УДК 004.912:003.035

С.І. Лысы, Ю.С. Гецэвіч

ГЕНЕРАЦЫЯ НАЦЫЯНАЛЬНАЙ ТРАНСКРЫПЦЫІ ТЭКСТАЎ НА БЕЛАРУСКАЙ МОВЕ

Прапануецца алгарытм аўтаматызаванай генерацыі нацыянальнай транскрыпцыі тэкстаў на беларускай мове, заснаваны на метадзе пераўтварэння «графема – фанема» фанетычнага працэсара сістэмы сінтэзу маўлення па тэксце. Апісваецца прататып сістэмы генерацыі транскрыпцый, распрацаваны на аснове апісанага алгарытму, які дазволіў аўтаматызавана згенераваць масіў адпаведнасцяў «слова – транскрыпцыя» для першага поўнага арфаэпічнага слоўніка беларускай мовы.

Уводзіны

Многія навуковыя і адукацыйныя задачы, тым ці іншым чынам звязаныя з вусным маўленнем, патрабуюць перадачы яго на пісьме пры дапамозе пэўнай знакавай сістэмы. Агульнапрынятым спосабам перадачы мовы ў фанетыцы з'яўляецца транскрыпцыя. Шырокае прымяненне транскрыпцыі тлумачыцца тым, што яна не патрабуе ніякіх спецыяльных прыстасаванняў і можа выдавацца друкарскім спосабам, бо нагадвае звычайнае пісьмо, толькі заснаванае не на арфаграфічных правілах, а на адназначнай адпаведнасці літар і гукаў, якой няма пры звычайным пісьме [1].

Для транскрыбавання тэкстаў існуе мноства розных фарматаў і сімвальных сістэм, якія можна аднесці да аднаго з двух тыпаў – міжнароднай ці нацыянальнай транскрыпцыі. У адрозненні ад міжнароднай транскрыпцыі, якая мусіць мець пэўную ўніверсальнасць адносна разнастайнасці моў і звычайна грунтуецца на лацінскім алфавіце, нацыянальная транскрыпцыя альбо цалкам засноўваецца на алфавіце нацыянальнай мовы, альбо спецыяльным чынам адаптуе яго [2]. У беларускім мовазнаўстве традыцыйна выкарыстоўваюць нацыянальную транскрыпцыю, у аснову якой пакладзена кірыліца. Адзін з варыянтаў беларускай нацыянальнай транскрыпцыі прыводзіцца ў кнізе «Фанетыка беларускай літаратурнай мовы» [1].

Трэба адзначыць, што транскрыбаванне аўдыязапісаў з маўленнем і друкаваных або электронных тэкстаў уручную з'яўляецца надзвычай працаёмкай задачай, а ў літаратуры не сустракаецца інфармацыі пра аўтаматызацыю гэтага працэсу ў дачыненні да тэкстаў на беларускай мове, як не існуе і поўнага арфаэпічнага слоўніка беларускай мовы, які мог бы быць пакладзены ў аснову такой сістэмы [3]. Таму задачай, якую паставілі перад сабой аўтары артыкула, з'яўляецца распрацоўка алгарытму аўтаматызаванай генерацыі нацыянальнай транскрыпцыі тэкстаў на беларускай мове.

Дадзены алгарытм і распрацаваны на яго аснове прататып сістэмы генерацыі транскрыпцый мусіць паспрыяць у вырашэнні наступных задач:

– напісання першага поўнага арфаэпічнага слоўніка беларускай мовы – слоўніка, які адлюстроўвае сучасныя нормы літаратурнага вымаўлення слоў і пастаноўкі націскаў у іх;

– удасканалення алгарытмаў і рэсурсаў фанетычнага працэсара інтэрнэт-сістэмы сінтэзу маўлення па тэксце, распрацаванай супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі НАН Беларусі [4];

– укаранення аўтаматызаванай генерацыі транскрыпцый у адукацыйны працэс як для дапамогі школьнікам і студэнтам у вывучэнні норм беларускага вымаўлення, так і для замежнікаў, якія вывучаюць беларускую мову.

1. Генерацыя транскрыпцый у сістэмах сінтэзу маўлення па тэксце

Алгарытмы генерацыі транскрыпцый з'яўляюцца неад'емнай часткай сістэм сінтэзу маўлення па тэксце (ССМТ), пад якімі разумеюць сістэмы, здольныя пераўтвараць друкаваны тэкст у адпаведны маўленчы сігнал. Такія сістэмы звычайна складаюцца з шэрагу працэсараў, якія паслядоўна апрацоўваюць электронны тэкст (мал. 1). Паколькі структура ССМТ з'яўляецца

паслядоўнай, то некарэктная праца кожнага з працэсараў істотна ўплывае як на працу наступных працэсараў, так і на канчатковы вынік дзейнасці сістэмы [5].

