УДК-056.262; 316.772.5; 342.725

РАСПРАЦОЎКА І РАЗВІЦЦЁ ЛІНГВІСТЫЧНАЙ БАЗЫ ВЕДАЎ ЮРЫДЫЧНАЙ ТЭМАТЫКІ ДЛЯ СІСТЭМ МАШЫННАГА ПЕРАКЛАДУ І СІНТЭЗУ ВУСНАГА МАЎЛЕННЯ

Ю. С. Гецэвіч[1], В. В. Варановіч[2], А. У. Бабкоў[1], М.В. Супрунчук[3]
[1]Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск;
[2]Беларускі дзяржаўны ўніверсітэт, Мінск;
[3]Мінскі дзяржаўны лінгвістычны ўніверсітэт, Мінск

Падымаецца пытанне даступнасці юрыдычных тэкстаў на дзвюх дзяржаўных мовах для грамадзян розных сацыяльных груп, у тым ліку людзей з інваліднасцю па зроку. Увага акцэнтуецца на кодэксах Рэспублікі Беларусь. Паказваюцца звязаныя з распаўсюджваннем інтэрнэту змяненні ў адносінах грамадства да людзей, якія слаба бачаць. Звяртаецца ўвага на комплекснасць паняцця «даступнасць заканадаўчых тэкстаў» у грамадстве; разглядаюцца магчымасці машыннага перакладу і выкарыстання партала corpus.by у мэтах аўтаматызацыі лінгвістычнай вычыткі і апрацоўкі тэкстаў. Паказваецца працэс падрыхтоўкі адпаведнага тэкставага матэрыялу для стварэння электроннага двухмоўнага корпусу юрыдычнай тэматыкі на беларускай і рускай мовах.

Сучасная моўная сітуацыя ў Рэспубліцы Беларусь характарызуецца найперш як дзяржаўнае двухмоўе: на заканадаўчым узроўні замацаваны дзве дзяржаўныя мовы – руская і беларуская. У сітуацыі білінгвізму вельмі важным з'яўляецца забеспячэнне даступнасці тэкстаў рознага прызначэння на абедзвюх дзяржаўных мовах. Даступнасць заканадаўства для носьбітаў розных моў павінна стаць прыярытэтнай задачай для навукоўцаў розных галін – юрыстаў, лінгвістаў, спецыялістаў у інфармацыйных тэхналогіях. Паняцце "даступнасць заканадаўства" на сённяшні дзень атрымала асэнсаванне і ўжо звязваецца не толькі з якасцю вербальнага адлюстравання нормы, але і з публічнасцю нарматворчай і правапрымяняльнай дзейнасці, сістэматызацыяй заканадаўства, афіцыйным растлумачэннем зместу нарматыўных актаў [1].

У розных нарматыўных актах даступнасць узгадваецца ў кантэксце патрабаванняў да якасці іх моўнага выяўлення. Напрыклад, у Законе Рэспублікі Беларусь ад 17 ліпеня 2018 г. «Аб нарматыўных прававых актах» у ліку патрабаванняў да мовы дакументаў пазначаны і такія, як яснасць, прастата і даступнасць (URL: https://pravo.by/document/?guid=3871&p0=H11800130). Забеспячэнне даступнасці і зразумеласці для кожнага грамадзяніна – гэта фактычна канчатковая мэта працэсу ўдасканалення заканадаўства ў любой прававой дзяржаве. Аднак дадзеная мэта заўсёды будзе заставацца недасягальнай без уліку моўнай сітуацыі ў краіне, без арыентавання на кожнага адрасата прававой інфармацыі як на моўную асобу. Інакш кажучы, якім бы ні быў заканадаўчы тэкст дасканала апрацаваным з пункту гледжання дакладнасці, лагічнасці, яснасці і г. д., як бы добра ні былі забяспечаны ўсе неабходныя ўмовы для азнаямлення з ім, даступнасць тэксту ўсё роўна не будзе поўнай, калі ён створаны і існуе толькі на адной мове ў білінгвальным грамадстве. У Рэспубліцы Беларусь, нягледзячы на дзяржаўнае двухмоўе, пераважная большасць заканадаўчых дакументаў рэалізавана толькі на адной мове. Так, з 26 кодэксаў Рэспублікі Беларусь, тэксты якіх прадстаўлены на Нацыянальным прававым інтэрнэт-партале pravo.by, 25 афіцыйна прынятыя толькі па-руску, 1 – толькі па-беларуску (URL: http://pravo.by/pravovaya-informatsiya/normativnye-dokumenty/kodeksy-respubliki-belarus/). У 2019 годзе пры Нацыянальным цэнтры прававой інфармацыі быў створаны экспертны савет па перакладзе заканадаўчых актаў на беларускую мову (URL:

