УДК [004.522+004.934]

Ю.С. Гецэвіч

АЎТАМАТЫЗАВАНАЯ АПРАЦОЎКА СІМВАЛЬНЫХ ВЫРАЗАЎ У ТЭКСТАХ ДЛЯ СІСТЭМЫ СІНТЭЗУ БЕЛАРУСКАГА МАЎЛЕННЯ

Разглядаюцца праграмныя сродкі, якія дазваляюць знаходзіць і апрацоўваць новыя словы, невядомыя сімвальныя канструкцыі, словы-амографы ва ўваходным тэксце. Апісваецца магчымасць мадыфікацыі лексіка-граматычных катэгорый і націскаў слоў у электронным слоўніку для сістэмы сінтэзу беларускага маўлення.

Уводзіны

Сістэма сінтэзу маўлення Multiphone [1, 2] была ўбудавана ў склад семантычнай пытальна-адказнай сістэмы [3]. Натуральна-маўленчы інтэрфейс прымаў тэкставыя дадзеныя на ўваход і генераваў карыстальніку адказ у выглядзе зразумелага тэкставага паслання, якое агучвалася праз блокі сінтэзу маўлення. Выявілася, што карыстальнік успрымаў адказ цалкам правільна толькі тады, калі тэкставае пасланне, якое агучвалася, складалася з вядомых слоў для электроннага граматычнага слоўніка (ЭГС) і з правільна абраных яго амаграфічных слоў у сістэме сінтэзу маўлення.

Адным з фундаментальных блокаў сістэмы сінтэзу маўлення з'яўляецца лінгвістычны працэсар (ЛП) [4]. Ён выкарыстоўвае ЭГС для расстаноўкі націскаў і лексіка-граматычных катэгорый у словах, вырашае складаныя лінгвістычныя задачы па распазнаванні тэкставых канструкцый [5]. Усяго ў сістэме сінтэзу беларускага маўлення ЭГС (пабудаваны на базе слоўніка [6]) утрымлівае больш за два мільёны запісаў. Беларускі электронны слоўнік дазваляе вырашыць праблему вызначэння граматычных характарыстык і націскаў уваходных слоў на даволі высокім узроўні, але гэты слоўнік не канчатковы, яго патрэбна папаўняць новымі словамі, адсутнымі ў слоўніку. Больш за тое, слоўнік можа ўтрымліваць *словы-амографы* (аднолькавыя паводле напісання, але розныя паводле прамаўлення ў залежнасці ад сэнсавага кантэксту), напрыклад: *ён сказ**а**ў* і *шмат ск**а**заў у тэксце* (літары пад націскам падкрэслены).

Напрыклад, у першай кнізе вядомага твора Уладзіміра Караткевіча «Каласы пад сярпом тваім» налічваецца амаль 80 000 слоў, 4 341 абзац ці 14 522 сказы. Пасля апрацоўкі тэксту лінгвістычным працэсарам выявілася, што ў тэксце ёсць 2 883 невядомыя слоўныя выразы, 74 нераспазнаныя літарна-сімвальныя канструкцыі, 535 слоў-амографаў. Можна сказаць, што на кожныя 35 слоў (у сярэднім 5 сказаў) будзе прысутнічаць невядомае слова, а на кожныя 137 слоў (у сярэднім 8 абзацаў) – слова-амограф. Нераспазнаныя літарна-сімвальныя канструкцыі сведчаць пра тое, што лінгвістычны працэсар у большасці выпадкаў разаб'е іх на склады для акустычнага працэсара, гэта значыць, што сістэма сінтэзу маўлення прачытае іх па скаладах – апрацуе як невядомае слова. Калі падлічыць агульную суму складаных тэкставых момантаў – 3492, атрымаем, што кожнае 23-е слова можа быць няправільна (ці незразумела) агучана. Паколькі сярэдняя хуткасць прамовы аднаго слова 0,6 с, то кожныя 14 с у сярэднім карыстальнік можа чуць незразумелыя вымаўленні слоў. Такое чытанне прымушае слухача прыкладаць намаганні, каб зразумець, пра што ідзе гаворка ў тэксце.

Таму сістэма сінтэзу беларускага маўлення павінна мець праграмныя сродкі, якія маглі б папярэдне апрацоўваць тэксты, папаўняць ЭГС, выяўляць немагчымыя для апрацоўкі ЛП сімвальныя выразы, прапаноўваць сродкі для зняцця амаграфіі слоў ва ўваходным тэксце, даваць доступ да статыстычных дадзеных вынікаў працы ЛП, каб можна было вылічыць сярэдні часавы адрэзак, калі карыстальнік можа пачуць незразумелы для яго выраз.

1. Апісанне праграмнага інтэрфейса сістэмы ўдасканалення электроннага граматычнага слоўніка, лінгвістычнага працэсара і зняцця амаграфіі ў тэксце

ЛП апрацоўвае ўваходны арфаграфічны тэкст у наступнай паслядоўнасці: ачыстка тэксту, пераўтварэнне ўмоўных знакаў і абазначэнняў (абрэвіятур, лікаў і інш.), зняцце шмат-

сэнсоўнасці з сімвалаў алфавіта і са знакаў прыпынку (напрыклад, для скарачэнняў з кропкай у канцы), расстаноўка слоўных націскаў і граматычных прыкмет словаформаў. Для названых апрацовак выкарыстоўваюцца спісы моўных рэсурсаў:

– сімвалаў для ачысткі ўваходнага тэксту;
– скарачэнняў;
– выключэнняў для скарачэнняў вялікімі літарамі;
– выключэнняў для замежных слоў;
– «лік-лічэбнік»;
– «прыназоўнік-склон»;
– прыназоўнікаў;
– часціцаў, а таксама ЭГС.

