А.А. Барковіч (г. Мінск)

ЭКСТРАЛІНГВІСТЫЧНЫ КАНТЭКСТ РЭПРЭЗЕНТАТЫЎНАСЦІ БЕЛАРУСКАГА КОРПУСА ТЭКСТАЎ

Як вядома, корпусная лінгвістыка займаецца вывучэннем корпусаў тэкстаў як моўных аб'ектаў. У сучаснай лінгвістыцы корпусныя даследаванні ўжо маюць свае традыцыі, прызнаных лідэраў, свае навуковыя цэнтры, метады і праблематыку. У той жа час, на беларускай глебе корпусная лінгвістыка – адносна новы напрамак, які яшчэ толькі афармляецца, набывае абрысы. Усталяванне корпуснай лінгвістыкі на Беларусі, у пэўнай ступені, пытанне часу.

Развіццё корпуснай лінгвістыкі ў свеце пачалося ў 60-я гады XX ст. паралельна са з'яўленнем камп'ютарных тэхналогій. Менавіта тады ўзнік корпусны стандарт у 1 млн словаўжыванняў. Сёння асобныя корпусныя рэсурсы дасягаюць некалькі мільярдаў адзінак, многія з іх прадстаўленыя ў Інтэрнэце. Стварэнне беларускага корпуса нават у 1 млн словаўжыванняў, тое над чым у апошнія 5 год працуюць беларускія навукоўцы, і яго з'яўленне ў Інтэрнэце – крок хаця і вельмі запаздалы, але неабходны. Беларуская корпусная лінгвістыка адстае ад сусветных тэндэнцый развіцця вельмі істотна, але не назаўсёды. Што дазваляе глядзець наперад з аптымізмам?

Уласцівая корпуснай лінгвістыцы падтрымка інфармацыйнымі і праграмнымі сродкамі дазволяе ў параўнальна кароткі тэрмін дасягнуць сусветных стандартаў у праграмным забеспячэнні корпусных даследаванняў нават з нуля. Усё дастаткова проста: навукова-тэхнічны прагрэс у геаметрычнай прагрэсіі скарачае дыстанцыю паміж віртуальнай і матэрыяльнай рэчаіснасцю. Глабалізацыя, у тым ліку ў лінгвістычнай сферы дазваляе ўжо сёння вырашаць амаль любыя пытанні праграмнага забеспячэння пры мінімальных выдатках: многае ўжо зроблена і складзеныя папярэднікамі праграмы нецяжка прыстасаваць ці дапрацаваць. Выкарыстанне замежных распрацовак і набыццё пэўных ліцэнзійных прадуктаў дазволіла, напрыклад, польскім навукоўцам навучыцца эфектыўнаму выкарыстанню корпусных метадаў і выпрацаваць уласную стратэгію развіцця корпуснай лінгвістыкі на польскім матэрыяле.

У бліжэйшым асяроддзі, у Расіі і на Украіне, развіццю корпуснай лінгвістыкі ўдзяляецца належная ўвага, і шмат чаго, у выніку, зроблена. Асаблівай павагі заслугоўвае распрацоўка і ўдасканаленне Нацыянальнага корпуса рускай мовы (Национального корпуса русского языка) [6]. Названы праект знаходзіцца на ўзроўні замежных узораў.

Акрамя праграмных сродкаў развіццё корпуснай лінгвістыкі патрабуе эмпірычных матэрыялаў: слоўнікаў і тэкстаў. Акадэмічныя слоўнікі сучаснай беларускай мовы ёсць, але для корпусных распрацовак патрабуецца іх прадстаўленасць на машынных носьбітах і максімальна спрошчаны доступ неабмежаванага кола навукоўцаў і проста патэнцыяльных карыстальнікаў да адпаведных электронных рэсурсаў. Пакуль што слоўнік-эталон беларускай мовы

