УДК 81'33 + 004.6

ГРАФІЧНЫЯ МАРКЁРЫ ДЛЯ АЎТАМАТЫЗАВАНАЙ ІДЭНТЫФІКАЦЫІ ЎВАХОДЖАННЯ БЕЛАРУСКАМОЎНЫХ ФРАГМЕНТАЎ У ЗМЕШАНЫ БЕЛАРУСКА-РУСКІ ТЭКСТ[1]

А.Ю. СТАНКЕВІЧ
(Гродзенскі дзяржаўны ўніверсітэт імя Я. Купалы)
канд. філал. навук, дац. І.І. БУБНОВІЧ
(Гродзенскі дзяржаўны ўніверсітэт імя Я. Купалы)

На падставе лінгвастатыстычнага аналізу вызначаецца комплекс графічных маркёраў-ідэнтыфікатараў уключэнняў беларускамоўных фрагментаў у змешаны беларуска-рускі тэкст. Прыводзяцца алгарытм вызначэння тэставых масіваў для рускай і беларускай моў, а таксама схемы графічных маркёраў. Вызначаны комплекс графічных маркёраў можа быць выкарыстаны для аўтаматызацыі разметкі беларускамоўных украпленняў пры стварэнні паўнатэкставых электронных моўных рэсурсаў.

Ключавыя словы: *электронныя моўныя рэсурсы, корпусныя тэхналогіі, разметка, лінгвістычнае забеспячэнне, графічны маркёр, беларуска-рускае двухмоўе, беларускамоўныя ўкрапленні, змешаны беларуска-рускі тэкст.*

Уводзіны. Традыцыйна лічыцца, што на Беларусі з даўніх часоў развіццё беларускай літаратурнай мовы адбываецца ва ўмовах білінгвізму ці шматмоўя. Вынікам гэтага з'яўляюцца шматлікія пазычанні з розных моў. Праблемы вызначэння пазычанняў, асноўных тэндэнцый іх засваення, асаблівасці іх семантычнай, фанетычнай і марфалагічнай адаптацыі, а таксама іншыя пытанні, звязаныя з узаемадзеяннем беларускай мовы з іншымі мовамі, закранаюцца ў працах многіх беларускіх і замежных навукоўцаў. Адным з першых даследаваў асаблівасці беларускай мовы на ўсіх яе ўзроўнях, вызначыў і падаў пералік пазычанняў з розных моў Я.Ф. Карскі [1; 2]. Пытанні білінгвізму, полілінгвізму і ўзаемадзеяння моў на Беларусі ў розныя часы даследаваліся М.Г. Булахавым [3; 4], А.І. Жураўскім [5], А.А. Гіруцкім [6], Г.Ф. Вештарт [7] і інш. У многіх даследаваннях канца XX – пачатку XXI стст. праблемы беларуска-рускага білінгвізму аналізуюцца з пункту гледжання сацыялінгвістыкі і псіхалінгвістыкі. Сацыялагічны аспект узаемадачыненняў даследуецца Ю.Б. Караковым [8], Н.Б. Мячкоўскай [9]. Знешнелінгвістычныя фактары ўлічваюцца пры аналізе моўнай сітуацыі ў Беларусі М.І. Канюшкевіч [10; 11]. Як феномен беларускага грамадства Г.А. Цыхуном разглядаецца "трасянка" [12].

Шмат увагі надаецца навукоўцамі супастаўляльнаму аналізу рускай і беларускай моў, вызначаюцца іх аднолькавыя і адрозныя рысы [13 – 15]. Спецыфіку білінгвізму ў Беларусі, своеасаблівасць моўнай сітуацыі, а таксама функцыянаванне рускай мовы ва ўмовах руска-беларускага двухмоўя, фактары, што спрыяюць узаемапранікненню гэтых блізкароднасных моў, аналізуе В.Д. Старычонак [16]. Праблема ўзаемадзеяння моў закранаецца і ў артыкуле В.П. Маеўскай, Р.С. Сідарэнка [17]. Імі адзначаецца, што «блізкасць моў стварае ілюзію падабенства, а лёгкасць узаемаразумення, адэкватнасць разумення, нягледзячы на наяўнасць у маўленні шматлікіх адхіленняў ад нормаў, гэту ілюзію падтрымліваюць і паглыбляюць» ([17, с. 12]; пераклад наш).

Варта адзначыць, што ў многіх працах прасочваюцца асаблівасці ўзаемадзеяння паміж беларускай і рускай мовамі на розных моўных узроўнях. Так, пытанні фанетычнай інтэрферэнцыі разглядаюца Л.П. Новікавай [18], акцэнталагічныя адметнасці ўсходнеславянскіх моў аналізуюцца Л.М. Вардамацкім [19]. А.А. Мятлюк вызначае асаблівасці маўлення білінгва [20].

Асаблівую цікавасць навукоўцаў выклікае марфалагічны ўзровень, які даследуецца ў шэрагу прац [21; 22], прасочваюцца адрозненні паміж рускай і беларускай мовамі ў працэсе іх гістарычнага развіцця, асаблівая ўвага надаецца фарміраванню адрозненняў на граматычным узроўні [23; 24], звяртаецца ўвага на адметнасці часцін мовы, напрыклад, І.А. Кісялёў аналізуе часціцы [25], Л.Г. Машчэнская даследуе катэгорыю роду назоўнікаў у рускай і беларускай мовах і ўплыў беларускай мовы на маўленне беларусаў, якія гавораць па-руску [26].