https://pravo.by/novosti/novosti-pravo-by/2019/november/42031/). На сённяшні дзень на партале pravo.by размешчаны пераклады на беларускую мову 10-ці кодэксаў Рэспублікі Беларусь (URL: pravovaya-informatsiya/normativnye-dokumenty/kodeksy-respubliki-belarus/). Варта адзначыць, што ўсе прадстаўленыя пераклады маюць пазнаку “неафіцыйны пераклад”.

Акрамя праблемы даступнасці заканадаўства на дзвюх дзяржаўных мовах Рэспублікі Беларусь, існуе праблема даступнасці тэкстаў для невідушчых і слабабачачых людзей. У Беларусі на 1 жніўня 2018 г. пражывала 567,5 тыс. чалавек з інваліднасцю, зарэгістраваных у органах па працы, занятасці і сацыяльнай абароне, што складае 6 % ад агульнай колькасці насельніцтва рэспублікі. Калі дадаць да гэтага 1,9 млн беларусаў старэйшых за 60 гадоў, большасць з якіх слаба бачыць, то стане відавочна, што дадзеная праблема датычыцца чвэрці жыхароў Беларусі.

Раней людзі, якія слаба бачаць, былі амаль цалкам ізаляваныя ад грамадства. Цяпер грамадства пачынае разумець: людзі з інваліднасцю па зроку не існуюць асобна, а з’яўляюцца яго часткай. Камп’ютар дазваляе ім не замыкацца ў сваім вузкім асяроддзі. 18 кастрычніка 2016 г. Беларусь ратыфікавала міжнародную Канвенцыю аб правах інвалідаў, у артыкуле 9 якой так гаворыцца аб інтэрнэт-даступнасці: «Дзяржавы-ўдзельніцы… прымаюць належныя меры для таго, каб:

развіваць іншыя… формы аказання дапамогі і падтрымкі людзям з інваліднасцю для забеспячэння ім доступу да інфармацыі;

спрыяць доступу людзей з інваліднасцю да новых інфармацыйна-камунікацыйных тэхналогій і сістэм, у тым ліку да Інтэрнэту;

садзейнічаць праектаванню, распрацоўцы, вытворчасці і распаўсюджванню даступных інфармацыйна-камунікацыйных тэхналогій, каб даступнасць такіх тэхналогій і сістэм дасягалася з мінімальнымі выдаткамі» (URL: http: //mintrud.gov.by/ru/new_url_369854369).

4 снежня 2017 г. Беларусь у рамках Праграмы развіцця ААН таксама падпісала Інфармацыйную стратэгію Беларусі па паўнапраўным уключэнні (інклюзіі) людзей з інваліднасцю ў грамадства.

А з 1 студзеня 2019 г. уступіла ў сілу пастанова Савета міністраў Рэспублікі Беларусь № 797 ад 23 кастрычніка 2017 г., паводле якой унесены змяненне і дапаўненні ў Палажэнне аб парадку функцыянавання інтэрнэт-сайтаў дзяржаўных органаў і арганізацый (URL: http://etalonline.by/document/?regnum=c21700797&q_id=643794). Удакладнена, што інфармацыя на інтэрнэт-сайтах размяшчаецца на рускай і (або) беларускай мовах, а пры неабходнасці таксама на адной або некалькіх замежных мовах. Прымяненне і беларускай, і рускай моў патрэбнае пры размяшчэнні ўсёй інфармацыі, якая з’яўляецца абавязковай для галоўнай старонкі інтэрнэт-сайта. Пры гэтым інтэрнэт-сайт павінен падтрымліваць версію для людзей з інваліднасцю па зроку і быць сумяшчальным з рознымі інтэрнэт-браўзерамі.

Так, з мэтай забеспячэння агульнадаступнасці інфармацыі былі распрацаваны версіі інтэрнэт-сайта 4-й Гарадской клінічнай бальніцы імя М.Я. Саўчанкі на рускай і беларускай мовах, і абедзве версіі падтрымліваюць функцыянальнасць для людзей з інваліднасцю па зроку, якая дазваляе: уключаць ці выключаць галасавыя падказкі, што прагаворваюць назвы раздзелаў інтэрнэт-сайта, кнопак і пунктаў меню; агучваць навіны па-беларуску і па-руску з дапамогай усталяванай сістэмы сінтэзу маўлення; а таксама выбіраць камфортныя для ўспрымання памер шрыфту і каляровую палітру (URL: https://4gkb.by).