немагчыма знайсці ў Інтэрнэце. У «сусветным павуцінні», з іншага боку, можна знайсці аматарскія распрацоўкі і прататыпы слоўнікаў беларускай мовы рознага кшталту. У кожнага карыстальніка беларускамоўнага Інтэрнэта ў выніку – свой слоўнік: на базе тарашкевіцы ці трасянкі, нярэдка на лацініцы. Адсутнасць (у электронным выглядзе і свабодным доступе) нарматыўных слоўнікаў, на жаль, перашкаджае стварэнню, напрыклад, Інтэрнэт-рэсурсаў адпаведных крытэрыям нацыянальнага корпуса. Наяўнасць жа ў Інтэрнэце ненарматыўнай беларускай мовы і адсутнасць электроннага эталона, аказвае проста катастрафічны ўплыў на стан беларускай гутарковай мовы і ў далейшым, між іншым, звузіць магчымасці выкарыстання сённяшней беларускай мовы Інтэрнэта як крыніцы любога корпуса ўзуальнай беларускай мовы.

Збіранне тэкстаў на беларускай мове, у прынцыпе, не ўяўляе сабой невырашальнай задачы. Ёсць пытанне аўтарскага права. Але, па-першае, абмежаванні аўтарскага права ўжо амаль не закранаюць класіку беларускай літаратуры, па-другое, можа і не трэба імкнуцца спяшацца творы спрэчнай якасці ўключаць у прадстаўнічы корпус, калі аўтар супраць. Ёсць архівы перыядычных выданняў, спецыяльная літаратура. Ужо сёння праграма пошуку па беларускіх тэкстах, размешчаных, напрыклад, на сайце газеты «Звязда» (www.zvyazda.minsk.by) альбо на «Беларускай палічцы» (www.knihi.com), дае тыя ж магчымасці, што і любы неразмечаны (неанатаваны) камп'ютарны корпус, а значыць, можа выкарыстоўвацца для разнастайных лінгвістычных мэтаў [3, 4]. Уключэнне надрукаваных у дзяржаўных сродках масавай інфармацыі і выдавецтвах артыкулаў і мастацкіх твораў, відавочна, пры выкарыстанні такіх тэкстаў у нацыянальным корпусным праекце не будзе выклікаць асаблівых цяжкасцей з пункту гледжання аўтарскага права.

Дарэчы, маюцца прыклады паспяховай працы беларускіх даследчыкаў над корпусным праектам. Корпус, створаны ў навукова-даследчай лабараторыі інтэлектуальных інфармацыйных сістэм на факультэце прыкладной матэматыкі Белдзяржуніверсітэта (загадчык лабараторыі доктар фізіка-матэматычных навук прафесар І. В. Соўпель) па заказе Міністэрства інфармацыі, задумваўся як паралельны руска-беларускі корпус матэрыялаў пасяджэнняў Палаты прадстаўнікоў Нацыянальнага Сходу Рэспублікі Беларусь, заканадаўчых актаў і пад., а рэальна, пасля пашырэння зыходнай тэкставай базы, стаў рэпрэзентатыўным корпусам беларускай мовы, з якога пры дапамозе сродкаў камп'ютарнай падтрымкі лёгка вылучаюцца субкорпусы, прыдатныя для розных прыватных даследчых мэтаў. Хочацца спадзявацца, што названы корпус калі-небудзь з'явіцца ў Інтэрнэце.

Дзейнасць па стварэнні ці кампіляцыі корпуса тэкстаў вельмі разнастайная. Крытэрый адбору тэкстаў для свайго корпуса стваральнік задае, зыходзячы з мэтаў сваёй практычнай ці навуковай дзейнасці: паказчыкам рэпрэзентатыўнасці для яго будзе служыць патрабаванне максімальна адлюстраваць у створаным корпусе вывучаемую з'яву. Нават аб'яднанне ў адным файле адабраных па пэўных крытэрыях тэкстаў адбываецца, як правіла, з улікам папярэдне вызначанай мэты,

што немагчыма без усведамлення металінгвістычных умоў стварэння і фунцыяніравання абраных матэрыялаў.