Беларускі ЭГС пабудаваны ў выглядзе табліцы, якая змяшчае запісы слоў з пазначанымі для іх праз тэгі лексікаграфічнымі катэгорыямі (ЛГК), націскамі (абазначаюцца праз знак «+» ці «=») і прыярытэтамі атрымання слоў з ЭГС лінгвістычным працэсарам (табл. 1). Далей у артыкуле, калі гэта не будзе перашкаджаць сэнсу, ЭГС будзем называць проста *слоўнік*.

Табліца 1

Фрагмент ЭГС з пазнакай націскаў (+), тэгаў і прыярытэтаў слоў

Слова	Тэг	Прыярытэт
…		
зака+зчык	*NNAMO*	1
зака+зчыка	*NNAMG*	1
зака+зчыку	*NNAMD*	1
зака+зчыка	*NNAMA*	1
зака+зчыкам	*NNAMI*	1
…	…	…
ска+з	*NNIMA*	1
ска+зам	*NNIMI*	1
ска+зе	*NNIMR*	1
ска+зы	*NNIMPO*	1
ска+заў	*NNIMPG*	2
…	…	…
скажы+	*VPM2*	1
скажы+це	*VPM2P*	1
сказа+ла	*VPIPF*	1
сказа+ла	*VPIPN*	1
…	…	…

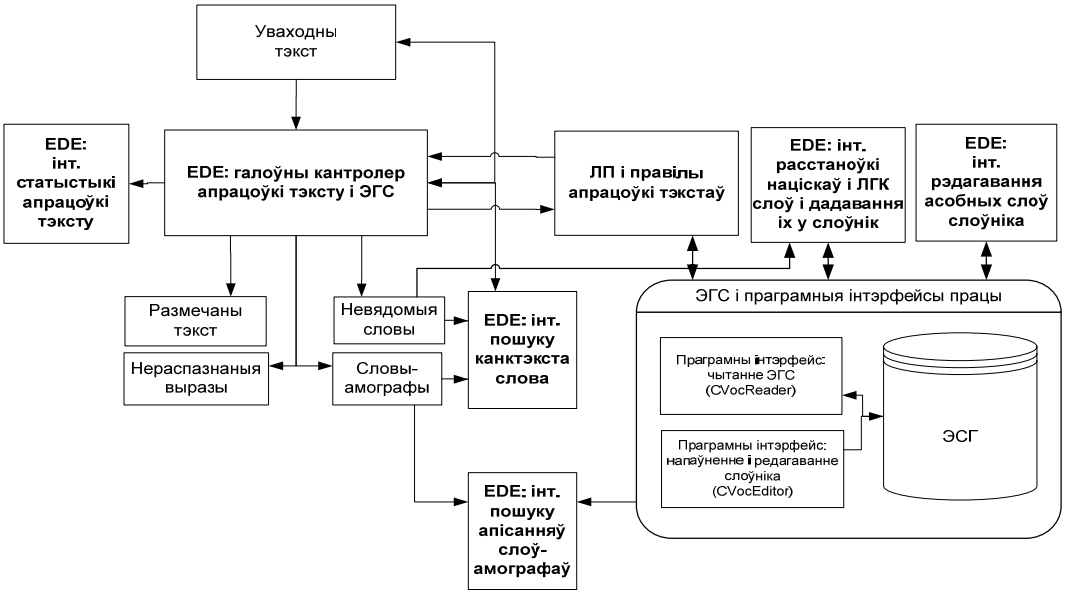
Напрыклад, ЛП па запыце на слова *сказаў* выдасць слова *сказа+ў*, бо першым з ЭГС будзе абрана слова з найменшым лікам сярод усіх прыярытэтаў слоў-амографаў. Калі б прыярытэты былі роўныя, напрыклад *сказа+ў*_1 і *ска+заў*_1, то было б выдадзена слова з абазначанымі частковымі націскамі для ўсіх месцаў у слове, дзе стаялі поўныя націскі, а менавіта *ска=за=ў* (слова будзе прачытана па складах). Заўважым, што прыярытэты найбольш ужывальных слоў-амографаў у ЭГС пастаўленыя паводле атрыманай статыстыкі [7] для беларускай мовы.

Для ўдасканалення працы ЛП і ЭГС быў распрацаваны аўтарам спецыяльны праграмны інтэрфейс – Expert Dictionary Editor (EDE) (мал. 1). Ён складаецца з галоўнага кантролера і спецыялізаваных інтэрфейсных блокаў:

– пошуку канктэксту слова;
– расстаноўкі націскаў і ЛГК слоў, дадавання іх у слоўнік;
– рэдагавання асобных слоў слоўніка;
– статыстыкі апрацоўкі тэксту;
– пошуку апісанняў слоў-амографаў у ЭГС.

Дадзеныя ЭГС даступныя і для чытання, і для мадыфікацыі праз адпаведныя блокі чытання (клас CVocReader), папаўнення, выдалення і рэдагавання (клас CVocEditor)

лінгвістычным працэсарам [8] і EDE-інтэрфейснымі блокамі (акрамя пошука кантэксту слова і статыстыкі апрацоўкі тэксту).



Мал. 1. Узаемадзеянне граматычнага слоўніка з ЛП і праграмай папаўнення слоўніка

Уваходны тэкст апрацоўваецца спецыяльнымі алгарытмамі ЛП. На выхадзе атрымліваюцца рэзультаты: размечаны тэкст, нераспазнаныя выразы, невядомыя словы, словы-амографы, статыстыка. У статыстыку ўваходзяць усе неабходныя дадзеныя для вылічвання сярэдняй колькасці слоў, у якой карыстальнік можа чуць адно незразумелае вымаўленне слова, па формуле

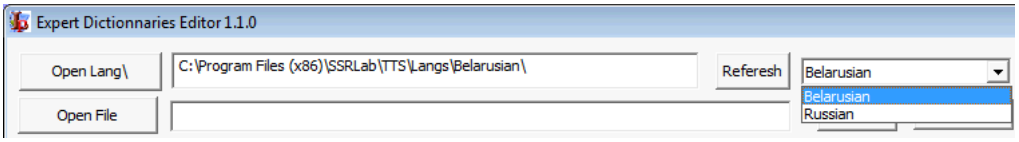
$$T = \frac{W}{S_{new} + S_{undef} + S_{hom}} \cdot t_0 , \qquad (1)$$

дзе S_{new}, S_{undef}, S_{hom} – колькасці адпаведна распазнаных нязнойдзеных дадзеных, нераспазнаных дадзеных, слоў-амографаў; t_0 – сярэдняя хуткасць прамовы аднаго слова (бярэцца з вынікаў працы ўсёй сістэмы сінтэзу маўлення па тэксце ў 2 500–3 000 слоў, звычайна ў інтэрвале 0,5–1 с); W – агульная колькасць апрацаваных слоў праз ЛП.

