
        <Form tag="PFS" slouniki="dzsl2007">мяла+</Form>¶
        <Form tag="PNS" slouniki="dzsl2007">мяло+</Form>¶
        <Form tag="PXP" slouniki="dzsl2007">мялі+</Form>¶
        <Form tag="I2S" slouniki="dzsl2007">мяці+</Form>¶
        <Form tag="I2P" slouniki="dzsl2007">мяці+це</Form>¶
        <Form tag="RG" type="potential">мятучы+</Form>¶
        <Form tag="RG" type="potential">мё+ўшы</Form>¶
    </Variant>¶
</Paradigm>¶
```


```
<Paradigm pdgId="1114391" lemma="ме+сці" tag="VDMN1">¶
    <Variant id="a" lemma="ме+сці" pravapis="A1957,A2008">¶
        <Form tag="0" slouniki="dzsl2007,sbm2012,tsblm1996,tsbm1984,krapivabr2012">ме+сці</Form>
        <Form tag="R1S" slouniki="dzsl2007,sbm2012">мяту+</Form>¶
        <Form tag="R2S" slouniki="dzsl2007,sbm2012">мяце+ш</Form>¶
        <Form tag="R3S" slouniki="dzsl2007,sbm2012">мяце+</Form>¶
        <Form tag="R1P" type="nonstandard">мяце+м</Form>¶
        <Form tag="R1P" slouniki="dzsl2007,sbm2012">мяцё+м</Form>¶
        <Form tag="R2P" type="nonstandard">мяце+це</Form>¶
        <Form tag="R2P" slouniki="dzsl2007,sbm2012">мецяце+</Form>¶
        <Form tag="R3P" slouniki="dzsl2007,sbm2012">мяту+ць</Form>¶
        <Form tag="PMS" slouniki="dzsl2007">мёў</Form>¶
        <Form tag="PFS" slouniki="dzsl2007">мяла+</Form>¶
        <Form tag="PNS" slouniki="dzsl2007">мяло+</Form>¶
        <Form tag="PXP" slouniki="dzsl2007">мялі+</Form>¶
        <Form tag="I2S" slouniki="dzsl2007">мяці+</Form>¶
        <Form tag="I2P" slouniki="dzsl2007">мяці+це</Form>¶
        <Form tag="RG" type="potential">мятучы+</Form>¶
        <Form tag="RG" type="potential">мё+ўшы</Form>¶
    </Variant>¶
</Paradigm>¶
```


Для разметкі польскіх тэкстаў трэба мець граматычную базу польскай мовы. Аднак адной польскай універсальнай базы не існуе. Ёсць шмат камп'ютарных прылад (гл. http://clip.ipipan.waw.pl/LR), якімі кожны стваральнік корпуса можа ў любы момант скарыстацца, у залежнасці ад апрацаванай канцэпцыі структуры свайго корпуса. Для прыкладу, сінтаксічную апрацоўку можна зрабіць пры дапамозе прылады *Skarbnica* (http://zil.ipipan.waw.pl/Sk%C5%82adnica W), а марфалагічную пры дапамозе прылады *orfeusz* (http://morfeusz.sgjp.pl/).

Калі ёсць базы ў вольным доступе, то можна скарыстацца імі альбо шукаць партнёраў, якія маюць падобныя распрацоўкі.

Для падрыхтоўкі канцавой версіі звестак корпусу таксама патрэбен кампілятар. Для *Беларускага N-корпусу* быў распрацаваны такі кампілятар і яго можна адаптаваць пад патрэбы беларуска-польскага пара-

лельнага корпусу. Кампілятар утрымлівае ўсе *граматычныя тэгі*, чытае *граматычную базу*, чытае ўсе тэксты са знятай і не знятай аманіміяй. Для кожнага слова вызначаецца граматычны тэг і лема, тэкст дзеліцца на асобныя параграфы (якія потым змешваюцца ў выпадковым парадку, каб немагчыма было аднавіць зыходны тэкст – магчыма, гэта можа часткова вырашыць праблему аўтарскіх правоў). Вынікі захоўваюцца ў базу, з якой працуе *інтэрфейс пошуку.*

У якасці базы выкарыстоўваецца база *Apache Lucene*. У адрозненне ад звычайных SQL баз (*PostgreSQL, Oracle, MariaDB* і г.д.) база *Apache Lucene* значна больш падыходзіць для захоўвання інфармацыі з нявызначанай колькасцю атрыбутаў для кожнага аб'екта (бо колькасць слоў у абзацы можа быць рознай і на кожнае слова можа прыпадаць розная колькасць граматычных пазнак) і для сцэнарыя выкарыстання "толькі чытанне без транзакцый".