Супастаўленне рускай і беларускай моў на лексічным узроўні робіцца многімі даследчыкамі: асаблівасці ўзаемадзеяння і ўзаемаўплыву разглядаюцца С.М. Грабчыкавым [27], А.Я. Міхневічам і А.А. Гіруцкім [28], закранаюцца Б.Ю. Норманам [29]. Аналізу фармавання і станаўлення лексікі ўсходнеславянскіх моў прысвечаны працы У.В. Анічэнкі [30], І.С. Козырава [31]. Сінтаксічныя асаблівасці рускай і беларускай моў вызначаюцца і параўноўваюцца В.І. Баркоўскім [32], М.І. Канюшкевіч [33; 34], П.П. Шубам [35]), Л.М. Чумак [36].

[1] Падрыхтавана ў межах праекта Дзяржаўнай праграмы навуковых даследаванняў «Эканоміка і гуманітарнае развіццё беларускага грамадства» на 2016 – 2020 гг. (дагавор № А70-16 ад 04.01.2016).

Праблеме ўзаемадзеяння беларускай і рускай моў на Беларусі прысвечана праца калектыву аўтараў «Русский язык в Белоруссии» [37], у якой падкрэсліваецца неабходнасць параўнальна-тыпалагічнага апісання нацыянальнай і рускай моў для ажыццяўлення ўсіх прыкладных работ, звязаных з нацыянальна-рускім двухмоўем [37, с. 9], а таксама выяўлення міжмоўнай інтэрферэнцыі [37, с. 10], якая назіраецца ў Беларусі. Пры гэтым звяртаецца ўвага на тое, што «ва ўмовах нацыянальна-рускага двухмоўя <…> склаліся і рэальна існуюць тыя асаблівыя разнавіднасці рускай мовы, галоўнай адрознай рысай якіх з'яўляецца наяўнасць у іх фанетычных, граматычных, лексіка-семантычных і стылістычных падсістэмах пэўнай сукупнасці іншанацыянальных элементаў» ([37, с. 11]; пераклад наш). Навукоўцы вылучаюць «белорусский нациолект русского языка» [37, с. 12]. Аўтарамі даследуецца не толькі рэальная інтэрферэнцыя, але і «патэнцыяльная», пад якой разумеюцца памылкі, якія могуць быць прадбачаны, бо абумоўлены разыходжаннямі кантактуючых моў [37, с. 60]. Выяўляюцца і называюцца прычыны лексічнай інтэрферэнцыі, адзначанай навукоўцамі, і яе вынікі. Асноўнай прычынай называецца поўнае супадзенне лексіка-семантычных і тэматычных груп слоў у рускай і беларускай мовах, пры якім магчымы перанос лексем з адной мовы ў другую пры маўленні на народнай мове.

Такім чынам, даследаванне ўзаемадзеяння беларускай і рускай моў мае працяглую гісторыю, аднак на сённяшні дзень няма комплексных прац па аўтаматызацыі пошуку беларускамоўных украплення ў рускамоўных тэкстах. Адметнасць нашага даследавання заключаецца ў распрацоўцы алгарытму фарміравання тэставых масіваў дадзеных для беларускай і рускай моў, а таксама ў вызначэнні графічных маркёраў для аўтаматызаванай ідэнтыфікацыі ўваходжання беларускамоўных фрагментаў у змешаны тэкст.

Асноўная частка. Мэтай нашага артыкула з'яўляецца вызначэнне мноства графічных маркёраў, наяўнасць якіх дазваляе ажыццяўляць аўтаматызаваную ідэнтыфікацыю ўключэнняў беларускамоўных фрагментаў (словаформаў і іх паслядоўнасцей, у тым ліку і роўных абзацу) у змешаны тэкст. Пад змешаным тэкстам тут разумеем рускамоўны тэкст з беларускамоўнымі фрагментамі, пры гэтым доля рускай мовы можа быць роўнай ці нязначна перавышаць долю беларускага. Такім чынам, наша задача адрозніваецца ад задачы пабудовы гэсэра мовы (language guesser), г.зн. універсальнага дэтэктара мовы тэксту.

У працы мы ўводзім паняцце ***каэфіцыента адрознівальнай сілы маркёра*** (далей – КАС). Мы прынялі наступныя патрабаванні да тэставых масіваў дадзеных, прыдатных для разліку КАС:

– масіў адлюстроўвае сістэму словазмянення адпаведнай мовы;

– у масіве знятыя паўторы амафармаў;

– масіў не ўтрымлівае неаднаслоўных адзінак;

– масіў утрымлівае міжмоўныя амонімы (для пары руская – беларуская мова).

У адпаведнасці з прынятымі патрабаваннямі мы вызначылі наступныя тэставыя масівы:

– для рускай мовы: тэставы масіў аб'ёмам 2 436 182 словаформаў, вызначаны на аснове разгорнутага слоўніка А.А. Залізняка ад М. Хагена [38] (далей – ЗХ);

– для беларускай мовы: тэставы масіў аб'ёмам 1 091 225 словаформаў, вызначаны на аснове лексіка-семантычнай базы Беларускага N-корпуса [39] (далей – БН).

А. Вызначэнне масіву рускамоўных словаформаў

Зыходныя дадзеныя: разгорнуты слоўнік А. А. Залізняка ад М. Хагена (рэдакцыя 2014 г.) [38]. Аб'ём 4 159 394 словаформы для 142 792 лем.

Зыходныя дадзеныя зменены наступным чынам:

– выключаны словаформы, якія не ўжываюцца (адзначаныя ў слоўніку М. Хагена зорачкай);

– выключаны неаднаслоўныя ўваходы (*а ведь, а именно, а не только что* і г.д.);

– выключаны палі з дадзенымі марфалагічнай разметкі і кодамі-ідэнтыфікатарамі словаформаў;

– дададзены амонімы да беларускіх слоў з ***ё*** (да словаформаў *завез, лед, мед* дададзены словаформы з узноўленай літарай ***ё***: *завёз, лёд, мёд* і г. д.);

– дададзены адлюстраваныя ў друкаванай версіі слоўніка А. А. Залізняка [40] фіналі з ***ё*** *(-чьё, -чьём; -вёшенький, …, -вёшенькими; -аёшь(ся), -аёт(ся), -аём(ся), -аёт(ся), -аёте(сь); -вёшь(ся), -вёт(ся),-вём(ся), -вёт(ся), -вёте(сь); - юёшь(ся), -юёт(ся),-юет(ся), -юёте(сь)* і т. п.), частотныя словаформы з ***ё*** *(всё, всём, её, нём, своё, чём, чьё, чьём* і нек. інш.*),* некаторыя частотныя пачатковыя сегменты з ***ё*** (*четырёх-, трёх-, платёжн-* і нек. інш.);

– зняты паўторы амафармаў.