Інтэрфейсныя блокі EDE забяспечваюць выкананне аўтаматызаваных апрацовак для атрыманых рэзультатаў. Наступнае апісанне праграмы EDE будзе ўтрымліваць паслядоўны тыпавы цыкл працы спецыяліста па апрацоўцы літарна-сімвальных выразаў у тэксце.

2. Праца з тэкстамі ў праграме EDE

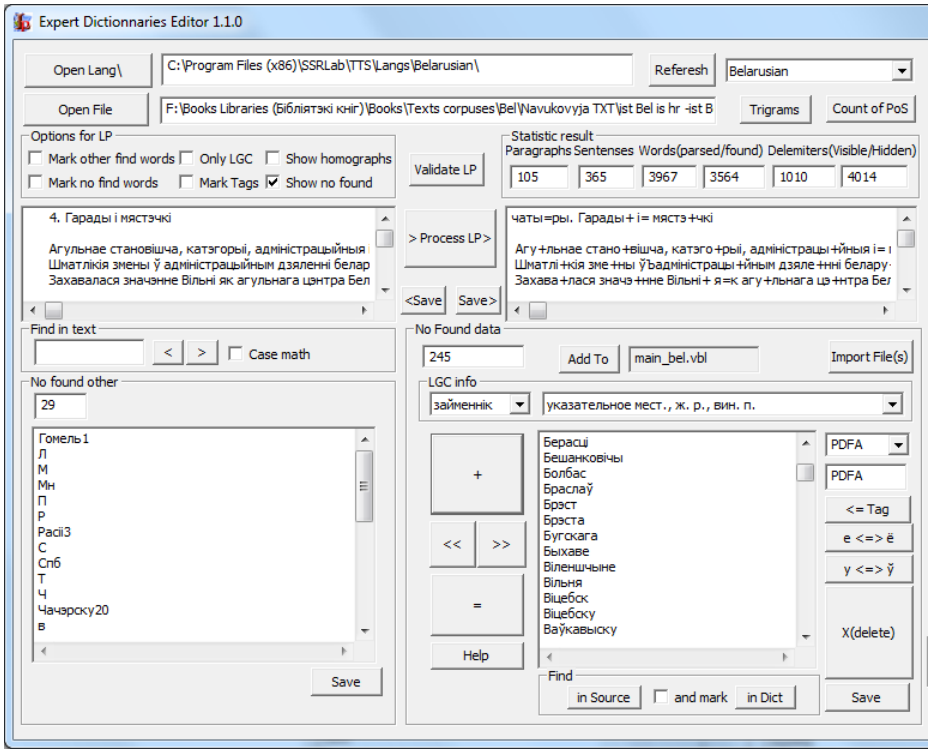
Перад апрацоўкай тэксту праграмай EDE патрэбна ўстанавіць неабходныя параметры для ЛП, які выкарыстоўваецца праграмай для апрацоўкі тэкстаў. Праз каманду Open Lang\ можна абраць неабходную тэчку з моўнымі рэсурсамі, а праз спіс выбару моў можна абраць канкрэтныя моўныя рэсурсы: беларускія (Belarusian; мал. 2).



Мал. 2. Выбар моўных рэсурсаў для ЛП

Адвольны тэкст у праграму можна ўвесці праз каманду адкрыцця файла Open File або адразу ж напісаць яго ў вакне для ўваходнага тэксту (мал. 3). Праз каманду Process LP ЛП апрацоўвае дадзеныя, вынікі змяшчаюцца ў тры акны:

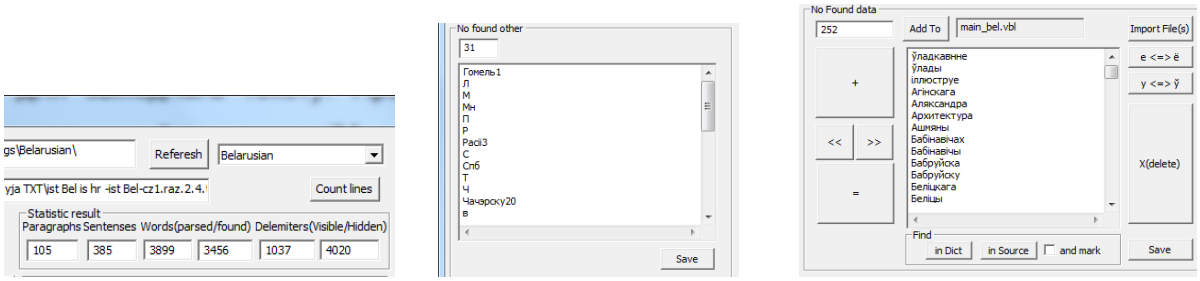
– размечаны тэкст (правае верхняе акно);
– нераспазнаныя выразы (левае ніжняе акно);
– невядомыя словы для ЭГС, але распазнаныя як словы (правае ніжняе акно).



Мал. 3. Прыклады апрацоўкі беларускага тэксту і размеркавання вынікаў апрацоўкі з магчымасцю захавання ў файлы праз каманду Save

Атрыманыя вынікі апрацоўкі можна захоўваць у тэкставыя файлы праз каманды кнопак Save, якія ёсць каля кожнага тэкставага акна. Імёны файлаў для вынікаў апрацоўкі праграма прапануе захоўваць з адпаведнымі постфіксамі: _result, _noFoundOTHER, _noFoundWORDS.