Адзін з асноўных фактараў, якія стрымліваюць практычнае забеспячэнне двухмоўя ў прававой сферы Рэспублікі Беларусь, – нявырашанасць праблемы якаснай і

хуткай лінгвістычнай апрацоўкі тэкстаў вялікага памеру, што сведчыць аб актуальнасці якаснага машыннага перакладу.

Атрымаць неафіцыйную беларускамоўную версію рускамоўнага нарматыўнага акта або неафіцыйную рускамоўную версію беларускамоўнага нарматыўнага акта сёння магчыма, калі звярнуцца на сайт Нацыянальнага цэнтра прававой інфармацыі Рэспублікі Беларусь (URL: http://etalonline.by/) і націснуць кнопку «Машынны пераклад». Пераклад ажыццяўляецца за некалькі хвілін, а яго якасць дазваляе ў цэлым зразумець дакумент, але звяртае на сябе ўвагу сэнсавая нераўнацэннасць лексічных адзінак або сінтаксічных канструкцый зыходнай рускай мовы і беларускай мовы, напрыклад (руская лексема – прапанаваная машынным перакладчыкам лексема – правільны аналаг): *хотя – хочучы – хоць; безвестно отсутствующий – невядома адсутны – адсутны без вестак; управлениями по труду – кіраваннямі па працы – упраўленнямі па працы; в противном случае – у агідным выпадку – у адваротным выпадку.*

Менавіта таму пастрэдагаванне тэкстаў лінгвістамі – гэта неабходная праца, якая можа быць аптымізавана пры дапамозе праграмнага забеспячэння (URL: http://corpus.by/), распрацаванага супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі (АІПІ) НАН Беларусі. У прыватнасці, пры перакладзе тэкстаў кодэксаў Рэспублікі Беларусь карыснымі з'яўляюцца сістэмы праверкі правапісу, статыстычнай і даведачнай інфармацыі аб сімвалах, статыстыкі ўжывання адвольных сімвальных паслядоўнасцей у электронным тэксце, пошуку і выпраўлення памылак у напісанні літар «у» і «ў», распазнавання і вылучэння ў тэксце амографаў [2].

Машынны перакладчык можа быць удасканалены, калі ён будзе мець магчымасць пастаянна напрацоўваць свае навыкі ў адной і той жа прадметнай галіне, грунтуючыся на правілах, якія выбіраюцца з зададзеных крыніц інфармацыі. А з удасканаленнем машыннага перакладчыка пачне вырашацца і праблема забеспячэння даступнасці заканадаўства для носьбітаў розных моў (у тым ліку сярод людзей з інваліднасцю па зроку) – адна з галоўных задач, што вырашаюцца ў кожнай дзяржаве ў мэтах інфармацыйнай бяспекі [1].

У лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі існуе шэраг распрацаваных сінтэзатараў беларускага і рускага маўлення, сярод якіх стацыянарны [3], мабільны і анлайн-сінтэзатар [4], што дазваляе людзям з інваліднасцю па зроку праз аўтаматычнае агучванне тэксту сінтэзатарам праслухоўваць патрэбны ім кантэнт на розных прыстасаваннях.

Дадзеныя праграмы таксама выкарыстоўваюцца для працы з юрыдычным даменам беларускай лексікі пры перакладзе з рускай мовы на беларускую і далейшым агучванні кодэксаў Рэспублікі Беларусь. Распрацавана канцэпцыя сістэмы машыннага перакладу, якая выкарыстоўвае нейронныя сеткі. У параўнанні з сістэмамі, заснаванымі на правілах і статыстыцы, сістэмы з нейроннымі сеткамі даюць значна лепшы вынік і якасць перакладзенага тэксту. На сённяшні дзень сістэмы з нейроннымі сеткамі для перакладу прымяняюць найбольш папулярныя сярод карыстальнікаў парталы Google, Яндэкс, Промт, Deeplearning і інш. Недахоп названых сістэм – выкарыстанне мовы-пасярэдніка (англійскай). Безумоўна, стварэнне асобнай сістэмы машыннага перакладу для пэўнай моўнай пары – больш прадуктыўны шлях для якаснай працы аўтаматычнага перакладу. І складанне корпуса паралельных тэкстаў для моўнай пары – неабходная ўмова пры стварэнні сістэмы машыннага перакладу, заснаванай на нейронных сетках. Такім чынам, для забеспячэння руска-беларускага перакладу патрэбна пэўная колькасць эталонных тэкстаў на дзвюх мовах, на якіх будзе навучацца нейронная сетка.