Аб'ём масіву (з міжмоўнымі амафармамі): 2 436 182 словаформаў.

Б. Вызначэнне масіву беларускамоўных словаформаў

Зыходныя дадзеныя: лексіка-граматычная база беларускага N-корпуса (рэдакцыя 2016 г.) [39]. Аб'ём: 1 840 835 словаформаў для 24 417 лем.

Зыходныя дадзеныя зменены наступным чынам:

– выключаны палі з дадзенымі марфалагічнай і акцэнталагічнай разметкі;

– зняты паўторы амафармаў.

Аб'ём масіву (з міжмоўнымі амафармамі): 1 091 225 словаформаў.

Прынцып прызначэння КАС маркёра такі: маркёр уключаем у слоўнік, а яго КАС прыпісваем значэнне 1, калі ipm (instances per million – ‘частата на мільён’) адзінак з гэтым маркёрам на БН большая ці роўная 100 (г. зн. 0,01% ад БН), а на ЗХ роўна 0; маркёр уключаем у слоўнік, а яго КАС прыпісваем значэнне 0,9, калі ipm адзінак з гэтым маркёрам на БН большая ці роўная 100, а на ЗХ менш або роўная 40 (г. зн. 0,004% ад ЗХ). У іншых выпадках мы праводзілі даследаванне правага / левага акружэння маркёра, якое пашыралася на 1 альбо 2 сімвалы ўправа / улева; у слоўнік уключалі атрыманыя ў выніку даследавання акружэння пашыраныя маркёры, якія не сустракаюцца на ЗХ і маюць на БН ipm большую ці роўную 100 (калі пашыраныя маркёры мелі ў сваім складзе літару ***ў*** або ***і,*** ім прызначалі КАС = 1, у іншых выпадках пашыраным маркёрам прызначалі КАС = 0.9). Абагульненне вышэйсказанага гл. у табліцы 1.

Табліца 1. – Схема прызначэння КАС графічным маркёрам

Тып маркёра	Ipm на ЗХ	Ipm на БН	Значэнне КАС
Просты маркёр	= 0	≥ 100	1
Просты маркёр	≤ 40	≥ 100	0.9
Пашыраны маркёр	= 0	≥ 100	1 (для маркёраў з ***і, ў)***
Пашыраны маркёр	= 0	≥ 100	0.9

Пры вызначэнні мноства графічных маркёраў-ідэнтыфікатараў беларускамоўных фрагментаў быў рэалізаваны падыход «ад экспертных ведаў»: прызначэнне КАС з апорай на тэставыя масівы дадзеных маркёрам, якія былі прапанаваны на падставе лінгвістычных крытэрыяў.

Адрозненні паміж беларускай і рускай мовамі праяўляюцца ў тэкстах на розных узроўнях. Ніжэй прыведзены графемы і спалучэнні графем, якія з пункту гледжання эксперта-лінгвіста могуць быць патэнцыяльнымі маркёрамі-ідэнтыфікатарамі; дадзены вынікі аналізу функцыянавання гэтых элементаў на масівах БН і ЗХ. Фармалізаванае апісанне графічных маркёраў, што разглядаюцца ў названым раздзеле, дадзена ў табліцы 2.

Табліца 2. – Схемы графічных маркёраў

№ радка	Схема	КАС
1	2	3
1	АБО (і (419126; 0), ў (146026; 0))	1
2	ПАСЛЯДОЎНАСЦЬ (літара беларускамоўнага алфавіта ў любым рэгістры, апостраф, знак галоснага малой літарай)	1
3	АБО (ёмі (289; 0), ёў (2838; 0)) АБО (інё (129; 0), ікё (108; 0), іё (959; 0))	1
4	АБО (ёбв (144; 0), ёрб (249; 0), ёпв (304; 0), ёпр (105; 0), ёга (107; 0), ёгр (141; 0), ётэ (163; 0), ётч (132; 0), ёа (183; 0)) АБО (ямё (120; 0), дзё (1841; 0), каё (109; 0), длё (265; 0),ынё (109; 0), ылё (137; 0), ялё (567; 0), лаё (134; 0), ваё (176; 0), ысё (120; 0), ыё (3042; 0), 'ё (585; 0), цё (2685; 0))	0.9
5	джаў (119; 0)	1
6	АБО (джаю (246; 0), джае (295; 0), джва (5672; 0), джац (101; 0), джус (169; 0), джэ (653; 0), джы (628; 0), джг (134; 0)) АБО (ярдж (143; 0), падж (230; 0), будж (328; 0), гадж (315; 0), судж (379; 0), годж (223; 0), вудж (127; 0), кодж (109; 0), ладж (1344; 0), пудж (267; 0), вадж (274; 0), тудж (296; 0), рудж (415; 0), водж (223; 0), эндж (284; 0), сюдж (105; 0), кудж (195; 0), седж (383; 0), родж (1188; 0), ждж (295; 0), здж (786; 0), эдж (774; 0))	0.9
7	АБО (дзеў (451; 0), дзі (12888; 0)) АБО (ідз (750; 0), ўдз (381; 0))	1
8	АБО (дзю_ (120; 0), дзел (1835; 0), дзен (1935; 0), дзев (283; 0), дзее (221; 0), дзеш (249; 0), дзею (142; 0), дзея (238; 0), дзюб (289; 0), дзес (652; 0), дзец (371; 0), дзё (1841; 0), дзь (1765; 0), дзя (5660; 0), дзг (136; 0)) АБО (падз (1405; 0), аадз (174; 0), лодз (262; 0), _адз (786; 0), гадз (638; 0), бадз (447; 0), хадз (215; 0), годз (180; 0), ледз (466; 0), вэдз (111; 0), вадз (372; 0), цадз (115; 0), ходз (626; 0), рэдз (275; 0), ундз (271; 0), садз (304; 0), водз (715; 0), эндз (192; 0), аедз (162; 0), дадз (614; 0), ведз (404; 0), андз (613; 0), 'едз (120; 0), седз (231; 0), задз (610; 0), медз (109; 0), родз (586; 0), мадз (332; 0), йдз (204; 0), рдз (1011; 0), ядз (1680; 0), ыдз (1041; 0), удз (2661; 0), бдз (497; 0), ьдз (174; 0), юдз (247; 0), здз (2310; 0), ддз (1476; 0))	0.9
9	шч (25699; 19)	0.9
10	АБО (жы (10182; 0), шы (24825; 0), чы (43714; 0))	1
11	АБО (жэ (3276; 14), шэ (2792; 5), чэ (8472; 0.4))	0.9