Мал. 1. Агульная схема працы сістэм сінтэзу маўлення па тэксце

Варта адзначыць, што кожны з працэсараў ССМТ змяшчае ў сабе шэраг алгарытмаў, якія могуць уяўляць самакаштоўнасць і быць укаранёнымі ў іншыя сістэмы. Такім чынам, праца па стварэнні і ўдасканаленні ССМТ разумее пад сабой распрацоўку алгарытмаў шырокага прымянення, якія могуць быць выкарыстаны для вырашэння іншых прыкладны́х задач , а праца па вырашэнні гэтых задач, у сваю чаргу, можа служыць для ўдасканалення альбо тэставання алгарытмаў ССМТ.

У аснову алгарытму генерацыі транскрыпцый тэкстаў на беларускай мове было вырашана пакласці метад пераўтварэння «графема – фанема», распрацаваны для фанетычнага працэсара ССМТ і апісаны ў артыкулах [6, 7]. Алгарытм пераўтварэння «графема – фанема» (альбо «літара – фанема») служыць для вызначэння паслядоўнасці фанем, якая адпавядае ўваходнаму арфаграфічнаму тэксту. Выдзяляюць наступныя падыходы да распрацоўкі такіх алгарытмаў [8]: выкарыстанне баз дадзеных; падыход, заснаваны на правілах; выкарыстанне метадаў кіруемых дадзеных і статыстычных метадаў.

Першы з пералічаных падыходаў прадугледжвае выкарыстанне адмысловых слоўнікаў, дзе змешчаны адпаведнасці арфаграфічнага запісу слова яго фанетычнаму запісу. Гэты падыход патрабуе вялікіх аб'ёмаў камп'ютарнай памяці, што становіцца асабліва крытычным у выпадку флектыўных моў (да якіх належыць беларуская), бо колькасць словаформаў такіх моў вельмі вялікая. Акрамя таго, ён патрабуе пастаяннай актуалізацыі і папаўнення базы дадзеных.

Сярод астатніх пералічаных падыходаў выбар шмат у чым залежыць ад тыпу арфаграфічнай сістэмы мовы, якая апрацоўваецца. Статыстычныя метады і метады кіруемых дадзеных найбольш шырока прымяняюцца для моў з так званай «глыбокай арфаграфіяй», у якой суадносіны напісання і вымаўлення надзвычай складаныя або непаслядоўныя. У выпадку, калі арфаграфія мовы мае адносна простыя і паслядоўныя суадносіны з вымаўленнем, больш мэтазгодна выкарыстоўваць падыход, заснаваны на правілах.

Так як арфаграфія беларускай мовы грунтуецца на фанетычным прынцыпе і мае даволі паслядоўныя суадносіны з вымаўленнем, у ССМТ для беларускай мовы звычайна выкарыстоў-

ваецца падыход, заснаваны на правілах. Спісы правілаў пераўтварэння «графема – фанема», якія выкарыстоўваюцца ў сінтэзатары маўлення па тэксце [4], з'яўляюцца прыдатнымі для ўбудавання ў сістэму генерацыі транскрыпцый (табл. 1).

Табліца 1

Фрагменты спісаў правілаў пераўтварэння «графема – фанема» для беларускай мовы

Агульныя правілы «графема – фанема»	Выключэнні з агульных правілаў	Змякчальныя графемы	Агульныя правілы змякчэння
Ж-ZH	Д(С)ТВ-С	Е	(Н)[ДЗЦЙ][ЕЁЮЯІЬ]
З-Z	(Д)[КСПТФХЦЧШ]-Т	Ё	(Н)[СЛЦЗ]
І-І	(Т)[БГДЗЖ]-D	Ю	(Л)[Л]
Й-J'	(З)ДЖ-ZH	Я	(М)[М]
К-К	(З)[КПСТФХЦЧШ]-S	І	([ЗСН])[Д]
…	…	…	…

Аднак фармат выніковых дадзеных алгарытму пераўтварэння «графема – фанема» з'яўляецца недастатковым для адназначнай генерацыі транскрыпцыі, таму было вырашана выкарыстоўваць яшчэ адзін алгарытм фанетычнага працэсара ССМТ, які дае больш дэталёвае фанетычнае апісанне тэксту, пераўтвараючы фанемны тэкст у алафонны.

2. Распрацоўка лінгвістычных рэсурсаў

Як было адзначана вышэй, алгарытмы ССМТ у працэсе працы здзяйсняюць пераўтварэнне арфаграфічнага электроннага тэксту ў транскрыпцыю, але дадзеная транскрыпцыя часта мае даволі спецыфічны, выкарыстоўваемы толькі ў самой сістэме, фармат. Ва ўнутраным прадстаўленні інтэрнэт-сінтэзатара маўлення [4] гэты фармат уяўляе сабой паслядоўнасць абазначэнняў алафонаў і паўз, падзеленых коскамі. Тэкст, прадстаўлены ў такім выглядзе, будзем называць алафонным тэкстам. Ніжэй прыведзены фрагмент арфаграфічнага тэксту і адпаведны яму алафонны тэкст.

Груша цвіла апошні год. Усе галіны яе, усе вялікія расохі, да апошняга пруціка, былі ўсыпаны бурным бела-ружовым цветам.