З гэтай мэтай у Лабараторыі распазнавання і сінтэзу маўлення ствараецца двухмоўны корпус кодэксаў Рэспублікі Беларусь (зацверджаны кодэкс на адной з дзвюх дзяржаўных моў з перакладам на другую мову). Такі корпус забяспечыць машыннае навучанне нейроннай сеткі найперш для аўтаматычнага перакладу тэкстаў афіцыйна-справавога стылю.

Пераклад кожнага кодэкса праходзіць некалькі этапаў:

1) Пераклад з дапамогай сістэмы машыннага перакладу на сайце НЦПІ.

2) Вычытка атрыманага перакладу пры дапамозе праграмнага забеспячэння, прадстаўленага на платформе corpus.by:

- праверка правапісу,
- аналіз статыстычнай і даведачнай інфармацыі аб сімвалах,
- аналіз статыстыкі ўжывання адвольных сімвальных паслядоўнасцей у электронным тэксце (падлік частотнасці словаформ),
- пошук і выпраўленне памылак у напісанні літар «у» і «ў»,
- распазнаванне і вылучэнне ў тэксце амографаў.

3) Ручная вычытка тэксту і выпраўленне памылак паслядоўна чатырма рэдактарамі, сярод якіх ёсць прафесійныя лінгвісты і юрысты. У працэсе вычыткі складаецца слоўнік найбольш частых (рэгулярных) замен слоў і словазлучэнняў. Такі кантэксталагічны слоўнік таксама плануецца выкарыстаць пры навучанні сістэмы машыннага перакладу для павышэння якасці перакладзенага тэксту.

У цяперашні час вядзецца праца над стварэннем корпусу перакладзеных і размешчаных у агульным доступе кодэксаў (URL: https://ssrlab.by/7804). Прымаюцца заўвагі і ўдасканальваюцца пераклады, рыхтуюцца матэрыялы для адсылкі корпусаў зацікаўленым арганізацыям і навукоўцам для дапрацоўкі. Акрамя перакладу, тэксты кодэксаў на беларускай і рускай мовах агучваюцца мадэрнізаваным анлайн-сінтэзатарам BarysBelHigh, які дае магчымасць атрымліваць больш чыстае гучанне за кошт выдалення часткі басовага спектра.

Падчас працы над перакладам кодэксаў былі выяўлены некаторыя асаблівасці беларускай граматыкі (у параўнанні з рускай), якія цяжка алгарытмізаваць у сістэме машыннага перакладу, напрыклад:

– розныя канчаткі назоўнікаў мужчынскага роду ў родным склоне (*-а* ці *-у*), прычым ужыванне аднаго з варыянтаў залежыць ад семантыкі слова: *рахунка* (у значэнні ‘*дакумент*’) і *рахунку* (у значэнні ‘*фінансавая аперацыя*’);

– граматычнае афармленне канструкцый са ступенямі параўнання прыметнікаў і прыслоўяў: *не ранее одного года – не раней за адзін год*;

– дзеепрыметнікі, дзеепрыметныя звароты часта ў беларускай мове перадаюцца даданымі сказамі: *организация, осуществляющая эксплуатацию жилищного фонда – арганізацыя, якая ажыццяўляе эксплуатацыю жыллёвага фонду;*

– анафарычныя сувязі, пры якіх машынны пераклад выкарыстоўвае памылковы род слова-заменніка, напрыклад: *пассажир – физическое лицо, имеющее проездной документ – пасажыр – фізічная асоба, *якое мае праязны дакумент; Решение принимает государство. Оно определяет порядок… – Рашэнне прымае дзяржава. * Яно вызначае парадак…*Такая ж праблема ёсць і з катэгорыяй ліку, напрыклад, са словамі *отношение – адносіны*, *зерновые – збожжа*.