Канчатак табліцы 2

1	2	3
12	чоў (635; 0)	1
13	АБО (чорн (138; 0), чос (126; 0), чот (213; 0)) АБО (пячо (178; 0), ашчо (267; 0), _шчо (167; 0))	0.9
14	ллю_ (192; 0)	0.9
15	нняў (697; 0)	1
16	АБО (нням (1256; 0), ннях (618; 0); нню_ (5982; 0)) АБО (чанне (232; 0), зенне (122; 0), энне (1090; 0); жання (129; 0), нення (308; 0), кання (399; 0), чання (169; 0), зання (99; 0), лення (531; 0), тання (219; 0), вання (4121; 0), нання (141; 0), дання (255; 0), яння (111; 0), эння (982; 0); ненню (171; 0), канню (262; 0), ленню (318; 0), танню (131; 0), ванню (3248; 0), данню (121; 0), чанню (117; 0), энню (550; 0))	0.9
17	АБО (цця_ (112; 0); ццю_ (174; 0)) АБО (ыцця (232; 0); ыццё (158; 0))	0.9
18	шш (90; 0)	1
19	чч (160; 1)	0.9
20	ць (45092; 0)	1
21	АБО (рыні (250; 0), рыпі (383; 0), рыці (355; 0), рыбі (335; 0), рыхі (307; 0), рызі (259; 0), рымі (974; 0), рыві (457; 0), рыкі (234; 0), рысі (117; 0), рылі (1206; 0), рыгі (251; 0), рыў (2218; 0), рыі (399; 0)) АБО (ітры (217; 0), ігры (219; 0), ібры (243; 0), іры (1045; 0), ўры (254; 0))	1
22	АБО (рыжэ (174; 0), рыйц (117; 0), рыйс (178; 0), рыйн (290; 0), рыбр (140; 0), рыцц (526; 0), рыце (981; 0), рыцы (162; 0), рыцэ (126; 0), рыць (548; 0), рыця (345; 0), рыжы (352; 0), рычэ (292; 0), рычы (297; 0), рыпо (148; 0), рыпл (492; 0), рыпе (226; 0), рыпя (149; 0), рыпт (124; 0), рыпу (281; 0), рыпр (234; 0), рызв (119; 0), рызм (317; 0), рымл (335; 0), рымн (106; 0), рымс (247; 0), рымя (344; 0), рыпы (158; 0), рыдз (313; 0), рыдр (110; 0), рынг (278; 0), рыхт (810; 0), рынн (104; 0), рыхо (311; 0), рыхв (120; 0), рыгв (154; 0), рыем (241; 0), рыен (314; 0), рыер (114; 0), рытэ (149; 0), рыкв (213; 0), рыкм (123; 0), рытр (171; 0), рытм (215; 0), рыкр (391; 0), рывя (412; 0), рыгр (212; 0), рышп (175; 0), рышл (122; 0), рыгл (335; 0), рышв (266; 0), рышы (269; 0), рышч (694; 0), рышт (620; 0), рышс (160; 0), рынт (109; 0), рынц (212; 0), рысв (193; 0), рысм (192; 0), рысп (212; 0), рысл (409; 0), рыа (440; 0), рыу (162; 0), рыр (1837; 0), рыя (3260; 0), рыф (1113; 0), рыё (560; 0), рыю (410; 0)) АБО (тэры (1135; 0), цыры (131; 0), _зры (334; 0), бкры (188; 0), ытры (205; 0), узры (156; 0), _фры (194; 0))	0.9
23	АБО (тыві (571; 0), тылі (513; 0), тымі (1883; 0), тыні (109; 0), тыкі (263; 0), тыў (2323; 0)) АБО (істы (3833; 0), ўты (340; 0), іты (1931; 0))	1
24	АБО (тыту (537; 0), тыта (213; 0), тынг (350; 0), тысц (153; 0), тыст (689; 0), тыву (154; 0), тыйн (174; 0), тыя (1990; 0), тып (936; 0), тыз (1907; 0), тыц (562; 0), тыф (643; 0)) АБО (суты (104; 0), энты (150; 0), рэты (425; 0), ыяты (249; 0), аэты (163; 0), экты (164; 0), абты (123; 0), эаты (102; 0))	0.9

Заўвага: Слупок «Схема» табліцы 2 утрымлівае спрошчаныя схемы графічных маркёраў; у дужках праз кропку з коскай указаны ipm маркёра на БН і ЗХ. Графічны маркёр суадносіцца з усёй словаформай або з яе фрагментам (апошняе часцей). Калі няма асобых заўваг, схема дае апісанне маркёра з ігнараваннем рэгістра. Значэнне КАС (адпаведны слупок табліцы 2) прызначаецца па схеме, дадзенай у табліцы 1.