Беларускі N-корпус (з прыкладна 251 млн. слоў) кампілюецца 15 хвілін на працэсары i7-4770. Кампіляцыя падкорпусу "неразабраныя тэксты" (прыкладна 174 млн. слоў) займае прыкладна той самы час.

У якасці інтэрфейсу пошуку можна адаптаваць інтэрфейс, распрацаваны для *Беларускага N-корпусу*. Інтэрфейс зроблены на *Java Web Application* з інтэрфейсам *Angular+Bootstrap* і выкарыстоўвае граматычныя тэгі, граматычную базу і базу корпусу, створаныя кампілятарам.

Такім чынам, сучасны стан корпусных даследаванняў у Рэспубліцы Беларусь і ў Рэспубліцы Польшча адкрываюць новыя магчымасці ў развіцці беларуска-польскіх моўных сувязей. Механізм корпусных даследаванняў, у якіх выкарыстоўваюцца не штучна сканструяваныя прыклады, а рэальныя кантэксты, перакладзеныя кваліфікаванымі перакладчыкамі і ўжытыя ў тэксце, значна павышае надзейнасць і дакладнасць вынікаў даследавання дзвюх славянскіх моў. Паралельны корпус дазволіць аўтаматычна будаваць перакладныя канкардансы, выдзяляць у дзвюх мовах групы слоў пэўных словаўтваральных і словазмяняльных тыпаў, граматычныя мадэлі паняццяў і іх эквівалентаў у перакладным тэксце, удакладняць значэнні лексічных адзінак, зафіксаваных у слоўніках, апісваць новыя, адсочваць дынаміку змен у лексічнай сістэме моў.

Стварэнне беларуска-польскага і польска-беларускага паралельных корпусаў знойдзе шырокае практычнае выкарыстанне ў вучэбным працэсе пры выкладанні беларускай і польскай моў у вышэйшых навучальных установах; у вучэбна-метадычнай рабоце пры адборы тэкставых прыкладаў для падручнікаў і вучэбных дапаможнікаў; у навукова-даследчай працы пры распрацоўцы складаных тэарэтычных пытанняў лексікалогіі, лексікаграфіі, пры правядзенні навуковых даследаванняў тэксту і яго рознаўзроўневага моўнага аналізу (статыстычнага, марфалагічнага, стылістычнага, семантычнага і г. д.); у лексікаграфічнай практыцы пры стварэнні слоўнікаў рознага тыпу; у перакладазнаў-

стве пры распрацоўцы пытанняў тэорыі і практыкі перакладу тэкстаў рознай жанравай і стылёвай прыналежнасці.

Бібліяграфія

Bańko [illegible] irosław, Górski Rafał. 2014. *Praktyczny przewodnik po korpusie języka polskiego*. W: *Praktyczny przewodnik po korpusach języków słowiańskich*. Red. [illegible]. Hebal-Jezierska. Warszawa: Wydział [illegible]olonistyki Uniwersytetu Warszawskiego. S. 11–28.

Bronikowska Renata, [illegible]rzyborska-Szulc Aldona. 2018. *Elektroniczny korpus tekstów polskich [illegible]V[illegible]i [illegible]V[illegible]wieku (do 1772 roku)*. „[illegible]race Naukowe Uniwersytetu Śląskiego w Katowicach" nr 3670: 129–135.

Bulojčyk Alâksandr, Koščanka Uladzìmìr. 2019. *Belaruskì bìblìjny korpus.* U: *Belaruskaâ mova ŭ sakral'naj sfery: gìstoryâ ì sučasnasc': matèryâly [illegible] ìžnar. navuk. kanf. ([illegible] ìnsk, 21 lûtaga 2018 g.)*. Rèd. S. Garanìn, Ì. Budz'ko, [illegible] [illegible]âtrova. [illegible] ìnsk:Belaruskaâ navuka. S. 76–79 [Булойчык Аляксандр, Кошчанка Уладзімір. 2019. *Беларускі Біблійны корпус.* У: *Беларуская мова ў сакральнай сферы: гісторыя і сучаснасць: матэрыялы Міжнар. навук. канф. (Мінск, 21 лютага 2018 г.)*. Рэд. С. Гаранін, І. Будзько, Т. Пятрова. Мінск: Беларуская навука. С. 76–79].