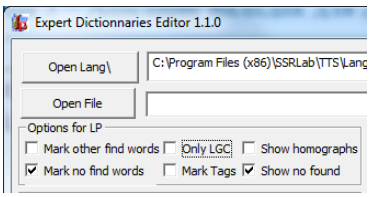
Пасля апрацоўкі тэксту можна паглядзець разнастайную статыстыку па яго структуры (мал. 4). Праграма выводзіць у згрупаванай вобласці Statistics Result колькасць параграфаў, сказаў, слоў (апрацаваных і знойдзеных), раздзяляльнікаў (бачных і нябачных) (мал. 4, *а*), у вобласці No found other – колькасць нераспазнаных выразаў (мал. 4, *б*), у вобласці No Found data – колькасць нязнойдзеных (распазнаных) слоў (мал. 4, *в*). Заўважым, што ў вобласці No Found data пры пэўных параметрах для ЛП (апісваецца ніжэй) можа быць выведзены спіс слоў-амографаў і яго статыстыка (мал. 10). Атрыманыя статыстычныя дадзеныя могуць быць выкарыстаны ў формуле (1).



а) *б)* *в)*

Мал. 4. Статыстыка па структуры тэксту і спісаў прысутнасці ў ім колькасці складаных для ўспрыняцця карыстальнікам выразаў: *а)* параграфаў, сказаў, слоў, раздзяляльнікаў; *б)* нераспазнаных выразаў; *в)* невядомых (але распазнаных) слоў для ЭГС

Карыстальнік можа задаваць спецыяльныя параметры для ЛП перад апрацоўкай тэксту праз выбар канкрэтных опцый, якія знаходзяцца ў вобласці наладак для ЛП Options for LP (мал. 5).



Мал. 5. Наладкі апрацоўкі тэксту

ЛП можа быць сканфігураваны паводле табл. 2.

Табліца 2

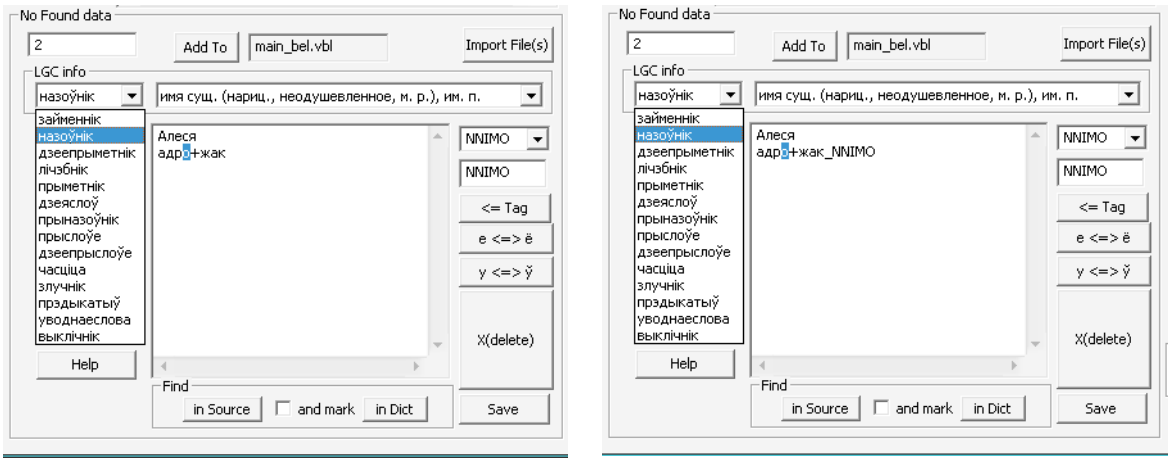
Наладкі для ЛП у праграме EDE

Опцыя наладкі	Апісанне працы ЛП
Абазначаць іншыя знойдзеныя словы (Mark other find words)	У акне апрацаванага тэксту выводзяцца ўсе знойдзеныя ў ЭГС словаформы і амографы слова. Напрыклад, слова *сказаў* транслюецца ў *ска=за=ў*{2, *ска+заў_NNIMPG*_1, *сказа+ў_VPIPM*_1} (розныя націскі, розныя часціны мовы), а слова *момант* – *мо+мант*{2, *мо+мант_NNIMO*_1, *мо+мант_NNIMA*_1} (аднолькавыя націскі, аднолькавыя часціны мовы, але розныя склоны)
Абазначаць нязнойдзеныя словы (Mark no find words)	У акне апрацаванага тэксту нязнойдзеныя словы ў слоўніку абазначаюцца зорачкамі. Напрыклад, слова з абдрукоўкай *маліко* (замест *малако*) транслюецца ў **ма=лі=ко=**
Паказваць толькі ЛГК (Only LGK)	У акне апрацаванага тэксту паказваюцца толькі назвы часцінаў мовы слоў. Напрыклад, выраз *мама мыла дом* апрацоўваецца ў *назоўнік дзеяслоў назоўнік*
Абазначаць тэгі (Mark Tags)	У акне апрацаванага тэксту паказваюцца словы з пазнакай часцінаў мовы праз тэгі і прыярытэт слова ў ЭГС. Напрыклад, выраз *беларуская шляхта* апрацоўваецца ў выраз *белару+ская_JJFO*_1 *шля+хта_NNIFO*_1
Паказваць амографы (Show homographs)	У акне нязнойдзеных дадзеных адлюстроўваюцца словы-амографы. Напрыклад, у беларускім слоўніку для слова *казачку* пазначана чатыры формы: *ка+зачку_NNIFA*_1, *казачку+_NNAMR*_1, *казачку+_NNAMD*_1, *каза+чку_NNAFA*_1, таму ЛП будзе апрацоўваць гэта слова па прынцыпу амографаў: калі не стаіць прыярытэт на нейкім канкрэтным слове, то на розных месцах поўных націскаў будуць пастаўлены частковыя (*ка=за=чку=*) і слова будзе выведзена ў спісе нязнойдзеных дадзеных
Паказваць нязнойдзеныя словы (Show no found)	У акне нязнойдзеных дадзеных будуць адлюстроўвацца нязнойдзеныя ў слоўніку словы. Напрыклад, слова *Алесь* няма ў слоўніку, як і шмат іншых імёнаў уласных, таму гэтае слова будзе апрацавана ў слова з частковымі націскамі *А=ле=сь* і выведзецца ў акно нязнойдзеных слоў