Самымі прыкметнымі і дакладнымі для ідэнтыфікацыі з'яўляюцца адрозненні на графічным узроўні, якія звязаны з ужываннем у беларускіх тэкстах графем, што адсутнічаюць у рускай мове: ***і*** (пры ўмове, что выпраўлены памылкі набору, дзе змешваюцца кірылічнае / лацінскае ***i***), ***ў***. Значэнне КАС для гэтых графем роўнае 1. Фармалізаванае апісанне маркёраў на аснове графем ***і, ў*** (табл. 2, рад. 1).

Знак апострафа, які апрыёры ўспрымаецца як спецыфічны для беларускамоўных тэкстаў, у рускамоўных тэкстах мае нізкую, але не нулявую частату. І ў рускамоўных, і ў беларускамоўных тэкстах знак апострафа выкарыстоўваецца ў наступных выпадках:

а) у складзе пазычанняў, пераважна ў імёнах уласных (*О'Хара (рус.), А'Хара (бел.)* і падобныя са службовым элементам *О' (рус.)* і *А' (бел.)*; *о'кей*, таксама *д'* ці *Д' (рус., бел.)*; *Кот-д'Ивуар, д'Артаньян, Д'Артаньян*;

б) у змешаных кірылічна-лацінскіх словаформах пры аддзяленні іншамоўнай часткі ад рускамоўнай фіналі (*e-mail'ом*).

Толькі ў беларускамоўных тэкстах знак апострафа выкарыстоўваецца ў пазіцыі «перад набраным у ніжнім рэгістры знакам **галоснага**». Задаўшы адпаведнае фармалізаванае апісанне, атрымліваем маркёр з КАС = 1. (табл. 2, рад. 2).

Паказчыкам беларускамоўных тэкстаў і беларускіх украпленняў у рускамоўных тэкстах з'яўляецца ***ё***, таму што для беларускай мовы выкарыстанне *ё* абавязковае, замена графемы *ё* на *е* не дапускаецца,

а ў рускай мове дадзеная графема дэ факта не мае такога статусу да сённяшняга часу. Да таго ж, у беларускай мове на *ё* пачынаюцца пазычаныя словы (*ёгурт, ёд, ёга, ётавы, ётаванне*), у рускай мове яму адпавядае *йо* (*йогурт, йод, йога, йотированный*). Як паказвае практыка, у рускамоўнай прэсе графема ***ё*** выкарыстоўваецца толькі спарадычна. Аднак графема ***ё*** мае высокую ipm на ЗХ (934); таму праводзім даследаванне яе правага / левага акружэння. (табл. 2, рад. 3, 4).

Дыграфы ***дж, дз, шч*** абазначаюць гукі, характэрныя беларускай мове, аднак калі спалучэнне ***шч*** нерэгулярнае для рускамоўных тэкстаў (мае ipm 19, сустракаецца на масіве ЗХ у словаформах па лексемах *веснушчатость, веснушчатый, кошчонка, сиводушчатый*), то спалучэнні ***дж*** и ***дз*** з'яўляюцца рэгулярнымі (маюць на масіве ЗХ ipm 1028 і 1046 адпаведна), і таму патрэбна даследаваць іх акружэнне.

Фармалізаванае апісанне маркёраў як вынік даследавання правага / левага акружэння спалучэнняў ***дж, дз*** змешчана ў табліцы 2, радкі 5 – 8, маркёр на аснове спалучэння ***шч*** – радок 9.

Маркіраванай прыкметай беларускай мовы могуць служыць склады ***жы*** (*жыццё*), ***шы*** (*шырыня, шыпець*), ***чы*** (*чытаць, чысты*), таму што ў рускай мове яны адсутнічаюць. Адзначым, што пералічаныя склады сустракаюцца як у спрадвечнабеларускай, так і ў пазычанай лексіцы, таму што ў працэсе фанетычнай адаптацыі іншамоўныя словы асімілююцца беларускай мовай і збліжаюцца з беларускай лексікай у гучанні і напісанні. Выключэнняў няма. У рускай мове, згодна з традыцыйным прынцыпам напісання, захоўваюцца склады ***жи***, ***чи***, ***ши*** (табл. 2, рад. 10).

Гіпотэза аб выкарыстанні ў якасці магчымых графічных маркёраў для аўтаматызаванай ідэнтыфікацыі беларускамоўнага тэкста складоў ***жэ*** (*жэст*), ***шэ*** (*шэршань, шэсць*), ***чэ*** (*чэмер*), ***чо*** (*чорны, чоканне*) пацвердзілася часткова. Склады ***жэ*** сустракаюцца ў рускай мове ў шэрагу слоў (ipm на ЗХ = 14). У асноўным, гэта формы абрэвіятуры *жэк* і формы вытворнага ад яе слова *жэковский* (*жэк, …, жэках; жэковский,… , жэковских*), а таксама такія склады выяўлены на стыку марфем (на марфемным шве) прыстаўкі *меж-* і кораня *этаж* (*межэтажный, …, межэтажных*). Склад ***шэ*** на ЗХ намі адзначаны ў 11 выпадках (ipm = 5), усе яны звязаны з выкарыстаннем дадзенага спалучэння ў пазычаных лексемах (*сэшэа, шэн, …, шэны*). Толькі ў адным прыкладзе (*эмчээс*) быў выяўлены на ЗХ склад ***чэ*** (ipm = 0.4). Фармалізаванае апісанне маркёраў на аснове спалучэнняў ***жэ, шэ, чэ*** гл. у табліцы 2, радок 11.