GH004,R022,U022,SH002,A323,/,C'002,V'002,I241,L002,A012,/,A221,P001,O012,SH002,N'004,I242,/,GH001,O032,T000,/,#P4,U203,S'001,E042,/,GH004,A233,L'002,I042,N004,Y323,/,J'012,A243,J'011,E040,/,#C3,U203,S'001,E043,/,V'012,A243,L'002,I043,K'002,I343,J'012,A342,/,R002,A222,S001,O023,H'002,I340,/,#C3,D004,A322,A221,P001,O012,SH002,N'004,A342,GH004,A231,/,P002,R012,U023,C'002,I342,K004,A330,/,#C3,B004,Y213,L'002,I041,/,W013,S001,Y021,P002,A312,N004,Y221,/,B002,U012,R001,N004,Y221,M001,/,B'002,E141,L004,A312,R002,U222,ZH002,O021,V012,Y211,M003,/,C'002,V'001,E042,T002,A321,M000,/,#P4

Для аўтаматызаванай генерацыі транскрыпцый было неабходна распрацаваць адмысловыя лінгвістычныя рэсурсы для канвертавання алафоннага тэксту ў транскрыпцыю – спісы адпаведнасцяў «алафон – транскрыпцыя». Трэба заўважыць, што кожны алафон у адпаведным тэксце абазначаецца кодам, які складаецца з адной, дзвюх ці трох лацінскіх літар, магчымага знака апострафа і трох арабскіх лічбаў. Гэтае абазначэнне дае інфармацыю пра шэраг характарыстык фанемы ў залежнасці ад кантэксту ў слове і тэксце. Разам алафонная база інтэрнэт-сінтэзатара маўлення налічвае каля 960 розных алафонаў, аднак аналіз адпаведнасці алафонаў і знакаў фанетычнай транскрыпцыі паказаў, што для адназначнага выбару транскрыпцыі дастаткова скарочаных абазначэнняў алафонаў, а менавіта назвы фанемы, знака мяккасці (пры неабходнасці) і першага індэкса. У сувязі з гэтым колькасць неабходных для працы сістэмы адпаведнасцяў «алафон – транскрыпцыя» знізілася да 99. Дадзены спіс быў распрацаваны лінгвістамі на аснове працы А.І. Падлужнага [1]. Фрагмент атрыманага спісу прыведзены ў табл. 2.

Табліца 2

Фрагмент спісу адпаведнасцяў «алафон – транскрыпцыя»

Скарочаны код алафона	Транскрыпцыя
A0	а́
A1	а̀
A2	а
A3	а
B0	б
B'0	б'
B1	б:
B'1	б':
…	…

3. Алгарытм генерацыі транскрыпцый тэкстаў на беларускай мове

Алгарытм генерацыі транскрыпцый дае магчымасць канвертаваць адвольны тэкст на беларускай мове ў яго фанетычнае прадстаўленне – транскрыпцыю. Гэты алгарытм з'яўляецца пашырэннем алгарытмаў, апублікаваных аўтарамі ў артыкулах [9, 10], і выкарыстоўвае алгарытмы пераўтварэння «графема – фанема» і «фанема – алафон», распрацаваныя для ССМТ і апісаныя ў артыкулах [6, 7, 11].

Уваходныя дадзеныя алгарытму: адвольны арфаграфічны тэкст на беларускай мове T_{bel}, дзе дапускаюцца наступныя пазнакі:

плюс /+/ або акут / ́ / – асноўнага націску (напрыклад, «чо+рны», «чо́рны»);

роўна /=/ або гравіс / ̀ / – пабочнага націску (напрыклад, «чырво=на-бе+лы», «чырво̀на-бе́лы»);

цыркумфлекс /^/ або сімвал /◡/ – аб'яднання двух слоў у адно фанетычнае слова (напрыклад, «па^не+бе», «па◡не́бе»).

Рэсурсы алгарытму:

– мноства адпаведнасцяў «знак прыпынку – інтанацыйная памета» $P = <<p_1, int_1>, ..., <p_k, int_k>>$, дзе p_k – k-ы знак прыпынку, int_k – k-я інтанацыйная памета, k – колькасць адпаведнасцяў;

– база дадзеных, якая змяшчае правілы пераўтварэння «графема – фанема»;

– база дадзеных, якая змяшчае правілы пераўтварэння «фанема – алафон»;

– база дадзеных, якая змяшчае адпаведнасці «алафон – транскрыпцыя».

Уваход.