Па змоўчванні ЛП настроены на апрацоўку тэксту і вывад нязнойдзеных слоў, бо абраная птушка Show no found. Апішам магчымасці апрацоўкі новых слоў для ЭГС праз праграму EDE. Пазней апішам магчымасць апрацоўкі знойдзеных амографаў у тэксце, калі абраная птушка Show homographs.

2.1. Апрацоўка распазнаных нязнойдзеных (новых) слоў у праграме EDE

Для дадавання новых слоў з націскамі і ЛГК распрацаваны спецыяльны набор клавішаў (мал. 6).



а) *б)*

Мал. 6. Інтэрфейс у праграме EDE:
абазначэнні ў новых словах націскаў *(а)* і ЛГК *(б)*

Клавішамі са знакамі стрэлачак ‘<<’ і ‘>>’ можна перамяшчаць сіні маркер па галосных невядомых слоў, па літары ‘ў’ і прастаўляць галоўны націск ‘+’ ці частковы націск ‘=’ праз націсканне адпаведных клавішаў ‘+’ і ‘=’, якія размешчаны зверху і знізу клавішаў са стрэлкамі. Выдаленне непатрэбнага націску на галоснай ажыццяўляецца праз клавішу X (delete) (мал. 6, *а*).

Слова будзем лічыць *вылучаным*, калі на ім стаіць мігаючы курсор ці любая літара слова вылучана сінім маркерам.

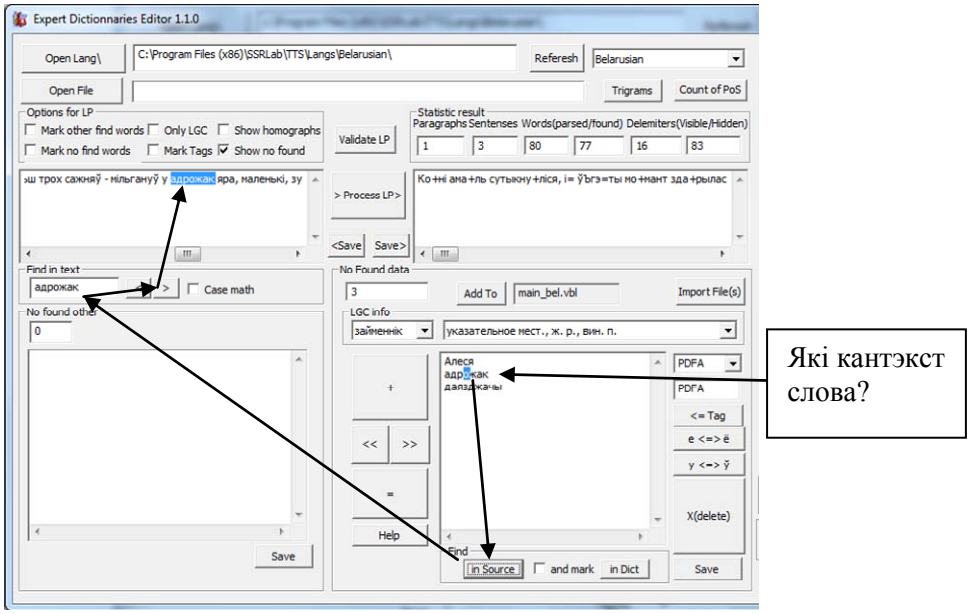
Новы тэг для слова пазначаецца праз выраз *слова_тэг*. Спачатку ў вобласці LGC info абіраецца часціна мовы, а пазней сістэма прапануе абраць іншыя магчымыя ЛГК для гэтай часціны мовы з правага спісу выбару. Сфармаваныя ЛГК адлюстроўоцца ў форме пяці-васьмі вялікіх лацінскіх літараў. Калі выбар тэгу карыстальнік палічыць правільным, то праз націск клавішы <=Tag сфармаваны тэг прыпішацца справа ад выдзеленага слова (мал. 6, *б*). Напрыклад, выраз *адро+жак_NNIMO* дадасца ў слоўнік у якасці слова *адрожак* з поўным націскам на другую галосную літару і з тэгам *NNIMO* (назоўнік, мужчынскі род, назоўны склон, неадушаўлёны, агульны).

У некаторых словах выдзеленыя галосныя ‘е’ ці ‘ў’ могуць быць хутка зменены карыстальнікам на ‘ё’ ці ‘у’ (ці адпаведна наадварот) праз клавішы ‘е<=>ё’, ‘у<=>ў’. Заўважым, што ў беларускі ЭГС словы з пачатковым ‘у’ мусяць быць змешчаныя з ‘ў’ бо ЛП любое слова з пачатковым ‘ў’ шукае ў ЭГС з прымусовай зменай на ‘у’.