Demenko Grażyna. 2015. *Korpusowe badania języka mówionego*. Warszawa: Akademicka Oficyna Wydawnicza [illegible]XI[illegible]

Graszewicz [illegible] aurycy. 2014. *Korpus wypowiedzi polskich polityków (KWPP)*. „Dziennikarstwo i [illegible] edia" t. 5: 205–214.

Koščanka Uladzìmìr, Kapyloŭ Ìgar, [illegible]ìklašèvìč Ìgar. 2009. *Korpus belaruskamoŭnyh navukovyh tèkstaŭ âk častka rèalìzacyì mìžnarodnaga praekta "Balticgrid [illegible]".* "Belaruskaâ lìngvìstyka" nr 53: 3–8 [Кошчанка Уладзімір, Капылоў Ігар, Міклашэвіч Ігар. 2009. *Корпус беларускамоўных навуковых тэкстаў як частка рэалізацыі міжнароднага праекта "Balticgrid [illegible]".* "Беларуская лінгвістыка*"* № 53: 3–8].

Koščanka Uladzìmìr. 2013. *Paralel'nyâ belaruska-ruskì ì ruska-belaruskì korpusy na sajce nacyânal'naga korpusu ruskaj movy.* U: *Problemy sovremennoj prikladnoj lingvistiki. Sbornik naučnyh statej.* [illegible]insk: [illegible]GLU. S. 41–47 [Кошчанка Уладзімір. 2013. *Паралельныя беларуска-рускі і руска-беларускі корпусы на сайце нацыянальнага корпусу рускай мовы*. У: *Проблемы современной прикладной лингвистики. Сборник научных статей*. Минск: МГЛУ. С. 41–47].

Łaziński [illegible]., Kuratczyk [illegible]. 2016. *Korpus polsko-rosyjski Uniwersytetu Warszawskiego*. W: *Polskojęzyczne korpusy równoległe = Polish-language Parallel Corpora*. 2016. Red. [illegible]. Gruszczyńska, A. Leńko-Szymańska. Warszawa: Uniwersytet Warszawski. Wydział Lingwistyki Stosowanej. Instytut Lingwistyki Stosowanej. S. 83–95.

Lazinskij [illegible]., Kuratčik [illegible]. 2015. *Pol'sko-russkij paralel'nyj korpus Varšavskogo universiteta i ego ispol'zovanie v lingvističeskom issledovanii.* V: *Âzyki metod. Russkij âzyk na grani metodologičeskogo sryva.* Red. D. Szumska, K. Ozga. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego. S. 95–103 [Лазинский М., Куратчик М. 2015. *Польско-русский паралельный корпус Варшавского университета и его использование в лингвистическом исследовании.* В: *Язык и метод. Русский язык на грани методологического срыва.* Red. D. Szumska, K. Ozga. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego. S. 95–103].

Narodowy Korpus Języka Polskiego. 2012. Red. A. [illegible]rzepiorkowski, [illegible]. Bańko, R. Górski, B. Lewandowska-[illegible]omaszczyk. Warszawa: Wydawnictwo Naukowe [illegible]WN.

[illegible]eljak-Łapińska A. 2016. *Białoruski N-korpus: W kierunku Białoruskiego Korpusu Narodowego*. „Acta Albartuhenica" nr 16: 203–210.

Polskojęzyczne korpusy równoległe = Polish-language Parallel Corpora. 2016. Red. [illegible]. Gruszczyńska, A. Leńko-Szymańska. Warszawa: Uniwersytet Warszawski. Wydział Lingwistyki Stosowanej. Instytut Lingwistyki Stosowanej.

□rzepiórkowski Adam. 2004. *Korpus □P□PAN – wersja wstępna*. Warszawa: I□I □AN.

Rubaško Natal'â, Nevmeržickaâ Galina, Sovpel' Igor'. 2006. *Komp'ûternyj fond belorusskogo âzyka i ego priloženiâ.* V: □*nformacionnye sistemy i tehnologii (□S□`2006): materialy □□□ eždunar.konf., □insk, 1–3 noâbrâ 2006 g.: v 2 č. Č. 2.* Red. A. Kurbackij. □insk: BGU. S. 70–71 [Рубашко Наталья, Невмержицкая Галина, Совпель Игорь. 2006. *Компьютерный фонд белорусского языка и его приложения.* В: *Информационные системы и технологии (□S□'2006): материалы □□□Междунар. конф., Минск, 1–3 ноября 2006 г.: в 2 ч. Ч. 2.* Ред. А. Курбацкий. Минск: БГУ. С. 70–71].