1. Вылучэнне слоў. Для вылучэння слоў з уваходнага тэксту T_{bel} выкарыстоўваецца шаблон слова Pt_w. Ужываючы сінтаксіс рэгулярных выразаў PCRE [12], дадзены шаблон можна прадставіць наступным чынам:

$$Pt_w = (\text{^}|[set1][set1set2]*)([\text{^}set1]*),$$

дзе $set1$ – мноства сімвалаў, з якіх можа пачынацца слова, $set2$ – мноства сімвалаў, з якіх можа складацца, але не можа пачынацца слова, $set1set2 = set1 \cup set2$. У склад мноства $set1$ уваходзяць літары беларускага алфавіту, у склад мноства $set2$ – злучок, апостраф, сімвалы націскаў і інш. Пры дапамозе шаблона Pt_w тэкст разбіваецца на фрагменты двух тыпаў: словы і сімвалы паміж словамі, альбо раздзяляльнікі. Атрыманыя фрагменты збіраюцца ў спіс $L = <<W_1, D_1>,…, <W_n, D_n>>$, дзе W_i – i-е слова, D_i – i-я паслядоўнасць сімвалаў паміж словамі, $i = 1...n$, n – колькасць вылучаных слоў. Спіс L перадаецца на інтанацыйную апрацоўку.

2. Інтанацыйная апрацоўка. У кожнай паслядоўнасці сімвалаў D_i са спісу L здзяйсняецца пошук знакаў прыпынку паводле спісу P. Калі ў паслядоўнасці сімвалаў D_i знойдзены знак

прыпынку, то адбываецца замена D_i на адпаведную знойдзенаму знаку прыпынку p_k інтанацыйную памету int_k. Калі ў паслядоўнасці сімвалаў D_i знакі прыпынку адсутнічаюць, то D_i замяняецца на памету адсутнасці знакаў прыпынку. Атрыманы спіс $L_{int} = <<W_1, int_1>,\ldots, <W_n, int_n>>$ перадаецца ў блок фанетычнай апрацоўкі.

3. Фанетычная апрацоўка. Кожнае слова W_i са спісу L_{int} канвертуецца ў алафонны запіс W_{ai} пры дапамозе функцый пераўтварэння «графема – фанема» і «фанема – алафон» сінтэзатара маўлення па тэксце. Такім чынам,

$$W_{ai} = phoneme\_to\_allophone(grapheme\_to\_phoneme(W_i)),$$

дзе *grapheme_to_phoneme*() – функцыя пераўтварэння «графема – фанема», якая прымае ў якасці аргумента арфаграфічны тэкст і вяртае фанемны тэкст; *phoneme_to_allophone*() – функцыя пераўтварэння «фанема – алафон», якая прымае ў якасці аргумента фанемны тэкст і вяртае алафонны тэкст.

Вынікам дадзенай канверсіі з'яўляецца спіс $L_a = <<W_{a1}, int_1>,\ldots,<W_{an}, int_n>>$, дзе W_{ai} – *i*-е слова ў алафонным запісе, int_i – *i*-я інтанацыйная памета.

4. Фарміраванне правілаў «алафон – транскрыпцыя». Адбываецца зварот да базы дадзеных, якая змяшчае адпаведнасці «алафон – транскрыпцыя». Фарміруецца мноства адпаведнасцяў «алафон – транскрыпцыя» $C = <C_1,\ldots,C_m>$, дзе $C_m = <a_m, tr_m>$, a_m – *m*-ны алафон, tr_m – адпаведны алафону a_m сімвал транскрыпцыі, m – колькасць алафонаў у базе.

5. Генерацыя транскрыпцый. Для кожнага алафоннага слова W_{ai} са спісу L_a выконваюцца крокі 5.1–5.4.

5.1. Вылучэнне алафонаў. У алафонным слове W_{ai} вылучаюцца сімвальныя паслядоўнасці, якія адпавядаюць шаблону Pt_a. Выкарыстоўваючы сінтаксіс рэгулярных выразаў PCRE [12], дадзены шаблон можна прадставіць наступным чынам:

$$Pt_a = [A–Z]\{1,3\}'?[0–9]\{3\}|> ,$$

што адпавядае форме запісу алафонаў і знакаў складападзелу ва ўнутраным прадстаўленні сінтэзатара маўлення па тэксце. Вылучаныя сімвальныя паслядоўнасці захоўваюцца ў спіс *A*.

5.2. Канверсія «алафон – транскрыпцыя». Для кожнага алафона *a* са спісу *A* знаходзіцца адпаведнік паводле спісу *C*. У выніку паслядоўнай апрацоўкі спісу *A* атрымліваем спіс $Tr = <Tr_1\ldots Tr_p>$, дзе Tr_p – транскрыпцыя *p*-га алафона слова W_{ai}, p – колькасць алафонаў у слове.

5.3. Атрыманне транскрыпцыі слова. Адбываецца канкатэнацыя транскрыпцый асобных алафонаў слова W_{ai} у транскрыпцыю W_{tri}.

5.4. Збор выніковых транскрыпцый у спіс. Алафоннае слова W_{ai} у спісе L_a замяняецца на транскрыпцыю W_{tri}.

Такім чынам, пасля апрацоўкі ўсіх слоў W_{ai} са спісу L_a атрымоўваецца спіс $L_{tr} = <<W_{tr1}, int_1>,\ldots,<W_{trn}, int_n>>$.