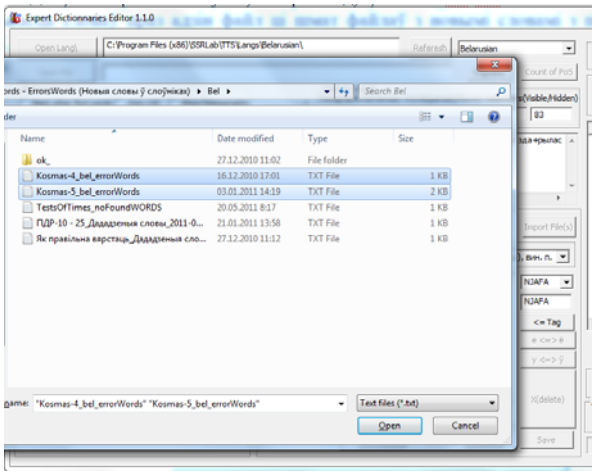
Клавіша in Source ў згрупаванай вобласці Find падсвечвае сінім маркерам канкрэтнае слова ва ўваходным тэксце, каб карыстальнік мог карэктна праставіць характарыстыкі слова адпаведна кантэксту ў тэксце (мал. 7). Птушка and Mark уключае аўтаматычнае сачэнне за невядомымі словамі ў тэксце акна ўводу.

Калі абазначэнні націскаў і ЛГК слоў скончаныя, то праз клавішу Add to словы дадаюцца ў абраны ЭГС, напрыклад у main_bel.vbl. Калі патрэбна дадаць вялікую колькасць слоў праз адзін файл ці шмат файлаў з новымі словамі з пазначанымі націскамі і ЛГК, можна скарыстацца клавішай Import File(s). Яна выклікае дыялог прапановы выбару файла ці файлаў са словамі (мал. 8).

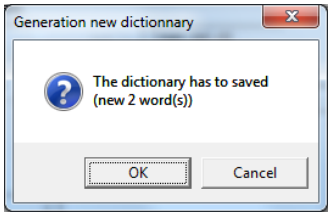
Праграма EDE прапануе дадаць у ЭГС толькі карэктныя словы з абазначанымі націскамі на галосных літарах, прычым галоўны націск мусіць быць толькі адзін, частковых націскаў можа быць некалькі (мал. 9). Пасля станоўчага адказу карыстальніка ў слоўнік дадаюцца абазначаныя словы, і слоўнік гатовы да далейшага выкарыстання. Калі яшчэ раз апрацаваць уваходны тэкст, то абазначаныя і дададзеныя ў слоўнік словы перастаюць паказвацца ў акне нязнойдзеных слоў.



Мал. 7. Ідэнтыфікацыя кантэксту невядомага слова



Мал. 8. Дыялог выбару файла(аў) са словамі для ЭГС



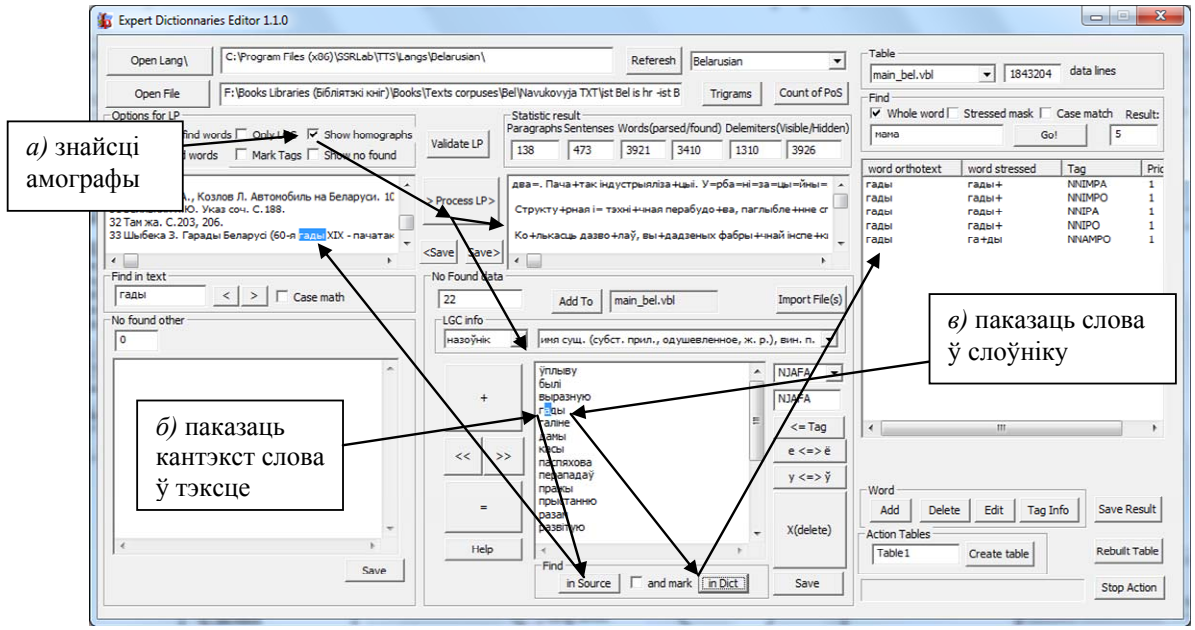
Мал. 9. Дыялог прапановы дадаць новыя абазначаныя словы ў слоўнік

2.2. Апрацоўка нераспазнаных выразаў

Часам у тэксце ёсць нераспазнаныя выразы для ЛП. Яны ўзнікаюць, па-першае, з-за таго, што ЛП да канца яшчэ не распрацаваны для ўсіх сімвальных выпадкаў (напрыклад, для ўсіх фарматаў даты, бо іх ёсць даволі шмат); па-другое, няма ніякага абмежавання на адвольны набор сімвалаў у падаваным тэксце, а здагадацца пра гэтую паслядоўнасць для агульнага выпадку немагчыма. Таму для нераспазнаных выразаў прадугледжана толькі аперацыя захавання ў файл Save (гл. мал. 4, *б*) і яго далейшае перадаванне распрацоўшчыкам праграмы EDE для ўдакладнення алгарытмаў працы EDE.

2.3. Апрацоўка слоў амографаў

Для пошуку амографаў у зыходным тэксце карыстальнік павінен абраць толькі птушку Show homographs (і адключыць птушку Show no found) у наладках ЛП і апрацаваць уваходны тэкст праз націск на Process LP. Амографы з’явяцца ў акне No Found data (мал. 10, *а*).