Экстралінгвістычны аспект корпусных распрацовак мае практычнае значэнне, аказваючы сур'ёзны ўплыў на даследчы працэс у корпусным рэчышчы. Паводле крытэрыяў адбору тэкстаў і рэпрэзентатыўнасці корпусы распадаюцца на два класы. У першы клас уваходзяць корпусы тэкстаў, мэта якіх – адлюстраваць аб'ектыўную карціну маўленчай дзейнасці. Толькі суцэльны, без суб'ектыўнага адбору корпус, напрыклад, нацыянальны корпус, дазваляе атрымаць поўную, сістэмную карціну стану мовы. Метадалогія пабудовы корпусаў першага тыпу так ці інакш грунтуецца на прынцыпе дэдукцыі – ад агульнага (аб'ектыўнай моўнай практыкі носьбітаў мовы) да абмежаванага корпуса тэкстаў, які адлюстроўвае гэта агульнае.

У другі клас уваходзяць корпусы, пабудаваныя для адлюстравання пэўнай з'явы. У гэтым выпадку прыходзіцца задаволіцца тымі ці іншымі выбаркамі з агульнай сукупнасці моўных дадзеных. Але гэтыя выбаркі павінны адпавядаць як універсальна-статыстычным, так і спецыяльна-прадметным крытэрыям якаснай прадстаўнічасці выбаркі для лінгвістычнага даследавання. Па гэтай прычыне метадалогія пабудовы корпусаў другога класа павінна займацца праблемай карэктнасці адлюстравання асобнага лінгвістычнага феномена ў корпусе тэкстаў, прызначаным адлюстраваць гэты феномен.

Спіс корпусаў, створаных замежнымі лінгвістамі для пэўнай прагматычнай мэты, проста бясконцы, але большая частка з іх, што характэрна, пазней выкарыстоўваецца для мэтаў, больш шырокіх за першапачатковыя.

Поўныя ці рэпрэзентатыўныя моўныя корпусы і корпусы, якія ствараюцца для прыватных лінгвістычных задач, рэдка існуюць незалежна – як правіла, мэты корпусаў першага і другога тыпу сінкрэтычныя. З гэтага лагічна вынікае відавочная і важная выснова аб выкарыстанні ўжо акумуляваных рэсурсаў. Нават калі для пэўнай мовы (напрыклад, беларускай) яшчэ не створаны прадстаўнічы агульнадаступны корпус, то прыватныя даследчыя корпусы трэба рабіць агульнадаступнымі, максімальна моўна нарматыўнымі, і найлепшы сродак для гэтага – Інтэрнэт. З іншага боку, агульнадаступнымі, вельмі пажадана, павінны быць нарматыўныя слоўнікі ў электронным выглядзе.

Пад рэпрэзентатыўнасцю разумеецца, у тым ліку, здольнасць корпуса тэкстаў адлюстроўваць усе ўласцівасці праблемнай вобласці, рэлевантныя для дадзенага тыпу лінгвістычнага даследавання, у пэўнай прапорцыі, якая вызначаецца частатой той ці іншай з'явы ў праблемнай вобласці. Гэта патрабаванне арыентуе «збіральнікаў» корпусаў тэкстаў на спецыялізацыю распрацаванага прадукта па ўзроўневым прынцыпе: фанетычныя, марфалагічныя, сінтаксічныя, лексічныя, тэкставыя і інш. корпусы [1]. Што, дарэчы, не выключае агульную арыентацыю моўнага матэрыяла на вызначаныя стандарты.