6. Генерацыя выніковых тэкстаў. Для кожнага элемента спісу L_{tr} адбываецца канкатэнацыя транскрыпцыі і інтанацыйнай паметы, затым адбываецца канкатэнацыя атрыманых фрагментаў у выніковы транскрыбаваны тэкст T_{tr}, які выдаецца карыстальніку.

У выпадку рэалізацыі дадзенага алгарытму для інтэрнэт-асяроддзя немалаважнымі з'яўляюцца крокі 7 і 8.

7. Захаванне дадзеных у архіве. Здзяйсняецца збор інфармацыі аб уведзеных карыстальнікам уваходных дадзеных (зыходны тэкст T_{bel}), атрыманых карыстальнікам выніковых дадзеных (выніковых транскрыпцый T_{tr}), інфармацыі аб адсутнасці тых ці іншых элементаў у базе «алафон – транскрыпцыя», а таксама аналітычнай інфармацыі (памераў тэксту, IP-адрас, даты і часу запыту і інш.). Адбываецца захаванне гэтых дадзеных на серверы.

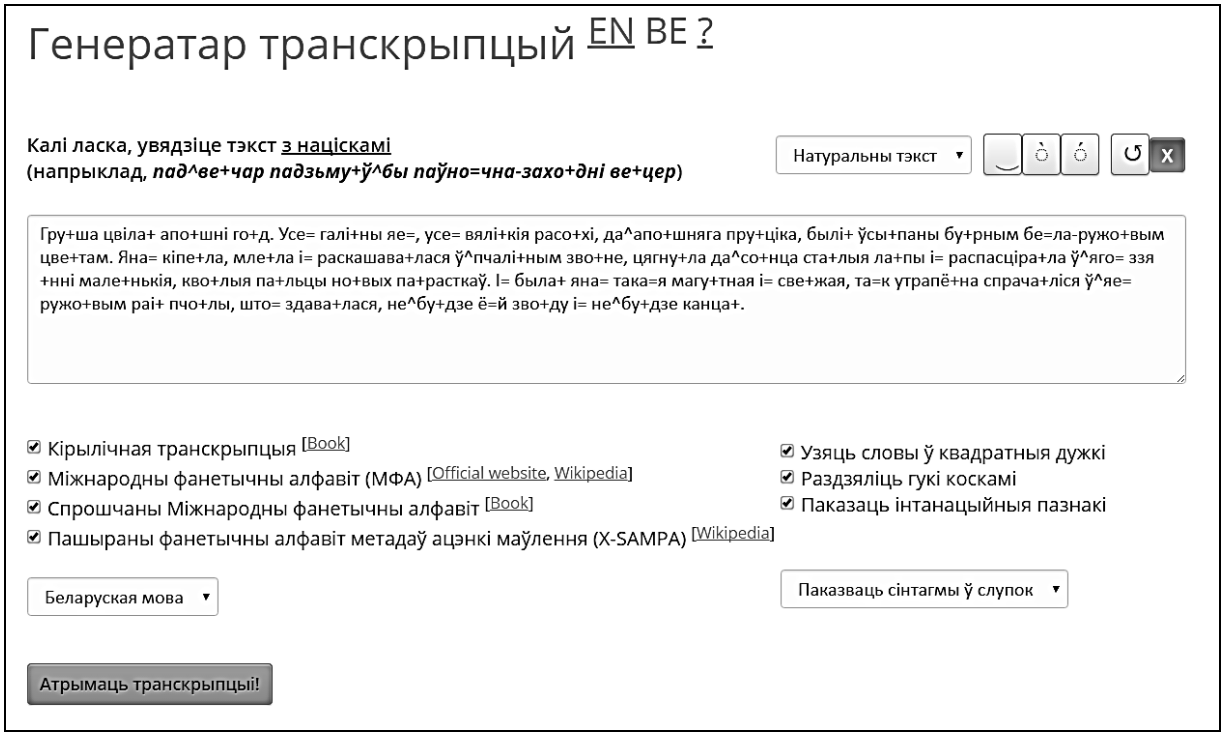
8. Рассылка апавяшчэнняў распрацоўшчыкам. На падставе пералічанай вышэй інфармацыі фарміруецца электронны ліст, які накіроўваецца распрацоўшчыкам для статыстычнага і аналітычнага аналізу, а таксама для вызначэння праблемных сітуацый пры працы з праграмай (памылкі ў лінгвістычных рэсурсах, алгарытмах і праграмных кодах, інш.) і аператыўнага пошуку шляхоў іх вырашэння.

Канец алгарытму.

У выніку апісанага вышэй алгарытму з адвольнага арфаграфічнага тэксту на беларускай мове можа быць згенеравана транскрыпцыя ў традыцыйным для беларускай мовы фармаце. Варта адзначыць, што прапанаваны алгарытм з'яўляецца прыдатным для пашырэння на іншыя фарматы транскрыпцый, бо алафонны запіс, распрацаваны для сінтэзатара маўлення па тэксце, прадстаўляе дастаткова інфармацыі для канверсіі яго ў розныя фарматы транскрыпцыі.