Мал. 10. Пошук і апрацоўка амографаў

Калі карыстальнік абярэ канкрэтнае слова, то яму важна ведаць кантэкст слова-амографа ў тэксце. Яго можна пабачыць праз апісаную вышэй магчымасць клавішы in Source ў згрупаванай вобласці Find (мал. 10, *б*). Таксама можна пабачыць, якія словаформы слоў-амографаў існуюць у слоўніку (мал. 10, *в*).

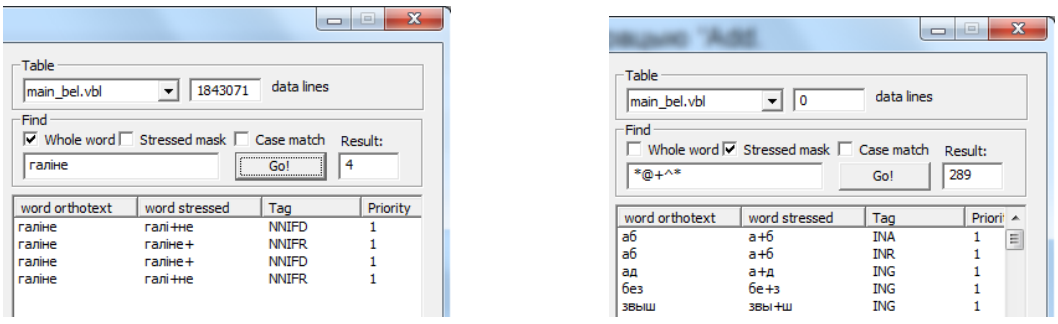
У карыстальніка з'яўляецца дзве магчымасці для апрацоўкі амографаў:

1. Можна ў зыходным тэксце пазначыць правільны націск у слове-амографе і перазахаваць зыходны тэкст праз Save. У выніку будзе атрыманы тэкст, у якім у канкрэтным слове вырашаная амаграфія. Так можна паступіць з усімі астатнімі словамі-амографамі. Такім чынам, амаграфія будзе знятая ва ўсім тэксце.

2. З увагі на тое, што ЛП лічыць амографамі словы, у якіх роўны прыярытэт і розныя галоўныя націскі, то, змяніўшы прыярытэт словаформы ў слова-амографа на больш высокі праз працу са слоўнікам наўпрост (гл. ніжэй), можна часткова зняць амаграфію слова ў тэксце. Гэта можа прымяняцца для экранізацыі малаўжывальных слоў-амографаў для канкрэтнай тэматыкі тэкстаў. Напрыклад, часцей будзе ўжывацца слова *гады+* для тэкстаў гістарычнай тэматыкі, чым слова *га+ды*, якое можа ўжывацца ў тэкстах па біялогіі ці ў гутарковым маўленні.

3. Праца наўпрост са слоўнікамі ў праграме EDE

Асновай ЭГС з'яўляецца файл з пашырэннем *.vbl, таму актыўны слоўнік можна пабачыць у згрупаванай вобласці Table. Побач з ім адлюстроўваецца колькасць запісаных у слоўнік слоў data lines (мал. 11, *а*).



а) *б)*

Мал. 11. Пошук у слоўніку: *а)* канкрэтнага слова; *б)* невыразнага слова

Згрупаваная вобласць Find дазваляе абраць неабходныя наладкі пошуку слова ў слоўніку (табл. 3). Слова можна шукаць, як яно напісана. З дапамогай масак пошуку можна знайсці словы з нейкай варыятыўнасцю адвольных ці канкрэтных галосных і зычных, націскных або ненаціскных літар. Апошні тып словаў будзем называць *невыразнымі*.

Табліца 3

Наладкі пошуку слова ў слоўніку («не» абазначае, што опцыя не абраная, «так» – абраная)

Опцыя ўключаная			Канфігурацыя пошуку
Цэлае слова	З улікам націску	Адрозніваць вялікія/малыя літары	
Так	Не	Не	Канкрэтнае слова шукаецца ў слоўніку з пераводам вялікіх літар у малыя без уліку націску
Так	Не	Так	Канкрэтнае слова шукаецца ў слоўніку без пераводу літар з вялікіх у малыя без уліку націску
Не	Не	Не	Невыразнае слова шукаецца ў слоўніку з пераводам вялікіх літар у малыя без уліку націску
Не	Не	Так	Невыразнае слова шукаецца ў слоўніку без пераводу вялікіх літар у малыя без уліку націску
Не	Так	Не	Невыразнае слова шукаецца ў слоўніку з пераводам вялікіх літар у малыя з улікам націску

Для задавання невыразнасці слоў, якія шукаюцца, былі распрацаваны кіруючыя сімвалы (табл. 4). Прыклады канструявання невыразных слоў прыводзяцца ў табл. 5.

Табліца 4

Апісанне ўжывання магчымых кіруючых сімвалаў для пабудовы маскі слова, якое шукаецца

Кіруючыя сімвалы	Тлумачэнне
+, =	Пошук з улікам націска (толькі пры націснутым Stressed Mask)
?	Любы адзін сімвал
*	Любая колькасць сімвалаў (0, 1, 2, …, n)
@	Любая адна галосная літара
^	Любая адна зычная літара
#	Любы адзін не галосны і не зычны сімвал

Табліца 5

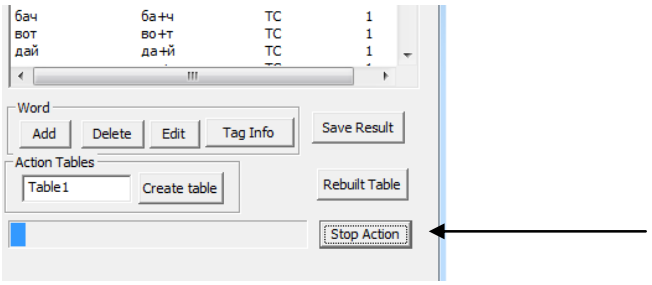
Прыклады пабудоў невыразных слоў для пошуку ў слоўніку