Двухлітарнае спалучэнне ***чо*** на ЗХ адлюстравана ў вялікай колькасці прыкладаў (ipm = 626), сярод якіх выразна прасочваецца тэндэнцыя да перадачы ***чо*** ў іншамоўных словах (*анчоус, гаучо, каприччо, лечо, пастиччо, ранчо, пончо*).

Таксама ***чо*** сустракаецца ў формах назоўнага і вінавальнага склонаў слоў з фіналлю *-чок* (*армячок, балычок, бардачок, барсучок, башлычок, башмачок, боровичок, бочок, брачок, бурундучок, бурачок, внучок, воротничок; молчок; облучок; сморчок*); у форме назоўнікаў мужчынскага роду адзіночнага ліку творнага склону на стыку кораня на ***-ч*** і канчатка ***-ом*** (*автотягачом, бирючом, бичом, богачом, бородачом, брюхачом, ветврачом, волосачом, головачом, копачом, кормачом, косачом, космачом, костромичом* і інш.); у форме назоўнікаў жаночага роду адзіночнага ліку творнага склону на стыку кораня на ***-ч*** і канчатка ***–ою / -ой*** (*камчою, кочой, кочою, кяманчой, кяманчою, кукарачой*); у форме назоўніка *плечо* (*плечом*); у кароткіх формах прыметнікаў ніякага роду адзіночнага ліку назоўнага склону і аманімічных ім прыслоўях (*горячо*).

У словах ***чо*** сустракаецца на стыку кораня, які заканчваецца на ***ч***, і суфікса: а) памяншальна-ласкальнага суфікса назоўніка ***-онок-*** (*бельчонок, барсучонок, барчонок, батрачонок, внучонок, волчонок*), б) памяншальна-ласкальнага і памяншальна-знічыжальнага суфікса назоўнікаў ***-онк-*** (*девчонка, казачонка, кепчонка, клячонка, кошчонка, мужичонка, мальчонка, лавчонка, собачонка, старичонка*) і ў прыслоўях, утвораных ад іх (*по-девчоночьи*), в) суфікса прыметнікаў ***-ов-*** (*алычовый, арчовый, грачовая, епанчовая, каланчовый, каракульчовый, парчовый, кумачовый, саранчовый, сургучовая, стосвечовая* і інш.).

Трэба ўлічыць таксама выпадкі выкарыстаня ***чо*** ў некаторых словах: *чокать* і вытворных ад яго (*зачокать, перечокаться*); *чопорный* (*почопорнее*); *чомга*; у вытворных ад *грач* (*грачовник, грачовый*); у слове *вечор* (уст.); у імёнах уласных (*Печорин*) і інш.

Фармалізаванае апісанне маркёраў як вынік даследавання правага / левага акружэння спалучэння ***чо*** змяшчаюць радкі 12–13 табліцы 2.

Адметным для беларускай мовы з'яўляецца фанетычнае падаўжэнне зычных ***л*** (*Купалле, застолле*), ***н*** (*насенне, здарэнне*), ***с*** (*калоссе*), ***ж*** (*раздарожжа, падарожжа*), ***ш*** (*зацішша, застрэшша, Замошша*), ***ч*** (*зарэчча, ноччу*), ***ц*** (*жыццё, быццё*), якое развілося ў выніку страты рэдукаваных у групе зычных перад наступным ***j.*** На пісьме фанетычнае падаўжэнне перадаецца праз падвоенае напісанне адпаведных графем. Шыпячыя ***ж, ш, ч*** зацвярдзелі, таму спалучаюцца з галоснымі непярэдняй зоны ўтварэння (*зарэчча, зацішша, падарожжа*). У рускай мове згаданым беларускім спалучэнням адпавядаюць спалучэнні зычных з ***ьj*** (*застолье, затишье* і г.д.). Патрэбна ўлічваць, што ў беларускай мове могуць сустракацца выпадкі падваення ***сс, нн*** на стыку марфем (*бясстрашны, дрэнны*), таму трэба абавязкова ўлічваць, якая графема будзе наступнай.

Апрыёры (без даследавання правага / левага акружэння) было зразумела, што ў беларускай мове маркіраванымі будуць спалучэнні ***лл, нн, сс, цц*** і літар ***е, і, я, ё, ю.*** Фармалізаванае апісанне маркёраў як вынік даследавання правага / левага акружэння спалучэння ***лл*[*еяюё*]** гл. табліца 2, радок 14, спалучэння ***нн*[*еяюё*]** (табл. 2, рад. 15–16), спалучэння ***цц*[*еяюё*]** (табл. 2, рад. 17).

Спалучэнні ***шш, чч*** спецыфічныя для беларускай мовы (ipm на ЗХ 0 и 1 адпаведна) (табл. 2, рад. 18, 19).

Не было выяўлена маркёраў са спалучэннямі ***сс, жж.***

У рускай мове адсутнічае спалучэнне ***ць***, якое адлюстроўвае адну з асноўных фанетычных рыс беларускай літаратурнай мовы – цеканне (*чытаць, скакаць*) і сустракаецца ў інфінітыве дзеясловаў, формах дзеясловаў 3 асобы множнага ліку цяперашняга і будучага простага часу і інш., таму з'яўляецца прыкметай частотнай і выразна маркіраванай (табл. 2, рад. 20).

Гіпотэза наконт складоў ***ры*** (*Вадохрышча*, *Марыя, фурыя*), ***ты*** (*тытан, ерэтык, тытунь*) як верагодных маркёраў беларускіх тэкстаў не спраўдзілася (ipm на ЗХ: ***ры*** = 14 408, ***ты*** = 22 593) (табл. 2, рад. 21–22, 23–24 адпаведна).