4. Прататып сістэмы генерацыі транскрыпцый

Для апрабацыі, тэставання і нагляднасці працаздольнасці апісанага вышэй алгарытму быў распрацаваны прататып сістэмы генерацыі транскрыпцый у форме вэб-сэрвісу «Генератар транскрыпцый», які даступны для вольнага выкарыстання ў Інтэрнэце [13]. Дадзены сэрвіс дае магчымасць канвертаваць адвольны тэкст на беларускай мове ў нацыянальную транскрыпцыю (мал. 2).



Мал. 2. Графічны карыстальніцкі інтэрфейс вэб-сэрвісу «Генератар транскрыпцый»

На ўваход сэрвіс можа прымаць як адвольны арфаграфічны тэкст, размечаны націскамі, так і тэкст у алафонным фармаце. Для зручнасці працэсу разметкі тэксту маюцца кнопкі ўстаўкі сімвалаў асноўнага націску, пабочнага націску і сімвала аб'яднання некалькіх слоў у адно фанетычнае слова. Сэрвіс дае магчымасць карыстальніку канвертаваць уведзены ім тэкст на беларускай мове ў фанетычную транскрыпцыю, дазваляючы здзяйсняць наладку фармату выніковых дадзеных. Адзначыўшы адпаведныя пункты ў наладках, карыстальнік можа атрымаць транскрыпцыю кожнага слова ў квадратных дужках або без іх; у транскрыпцыі кожнага слова фанемы могуць быць выведзены разам або падзелены коскамі. Калі адзначыць пункт «Паказаць інтанацыйныя пазнакі», выніковы тэкст будзе разбіты на інтанацыйныя фрагменты – сінтагмы. Для гэтага будуць выкарыстаны адмысловыя пазнакі (вертыкальная рыса ‘|’ ў выпадку коскі і падвоеная вертыкальная рыса ‘||’ ў канцы сказа ці абзаца). Пасля ўводу тэксту і адпаведных наладак патрэбна націснуць кнопку «Атрымаць транскрыпцыі!».

Прыклад працы прататыпа сістэмы генерацыі транскрыпцый тэкстаў на беларускай мове, які дэманструе працаздольнасць і карэктнасць распрацаваных алгарытмаў, прыведзены ў табл. 3.

Табліца 3

Прыклад выніковых транскрыпцый, атрыманых пры дапамозе вэб-сэрвісу «Генератар транскрыпцый»

Размечаны тэкст на беларускай мове	Гру+ша цвіла+ апо+шні го+д. Усе+ галі+ны яе+, усе+ вялі+кія расо+хі, да^апо+шняга пру+ціка, былі+ ўсы+паны бу+рным бе=ла-ружо+вым цве+там.
Транскрыпцыя	[γру́ша] [ц'в'іла́] [апо́шн'і] [γо́т] ‖ [ус'э́] [γал'і́ны] [йайэ́] | [ус'э́] [в'ал'і́к'ійа] [расо́х'і] | [даапо́шн'аγа] [пру́ц'іка] | [был'і́] [ўсы́паны] [бу́рным] [б'э̀ларужо́вым] [ц'в'э́там] ‖

5. Тэставанне алгарытму генерацыі транскрыпцый і ўдасканаленне лінгвістычных рэсурсаў

Для тэставання алгарытму генерацыі транскрыпцый быў распрацаваны іншы прататып – «Генератар арфаэпічнага слоўніка», задачай якога з'яўляецца канвертаванне спісу слоў альбо слоўнікавых артыкулаў у фармат арфаэпічнага слоўніка [14]. Пры дапамозе дадзенага інтэрнэт-сэрвісу быў апрацаваны масіў электронных слоў, складзены паводле «Слоўніка беларускай мовы» [15], які налічвае звыш 117 тысяч слоўнікавых артыкулаў. У табл. 4 прыведзены фрагменты спісаў уваходных і выніковых слоўнікавых артыкулаў.

Табліца 4

Фрагменты «Слоўніка беларускай мовы» і артыкулаў арфаэпічнага слоўніка, згенераваных на яго аснове пры дапамозе вэб-сэрвісу «Генератар арфаэпічнага слоўніка»

Фрагмент «Слоўніка беларускай мовы»	Фрагмент арфаэпічнага слоўніка
... **сакаляня́** *і* **сакалянё,** *РДМ* -ня́ці, *Т* –нём; *мн.* -ня́ты, *РВ* -ня́т, -ня́там, -ня́тамі, -ня́тах **сакаляня́тка,** -тку, -так **сакаля́тнік,** -ка, -ку, -каў **са́кас,** -са, -се **саката́нне,** -нні **саката́ць,** *незак*. сакачу́, сако́чаш, -ча, -чам, -чаце, -чуць **сакатлі́вы** **сакату́н,** -на́, -не́, -но́ў **сакату́ха,** -у́се, -у́х **са́ква,** -ве, -ваў **сакваяж,** -жа, -жы, -жаў **сакваяжны** **сакé,** *н*., *нескл*. ...	... **сакаляня́** [сакал'ан'а́] і **сакалянё** [сакал'ан'о́] **сакаляня́тка** [сакал'ан'а́тка] **сакаля́тнік** [сакал'а́тн'ік] **са́кас** [са́кас] **саката́нне** [саката́н':э] **саката́ць** [саката́ц'] **сакатлі́вы** [сакатл'і́вы] **сакату́н** [сакату́н] **сакату́ха** [сакату́ха] **са́ква** [са́ква] **сакваяж** [саквайа́ш] **сакваяжны** [саквайа́жны] **сакé** [сак'э́] ...