На ўзроні лексікі пэўную цяжкасць для перакладу выклікаюць некаторыя дзеясловы, напрыклад: *устанавливать, подвергать, предоставлять*. Аднаслоўныя адпаведнікі гэтых дзеясловаў (*устанаўліваць, падвяргаць, прадастаўляць*), якія падаюцца ў слоўніках, нехарактэрныя для беларускай лексічнай сістэмы, таму варта шукаць іншыя варыянты перакладу сказаў з такімі дзеясловамі. Акрамя таго, слоўнікі

сучаснай беларускай мовы часам не дыферэнцыруюць некаторыя тэрміны, падаючы аднолькавы пераклад для розных лексем рускай мовы, у той час як у заканадаўстве гэтыя тэрміны маюць адрозненне ў значэнні: *збудаванне* – адпаведнік для слоў *строение* і *сооружение, пазыка – заем* і *ссуда, утрыманне – содержание* і *иждевение*, у той час як у Грамадзянскім кодэксе сустракаецца выраз *пожизненное содержание гражданина с иждивением*. Для запаўнення такіх лексічных лакун неабходна сумесная праца юрыстаў і лінгвістаў, а таксама зварот да старажытных помнікаў юрыдычнай літаратуры на старабеларускай мове.

7 жніўня 2018 г. паміж АІПІ НАН Беларусі і Нацыянальным цэнтрам прававой інфармацыі Рэспублікі Беларусь быў заключаны дагавор аб супрацоўніцтве ў галіне навуковых распрацовак па развіцці і ўдасканаленні інфармацыйных тэхналогій і сістэм, нацыянальных інфармацыйных рэсурсаў і іх рацыянальным выкарыстанні, у тым ліку па выпрацоўцы карпусоў беларускай, рускай і англійскай юрыдычнай лексікі з дапамогай партала corpus.by. Лабараторыя распазнавання і сінтэзу маўлення плануе удзельнічаць у развіцці і напаўненні дзяржаўных інфармацыйна-прававых рэсурсаў, прапаноўваць для тэсціравання і наступнай практычнай эксплуатацыі сістэмы сінтэзу і распазнавання маўлення па тэксце, сістэмы машыннага перакладу тэкстаў юрыдычнага і медыцынскага дамена.

Такім чынам, ва ўмовах новага інфармацыйнага асяроддзя чалавечай цывілізацыі – інфасферы, якая цягне за сабой радыкальныя сацыяльныя змены і істотным чынам датычыцца жыцця практычна кожнага грамадзяніна, – у Беларусі робяцца вельмі сур’ёзныя і патрэбныя крокі ў бок даступнасці юрыдычных тэкстаў. Значную ролю пры гэтым адыгрывае забеспячэнне даступнасці інфармацыі для безбар’ернай навігацыі па заканадаўстве Рэспублікі Беларусь усіх катэгорый моўных носьбітаў незалежна ад іх асаблівасцей.

Спіс літаратуры

1. Гецэвіч, Ю.С. Выкарыстанне сістэм машыннага перакладу і сінтэзу маўлення для забеспячэння даступнасці заканадаўчых тэкстаў на розных мовах у Рэспубліцы Беларусь / Ю.С. Гецэвіч, А.А. Кірдун // Информационные технологии и право (Правовая информатизация – 2018) : материалы VI Междунар. науч.-практ. конф., Минск, 17 мая 2018 г. / под общ. ред. Е.И. Коваленко ; Нац. центр правовой информ. Респ. Беларусь. – Минск, 2018. – С. 123–128.

2. Гецэвіч, Ю.С. Вычытка і генерацыя тэкстаў вялікага памеру на беларускай мове / М.У. Марчык, Г.Р. Станіславенка, С.І. Лысы, Ю.С. Гецэвіч // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2017) : доклады XVI Международной конференции, Минск, 16 ноября 2017 г. / ОИПИ НАН Беларуси ; под науч. ред. А.В. Тузиков, Р.Б. Григянец, В.Н. Венгеров. — Минск : ОИПИ НАН Беларуси, 2017. — С. 305-310.

3. Гецевич, Ю.С. Система синтеза белорусской речи по тексту / Ю.С. Гецевич, Б.М. Лобанов // Речевые технологии. – 2010. – № 1. – С. 91–100.

4. Гецэвіч, Ю.С. Распрацоўка сінтэзатара беларускага і рускага маўленняў па тэксце для мабільных і інтэрнэт-платформаў / Ю.С. Гецэвіч, Д.А. Пакладок, Д.В. Брэк // Развитие информатизации и государственной системы научно-технической информации (РИНТИ–2012) : докл. XI Междунар. конф., Минск, 15 нояб. 2012 г. – Минск : ОИПИ НАН Беларуси, 2012. – С. 254–259.