Заключэнне. Вызначаны на падставе лінгвастатыстычнага аналізу комплекс графічных маркёраў можа быць выкарыстаны для аўтаматызацыі разметкі беларускамоўных украпленняў у рускамоўных тэкстах пры стварэнні паўнатэкставых электронных моўных рэсурсаў, якія адлюстроўваюць натуральнае ўзаемадзеянне беларускай і рускай моў ва ўмовах дзяржаўнага двухмоўя. Апрабацыя разметкі беларускамоўных украплення ў рускамоўных тэкстах была праведзена на базе дадзеных беларускіх СМІ. У перспектыве нашага далейшага даследавання – выяўленне маркёраў на базе спалучэнняў графем у іншамоўных словах.

ЛІТАРАТУРА

1. Карский, Е. Ф. Белорусы : в 3 т. / Е. Ф. Карский. – Т. 1: Введение к изучению языка и народной словесности. – Вильн а: Типогр. А.Т. Сыркина, 1904. – 466 с.
2. Карский, Е. Ф. Белорусы : в 3 т. / Е. Ф. Карский. – Т. 3: Очерки словесности белорусского племени. – Кн. 3 : Художественная литература на народном языке. – Петроград, 1922. – 456 с.
3. Булахаў, М. Г. Развіццё беларускай літаратурнай мовы ў XIX – XX ст. ва ўзаемаадносінах з іншымі славянскімі мовамі / М. Г. Булахаў. – Мінск : АН БССР, 1958. – 43 с.
4. Булахов, М. Г. Особенности интерференции русского и белорусского языков / М. Г. Булахов // Проблемы двуязычия и многоязычия / АН СССС, Науч. совет «Закономерности развития нац. яз. в связи с развитием соц. наций», Ин-т языкознания, Ин-т рус. яз, Ин-т яз. и литературы АН ТССР; отв. ред. П.А. Азимов. – М. : Наука, 1972. – С. 217–224.
5. Жураўскі, А. І. Двухмоўе і шматмоўе ў гісторыі Беларусі / А. І. Жураўскі // Пытанні білінгвізму і ўзаемадзеяння моў / АН БССР, Ін-т мовазнаўства імя Я. Коласа, БДУ; рэд. М.В. Бірыла, А.Я. Супрун. – Мінск : Навука і тэхніка, 1982. – С. 18–49.
6. Гируцкий, А. А. Белорусско-русский художественный билингвизм: типология и история, языковые процессы / А. А. Гируцкий; под ред. П. П. Шубы. – Минск : Университетское, 1990. – 175 с.
7. Вешторт, Г. Ф. Смешанные формы речи / Г. Ф. Вешторт // Типология двуязычия и многоязычия в Беларуси. – Мінск : Бел. навука, 1999. – С. 93–101.
8. Коряков, Ю. Б. Языковая ситуация в Белоруссии / Ю. Б. Коряков // Вопросы языкознания. – 2002. – № 2. – С. 109–127.
9. Мечковская, Н. Б. Исторические типы двуязычия и типология языковых конфликтов / Н. Б. Мечковская // Языковой контакт : сб. науч. ст. – Минск : РИВШ, 2015. – С. 125–137.
10. Конюшкевич, М. И. Языковая ситуация в Белоруссии и особенности функционирования русского и белорусского языков / М. И. Конюшкевич // Язык в контексте общественного развития = Language in the Context of Social Development. – М. : ИЯ РАН, 1994.– С. 213–221.
11. Конюшкевич, М. И. Социолингвистические особенности коммуникации в русско-белорусскоязычном социуме / М. И. Конюшкевич // Язык и межкультурные коммуникации : сб. науч. ст. / Мин-во образования РБ, БГПУ им. М. Танка, Вильнюсский педагогический ун-т ; редкол.: В.Д. Стариченок (отв. ред.) [и др.]. – Минск : БГПУ, 2007. – С. 237–239.
12. Цыхун, Г. А. «Трасянка» як аб'ект лінгвістычнага даследавання / Г. А. Цыхун // Беларуская мова ў другой палове XX стагоддзя: матэрыялы Міжнар. навук. канф. / рэдкал.: М.Р. Прыгодзіч (адк. рэд.) [і інш.] – Мінск : Белдзяржуніверсітэт, 1998. – С. 83–89.
13. Бордович, А. М. Сопоставительный курс русского и белорусского языков : учеб. пособие для филол. спец-й вузов / А. М. Бордович, А. А. Гируцкий, Л. В. Чернышова. – Минск : Университетское, 1999. – 223 с.
14. Кривицкий, А. А. Белорусский язык для говорящих по-русски / А. А. Кривицкий, А. Е. Михневич, А. И. Подлужный. – Мінск : Выш. шк., 1990. – 368 с.
15. Сопоставительное описание русского и белорусского языков: морфология / АН БССР, Ин-т языкознания им. Я. Коласа. – Мінск : Навука і тэхніка, 1990. – 336 с.
16. Стариченок, В. Д. Русский язык в Беларуси: состояние, песпективы / В. Д. Стариченок // Слово.ру: Балтийский акцент. – 2012. – № 2. – С. 78–80.
17. Маевская, В. П. Билингвальное и этнокультурное образование в Республике Беларусь / В. Л. Маевская, Р. С. Сидоренко // Русский язык и литература. – 2008. – № 3. – С. 12–17.
18. Новикова, Л. П. Фонетическая интерференция в условиях русско-белорусского двуязычия / Л. П. Новикова // Наука – образованию, производству, экономике: материалы XVII (64) Региональной науч.-практ. конф. преподавателей, научных сотрудников и аспирантов, Витебск, 14–15 марта 2012 г. : в 2 т. – Витебск, 2012. – Т. 1. – С. 186–187.
19. Вардомацкий, Л. М. Особенности ударения существительных в русском, белорусском и украинском языках: учеб. пособие для студентов филол. фак. пед. ин-тов / Л. М. Вардомацкий. – Мінск : Выш. шк., 1988. – 128 с.