Атрыманы шляхам аўтаматызаванай генерацыі масіў электронных слоў і адпаведных ім фанетычных транскрыпцый прайшоў дэтальную праверку экспертамі-лінгвістамі і быў пакладзены ў аснову першага поўнага арфаэпічнага слоўніка беларускай мовы, які на дадзены момант гатуецца да выдання. Вынікі экспертнай праверкі паказалі, што з 117 100 слоўнікавых артыкулаў толькі ў 2109 алгарытм дапусціў памылкі. Такім чынам, дакладнасць працы алгарытму можна ацаніць у 98,2 %.

Здзейсненае экспертамі-лінгвістамі тэставанне не толькі паказала высокі ўзровень карэктнасці працы алгарытму генерацыі транскрыпцый, але і дазволіла атрымаць спіс слоў, якія канвертуюцца некарэктна. Варта адзначыць, што вызначаныя экспертамі памылкі (табл. 5) у большасці сваёй з'яўляюцца сістэматычнымі і могуць быць умоўна падзелены на дзве групы:

– памылкі, выкліканыя некарэктнасцю правілаў пераўтварэння «графема – фанема» фанетычнага працэсара сінтэзатара маўлення па тэксце;

– памылкі ў словах, якія з'яўляюцца выключэннямі і не падпадаюць пад правілы.

Табліца 5

Фрагмент спісу памылак, знойдзеных пры тэставанні алгарытму генерацыі транскрыпцый экспертамі

Слова	Згенераваная транскрыпцыя	Карэктная транскрыпцыя	Апісанне памылкі
гу́зік	[ɣу́з'ік]	[ɡу́з'ік]	Выключэнні на «г-выбухное»
экза́мен	[эгза́м'эн]	[эɡза́м'эн]	
аджа́ць	[аж́а́ц']	[аджа́ц']	Фанетычныя з'явы на сутыку прыстаўкі і кораня
адзна́ка	[аz'на́ка]	[адзна́ка]	
з’і́нелы	[зйі́н'элы]	[з'йі́н'элы]	Фанетычныя з'явы пры наяўнасці апострафа пасля прыстаўкі
уз’ядна́нне	[узйадна́н':э]	[уз'йадна́н':э]	
звы̀шсветлавы́	[звы̀с'в'этлавы́]	[звы̀шс'в'этлавы́]	Фанетычныя з'явы на сутыку частак складанага слова
мà с-спе́ктр	[мà с':п'э́ктр]	[мà сс'п'э́ктр]	
òст-і́ндскі	[òсті́нцк'і]	[òсты́нцк'і]	
...	...	...	...

На падставе прадстаўленых экспертамі дадзеных была праведзена дадатковая карэктыроўка правілаў пераўтварэння «графема – фанема», па выніках якой колькасць памылак зменшылася да 340. Дакладнасць працы алгарытму на масіве электронных слоў «Слоўніка беларускай мовы» адпаведна узрасла да 99,7 %. Далейшыя планы працы прадугледжваюць як працяг карэктыроўкі правілаў, так і ўбудаванне спісу выключэнняў.

Скарэктаваныя правілы пераўтварэння «графема – фанема» былі ўбудаваны ў інтэрнэт-версію сінтэзатара маўлення па тэксце [4], што паспрыяла павышэнню якасці сінтэзаванага маўлення. У будучым плануецца ўкараніць дадзеныя карэктыроўкі ў мабільную і стацыянарную версіі сінтэзатараў беларускага маўлення.

Заключэнне

У артыкуле прапанавана ідэя выкарыстання алгарытмаў ССМТ для вырашэння розных камп'ютарна-лінгвістычных задач, што ўяўляе каштоўнасць як для той сферы, да якой адносіцца абраная задача, так і для самой ССМТ. Гэтая ідэя праілюстравана распрацаваным і апісаным у артыкуле алгарытмам генерацыі транскрыпцый адвольнага тэксту на беларускай мове, у аснову якога былі пакладзены алгарытмы пераўтварэння «графема – фанема» і «фанема – алафон» інтэрнэт-сінтэзатара маўлення па тэксце. На аснове дадзеных алгарытмаў быў распрацаваны прататып сістэмы генерацыі транскрыпцый, карэктнасць яго працы была пратэставана экспертамі-лінгвістамі. Вынікі тэставання сведчаць пра высокі ўзровень дакладнасці працы прататыпа (99,7 % карэктна транскрыбаваных слоў). Распрацаваны прататып сістэмы генерацыі транскрыпцый знайшоў непасрэднае прымяненне ў падрыхтоўцы масіву электронных слоў першага поўнага арфаэпічнага слоўніка беларускай мовы, і ў той жа час экспертная вычытка атрыманага масіву электронных слоў і транскрыпцый дала магчымасць удасканаліць алгарытмы і лінгвістычныя рэсурсы ССМТ.