Рэпрэзентатыўнасць любога корпуса, у прынцыпе, не абмяжоўваецца ўласналінгвістычнымі параметрамі. Так, у кожным канкрэтным выпадку можа ўзнікнуць патрэба ўлічыць стылістычны, часавы, аўтарскі і іншыя элементы

тэкставага масіву праблемнай вобласці. Экстралінгвістычныя характарыстыкі тэкстаў корпуса – своеасаблівая абалонка корпусных дадзеных. Менавіта ўлік такіх экстралінгвістычных фактараў на папярэднім этапе дазваляе дакладна фармуляваць задачы складальнікам корпусаў. Складаныя ўмовы функцыяніравання сучаснай беларускай літаратурнай мовы проста абавязваюць улічваць экстралінгвістычны кантэкст рэпрэзентатыўнасці беларускага корпуса тэкстаў.

Дакладна спраектаваная рэпрэзентатыўнасць ператварае або не ператварае набор тэкстаў на машынным носьбіце ва ўнікальнае слоўнае адзінства – корпус тэкстаў. Гэта ўласцівасць корпуса настолькі важная, што часам гавораць пра рэпрэзентатыўнасць як пра вынік працэсу самаарганізацыі корпуса, разглядаемы, безумоўна, як метафара. Тады, па ідэі, на якой грунтуецца корпусная лінгвістыка, корпус тэкстаў адлюстроўвае аб'ектыўную карціну маўленчай дзейнасці незалежна ад жадання яго стваральніка [7; 8].

Такім чынам, для паспяховага развіцця беларускай корпуснай лінгвістыкі ёсць не толькі перашкоды ў выглядзе спазнення на пяцьдзесят год і ўнікальнага гістарычнага і функцыянальнага рознагалосся мовы, але і магчымасці ўлічыць усе памылкі і выкарыстаць усе дасягненні складальнікаў корпусаў іншых моў. Свядомы ўлік экстралінгвістычнага кантэкста сённяшнега развіцця беларускай мовы дазволіць паспяхова вырашаць пытанні рэпрэзентатыўнасці не толькі будучага нацыянальнага корпуса беларускай мовы, але і шматлікіх спецыялізаваных корпусаў, стварэнне якіх з дапамогай сучасных тэхналогій хутка можа стаць рэчаіснасцю лінгвістычных даследаванняў у Беларусі.

СПІС ЛІТАРАТУРЫ

1. Баранов, А. Н. Введение в прикладную лингвистику / А. Н. Баранов. – М.: Эдиториал УРСС, 2001. – 358 с.
2. Беларуская мова ў Інтэрнэт [Электронны рэсурс]. – Рэжым доступу: http://mova.by.ru/. – Час доступу: 10.08.2010.
3. Беларуская палічка [Электронны рэсурс]. – Рэжым доступу: www.knihi.com. – Час доступу: 10.08.2010.
4. Звязда [Электронны рэсурс]. – Рэжым доступу: www.zvyazda.minsk.by. Час доступу: 10.08.2010.
5. Компьютерный фонд белорусского языка и перспективы создания белорусского лингвистического портала / Н. К. Рубашко, Г. П. Невмержицкая, И. В. Совпель // Слово и словарь. Vocabulum et vocabularium: сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2007. – С. 44–46.
6. Национальный корпус русского языка [Электронны рэсурс]. – Рэжым доступу: www.ruscorpora.ru. – Час доступу: 10.08.2010.
7. Holmes-Higgin, P. Assembling and Viewing a Corpus of Texts: Self-organisation, Logical Deduction and Spreading Activation as Metaphors / P. Holmes-Higgin, K. Ahmad // Euralex'96 Proceedings. – Stockholm: Stockholm Univ. Press, 1996. – P. 47–91.

8. McEnery, T. Corpus Linguistics / T. McEnery, A. Wilson [Electronic resource]. – Edinburgh: Edinburgh Univ. Press, 1999. – Mode of access: http://www.ling.lancs.ac.uk/staff/andrew/data.htm. – Date of access: 10.08.2006.