20. Метлюк, А. А. Взаимодействие просодических систем в речи билингва: учеб. пособие для ин-тов и фак. иностр. яз. / А. А. Метлюк. – Минск : Выш. шк., 1986. – 110 с.
21. Абабурка, М. В. Параўнальная граматыка беларускай і рускай моў: вучэбны дапаможнік для філал. фак. вышэйшых навучальных устаноў / М. В. Абабурка. – Мінск : Выш. шк., 1992. – 224 с.
22. Гурскі, М. І. Параўнальная граматыка рускай і беларускай моў: фанетыка і марфалогія: падручнік для філал. фак. вышэйшых навучальных устаноў / М. І. Гурскі. – Мінск : Выш. шк., 1972. – 262 с.
23. Трофимович, Т. Г. Сравнительно-историческая грамматика русского и белорусского языков : курс лекций / Т. Г. Трофимович. – Минск: БГПУ, 2006. – 179 с.
24. Бубновіч, І. І. Сістэма форм выражэння грамем роду ў беларускай і рускай мовах у аспекце дыяхраніі / І. І. Бубновіч // Карповские научные чтения : сб. науч. ст. : в 2 ч. ; редкол.: А.И. Головня (отв. ред.) [и др.] – Минск : Белорусский Дом печати, 2014. – Вып. 8. – Ч. 2. – С. 171–175.
25. Киселев, И. А. Частицы в современных восточнославянских языках / И. А. Киселев. – Минск : БГУ, 1976. – 160 с.
26. Мощенская, Л. Г. Как белорусы говорят по-русски? Варианты рода имен существительных в русской речи белорусов / Л. Г. Мощенская; ред. П. П. Шуба. – Минск : Университетское, 1992. – 158 с.
27. Грабчиков, С. М. Межъязыковые омонимы и паронимы. Опыт русско-белорусского словаря. Свыше 550 пар слов / С. М. Грабчиков. – Минск : БГУ, 1980. – 215 с.
28. Міхневіч, А. Я. Вазьмі маё слова...: Нататкі аб лексічным узаемаўплыве беларускай і рускай моў у кантэксце ўзаемадзеяння культур / А. Я. Міхневіч, А. А. Гіруцкі. – Мінск : Навука і тэхніка, 1990. – 87 с.
29. Норман, Б. Ю. Билингвизм и многоречие в Республике Беларусь / Б. Ю. Норман // Русский язык в многоречном социокультурном пространстве / отв. ред. Б. М. Гаспаров, И.А. Купина. – Екатеринбург : УрФУ, 2014. – С. 267–286.
30. Анічэнка, У. В. Гістарычная лексікалогія ўсходнеславянскіх моў : вучэб. дапам. для студ. і выклад. філал. фак. выш. навуч. уст. па спец. «Мовы народаў СССР» / У. В. Анічэнка. – Гомель : ГДУ, 1978. – 94 с.
31. Козырев, И. С. К вопросу сравнительно-исторической лексикологии русского и белорусского языков / И. С. Козырев. – Минск : МГПИ, 1980. – 74 с.
32. Борковский, В. И. Синтаксис сказок: русско-белорусские параллели / В. И. Борковский. – М. : Наука, 1981. – 233 с.
33. Конюшкевич, М. И. Синтаксис близкородственных языков / М. И. Конюшкевич. – Минск : Университетское, 1989. – 156 с.
34. Конюшкевич, М. И. Синтаксис русского и белорусского языков. Сходство и различия: пособ. для учителя / М. И. Конюшкевич, М. А. Корчиц, В. Л. Лещенко. – Мінск : Народная асвета, 1994. – 158 с.
35. Шуба, П. П. Русско-белорусские контакты в области синтаксиса / П. П. Шуба // Вестн. БГУ. Сер. 4, Филология. Журналистика. Педагогика. – 1973. – № 2. – С. 31–36.
36. Чумак, Л. Н. Синтаксис русского и белорусского языков в аспекте культурологии / Л. Н. Чумак. – Минск : Белгосуниверситет, 1997. – 196 с.
37. Михневич, А. Е. Русский язык в Белоруссии / А. Е. Михневич [и др.]; под ред. А. Е. Михневича. – Минск : Наука и техника, 1985. – 272 с.
38. Хаген, М. Развернутый словарь А. А. Зализняка. Полная парадигма. Морфология [Электронный ресурс] / М. Хаген. – Режим доступа: http://www.speakrus.ru/dict/#morph-paradigm. – Дата доступа: 04.05.2018.
39. Кошчанка, У. Лексіка-граматычная база для Беларускага N-корпуса. Зборка ад 10.08.2016 / У. Кошчанка, А. Булойчык, С. Какора. – Рэжым доступу: https://bnkorpus.info/nkorpus-grammar.zip. – Дата доступу: 04.05.2018.
40. Зализняк, А. А. Грамматический словарь русского языка: Словоизменение. Ок. 100 000 слов / А. А. Зализняк – М. : Рус. яз., 1980. – 879 с.

Паступіў 04.09.2018

GRAPHICAL MARKERS FOR AUTOMATED IDENTIFICATION OF BELARUSIAN INCLUSIONS IN A MIXED BELARUSIAN-RUSSIAN TEXT

A. STANKEVICH, I. BUBNOVICH

The linguostatistically defined complex of graphical markers for automated identification of Belarusian inclusions in a mixed Belarusian-Russian text is described. The algorithm of compiling the test corpora of Belarusian and Russian languages and the schemas of graphical markers are provided in the appendix to the article. The revealed complex of the graphical markers can be widely used as a component of linguistic support for creation of diverse full-text language resources in conditions of the Republic of Belarus.

Keywords: *electronic language resources, corpus technologies, annotation, linguistic support, graphical marker, Belarusian-Russian bilingualism, Belarussian inclusions, mixed Belarusian-Russian text.*