Спіс літаратуры

1. Фанетыка беларускай літаратурнай мовы / рэд. А.І. Падлужны. – Мінск : Навука і тэхніка, 1989. – 335 с.

2. Тошович, Б. Корреляционная грамматика сербского, хорватского и бошняцкого языков. Часть 1: Фонетика – Фонология – Просодия / Б. Тошович. – М. : Языки славянской культуры, 2011. – 640 с.

3. Стварэнне сэрвіса арфаэпічнага генератара слоўнікаў / Ю.С. Гецэвіч [і інш.] // Тези доповідей Міжнар. конф. «Діалекти в синхронії та діахронії: загальнослов’янський контекст»

(Київ, 2–4 квітня 2014 року) / за ред. П.Ю. Гриценка ; Ін-т укр. мови НАН України. – Київ : КММ, 2014 . – С. 101–106.

4. Text-to-Speech Synthesizer [Electronic resource]. – 2012. – Mode of access : http://corpus.by/TextToSpeechSynthesizer/. – Date of access : 20.01.2017.

5. Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск : Беларус. навука, 2008. – 342 с.

6. Гецэвіч, Ю.С. Фанетычная і алафонная апрацоўка тэксту ў сінтэзатары беларускага і рускага маўлення для мабільных платформаў / Ю.С. Гецэвіч, Б.М. Лабанаў, Д.А. Пакладок // Інфарматыка. – 2014. – № 2(42). – С. 25–35.

7. Гецевич, Ю.С. Алгоритмы преобразования «Буква – Фонема» двуязычного синтезатора речи / Ю.С. Гецевич, Б.М. Лобанов, Д.А. Покладок // Речевые технологии. – 2013. – № 3–4. – С. 95–108.

8. Taylor, P. Text-to-Speech Synthesis / P. Taylor. – N. Y. : Cambridge University Press, 2009. – 626 p.

9. The system of generation of phonetic transcriptions for input electronic texts in belarusian / Yu. Hetsevich [et al.] // Pattern Recognition and Information Processing : Proc. of the 12th Intern. Conf. (28–30 May, Minsk, Belarus). – Minsk : UIIP NASB, 2014. – С. 81–85.

10. Using text-to-speech synthesis algorithms for solving a task of automatic generation of orthoepic dictionary of Belarusian language / Yu. Hetsevich [et al.] // Тези доповідей Міжнар. конф. «Оброблення сигналів і зображень та розпізнавання образів» (Київ, 3–7 листопада 2014 року). – Київ : УкрОБРАЗ, 2014 . – С. 99–101.

11. Гецевич, Ю.С. Алгоритмы преобразования «Фонема – Аллофон» двуязычного синтезатора речи / Ю.С. Гецевич, Б.М. Лобанов, Д.А. Покладок // Речевые технологии. – 2013. – № 3–4. – С. 109–126.

12. Perl-compatible Regular Expressions (PCRE) [Electronic resource]. – 1997–2016. – Mode of access : http://www.pcre.org/original/doc/html/pcrepattern.html. – Date of access : 20.01.2017.

13. Transcription Generator [Electronic resource]. – 2014. – Mode of access : http://corpus.by/TranscriptionGenerator. – Date of access : 20.01.2017.

14. Orthoepic Dictionary Generator [Electronic resource]. – 2014. – Mode of access : http://corpus.by/OrthoepicDictionaryGenerator. – Date of access : 20.01.2017.

15. Слоўнік беларускай мовы / Нац. акад. навук Беларусі, Ін-т мовы і літ. імя Я. Коласа і Я. Купалы ; уклад. Н.П. Еўсіевіч [і інш.] ; навук. рэд. А.А. Лукашанец, В.П. Русак. – Мінск : Беларус. навука, 2012. – 916 с.

Паступіла 23.03.2017

Аб'яднаны інстытут праблем
інфарматыкі НАН Беларусі,
Мінск, Сурганава, 6
e-mail: stanislau.lysy@gmail.com,
yuras.hetsevich@newman.bas-net.by

S.I. Lysy, Yu.S. Hetsevich

GENERATING THE NATIONAL TRANSCRIPTION OF TEXTS IN BELARUSIAN

The paper proposes an algorithm for automatic national transcription generation of texts in Belarusian, which is based on «grapheme-to-phoneme» conversion method of phonetic processor in text-to-speech synthesis system. The paper describes a prototype of transcription generation system, which is developed on the basis of presented algorithm. This algorithm made it possible to generate automatically an array of correspondences «word – transcription» for the first full pronouncing dictionary of the Belarusian language.