Ryčkova Lûdmìla. 1997. *Kamp'ûtèrnaâ versìâ dramaturgìčnyh tvoraŭ Ânkì Kupaly: pryncypy stvarènnâ, napramkì vykarystannâ.* U: □ìžnarodnyâ kupalaŭskìâ čytannì: matèryâly navukovaj kanferèncyì. Grodna. 25–27 lìstapada 1997 g.* Grodna:Grdu. S. 257–261 [Рычкова Людміла. 1997. *Камп'ютэрная версія драматургічных твораў Янкі Купалы: прынцыпы стварэння, напрамкі выкарыстання*. У: *Міжнародныя купалаўскія чытанні: матэрыялы навуковай канферэнцыі. Гродна. 25–27 лістапада 1997 г.* Гродна: ГрДУ. С. 257–261].

Sânkevìč Nìna. 2917. *Kankardans âk srodak vyâŭlennâ asablìvascej movy mastackaj lìtaratury.* U: *Aktual'nye problemy sovremennoj prikladnoj lingvistiki.* □insk: □GLU. S. 248–254 [Сянкевіч Ніна. 2917. *Канкарданс як сродак выяўлення асаблівасцей мовы мастацкай літаратуры.* У: *Актуальные проблемы современной прикладной лингвистики*. Минск: МГЛУ. С. 248–254].

Sičinava Dmitrij. 2012. *Russko-belorusskij parallel'nyj korpus: opyt razrabotki.* V: *Karpovskie naučnye čteniâ: sb. nauč. st.* Vyp. 6: v 2 č. Č. 1. Red. A. Golovnâ. □insk: Belorusskij Dom pečati. S. 270–272 [Сичинава Дмитрий. 2012. *Русско-белорусский параллельный корпус: опыт разработки*. В: *Карповские научные чтения: сб. науч. ст.* Вып. 6: в 2 ч. Ч. 1. Ред. А. Головня. Минск: Белорусский Дом печати. С. 270–272].

Świdziński □arek. 2006. *Lingwistyka korpusowa w Polsce – źrodła, stan, perspektywy*. „LingVaria" nr 1: 23–34.

Yaskevich Alya□ey. 2014. *Praktyczny przewodnik po korpusie języka białoruskiego*. W: *Praktyczny przewodnik po korpusach języków słowiańskich*. Red. □. Hebal-Jezierska. Warszawa: Wydział □olonistyki Uniwersytetu Warszawskiego. S. 184–197.

Zubov Aleksandr, Košenko Vladimir. 2006. *Korpus tekstov belorusskogo âzyka.* U: □*rudy meždunarodnoj konferencii «Korpusnaâ lingvistika – 2006»*. Sankt-□eterburg: Izd-vo S.-□eterb. un-ta; Izd-vo RHGA. S. 119–120 [Зубов Александр, Кощенко Владимир. 2006. *Корпус текстов белорусского языка.* У: *Труды международной конференции «Корпусная лингвистика – 2006»*. Санкт-Петербург: Изд-во С.-Петерб. ун-та; Изд-во РХГА. С. 119–120].

ABSTRACT: □he article provides information on the development of corpus linguistics in Belarus and □oland. □he importance of creating parallel Belarusian-□olish and □olish-Belarusian parallel corpora is noted, the possible algorithm for building the corpora is given.

KEYWORDS: □olish language corpus, □olish-Belarusian parallel corpus, Belarusian N-Korpus, Corpus Albaruthenicum, Russian-Belarusian parallel corpus, Le□ical-grammatical database of the Belarusian language.

STRESZCZENIE: Artykuł zawiera informacje na temat rozwoju językoznawstwa korpusowego na Białorusi i w □olsce. Zaakcentowano znaczenie tworzenia korpusów równoległych białorusko-polskich i polsko-białoruskich i podano możliwy algorytm do budowy takich korpusów.

SŁOWA KLUCZOWE: polski korpus językowy, polsko-białoruski korpus równoległy, białoruski N-Korpus, korpus Albaruthenicum, rosyjsko-białoruski korpus równoległy, leksykalno-gramatyczna baza danych języka białoruskiego.

РЭЗЮМЭ: Артыкул змяшчае інфармацыю аб развіцці корпуснай лінгвістыкі ў Беларусі і Польшчы. Падкрэсліваецца важнасць стварэння паралельных беларуска-польскага і польска-беларускага корпусаў, пададзены магчымы алгарытм укладання такіх корпусаў.

КЛЮЧАВЫЯ СЛОВЫ: корпус польскай мовы, польска-беларускі паралельны корпус, Беларускі N-Korpus, Corpus Albaruthenicum, руска-беларускі паралельны корпус, Лексіка-граматычная база беларускай мовы.