Невыразнае слова	Тлумачэнне
??+??	Знайсці ўсе словы з чатырма сімваламі з націскам на другі
??+*	Знайсці ўсе словы з мінімум двума сімваламі з націскам на другі
@^@^@^+	Знайсці ўсе словы з шасцю сімваламі, з пабудовай слова «галосны, зычны, галосны, зычны, галосны, зычны» і з націскам на трэцюю галосную
#	Знайсці ўсе словы з любой колькасцю сімвалаў, сярод якіх мусяць быць сімвалы не галосныя і не зычныя
мам*	Знайсці ўсе словы з любой колькасцю сімвалаў, але з пачаткам на «мам». Прычым пры націснутым Stressed mask выберуцца словы з ненаціскной першай галоснай
мо=та*	Знайсці ўсе словы з любой колькасцю сімвалаў, але з пачаткам на «мота». Прычым пры нявыбраным Stressed mask выберуцца словы з сімвалам ‘=’ пасля літары ‘о’. Пры выбраным Stressed mask выберуцца словы з частковым націскам на літары ‘о’

Заўважым, што алгарытм пошуку не правярае словы на карэктнасць расстаноўкі націскаў на зычных літарах ці на іншых сімвалах, якія выпадкова ці невыпадкова могуць быць у словах (напрыклад, сімвал дэфіс ‘-’), таму часам у табл. 4 і 5 для абазначэння літары выкарыстоўваецца слова *сімвал*.

Калі наладкі абраныя і слова (канкрэтнае або невыразнае) уведзена, то праз каманду Go! праграма EDE звернецца да слоўніка і выведзе колькасць знойдзеных адказаў (Result:) і самі вынікі пошуку ў чатыры калонкі: слова, слова з пазначаным націскам, тэг, прыярытэт (гл. мал. 11).

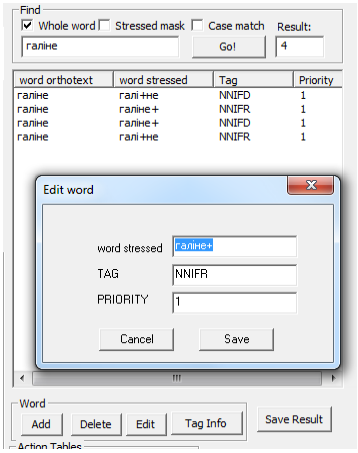
Пошук слова можа быць спынены праз каманду Stop Action (мал. 12).



Мал. 12. Спыненне пошуку ў слоўніку

У выніках пошуку над любым выбраным словам можа быць выканана дадатковая аперацыя (мал. 13):

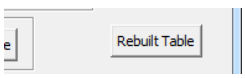
– выдаліць (Delete);
– рэдагаваць (Edit);
– паказаць расшыфроўку тэга (Tag Info).



Мал. 13. Прыклад рэдагавання словаформы

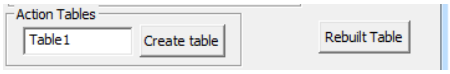
Дадаць новы запіс у слоўнік можна праз аперацыю Add, прычым ніякіх праверак над дадаваным словам праводзіцца не будзе.

У інтэрфейсе наўпростай працы са слоўнікам ёсць клавіша Rebuilt перабудавання асноўнага файла слоўніка (напрыклад, main_bel.vbl) і хэша слоўніка (напрыклад, main_bel.hsh) (мал. 14). Яе можна выкарыстоўваць у двух выпадках. Падчас таго, як словы са слоўніка выдаляюцца, для хуткасці працы ставіцца толькі пазнака на слове, што яно выдалена, а фізічна слова застаецца ў файле *.vbl, таму слоўнік патрэбна перабудоўваць, каб паменшыць яго файл. Таксама калі адбываецца пашкоджанне ці выпадковае выдаленне файла *.hsh слоўніка, то яго можна ўзнавіць, перабудаваўшы слоўнік.

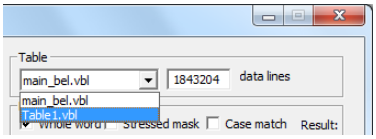


Мал. 14. Аперацыя перабудовы слоўніка Rebuilt

Для таго каб стварыць новы слоўнік, патрэбна скарыстацца аперацыяй Create table (мал. 15, *а*). Слоўнік ствараецца з напісаным іменем (па змоўчванні Table1.vbl), і ў яго дадаецца адно тэставае слова *ма+ма*.



а)



б)

Мал. 15. Прыклады працы з новым слоўнікам: *а)* стварэнне; *б)* актывацыя

Заключэнне

Праграма Expert Dictionary Editor дазваляе карыстальніку паляпшаць ЭГС і апрацоўваць тэксты для сінтэзатара беларускага маўлення Miltiphone: знаходзіць новыя і карэктаваць ужо дададзеныя словы, апрацоўваць словы-амографы і фіксаваць нераспазнаныя выразы ў тэксце, вылічваць прыкладную працягласць маўлення без памылак.

Важна адзначыць, што сістэма можа быць болей дапрацавана для зручнасці карыстальнікаў. Напрыклад, можна выкарыстаць разнастайныя падсвечванні колерам нязнойдзеных слоў і нераспазнаных выразаў у апрацаваным тэксце. Таксама магчыма распрацаваць аўтаматычную сістэму падказак націскаў і ЛГК для новых слоў паводле статыстычных дадзеных, якія даступныя са слоў ужо сфармаванага ЭГС, так, каб карыстальнік толькі згаджаўся ці не згаджаўся з прапанаванымі варыянтамі. У наўпростай працы са слоўнікам можна дадаць кнопку для праслухоўвання знойдзеных слоў, гэта дазволіць працаваць рэдактару слоўнікаў у меншым візуальным напружанні.

Аўтар удзячны навуковаму кіраўніку д. т. н. Б.М. Лабанаву за дапамогу ў правядзенні практычнай часткі і напісанні артыкула, магістру філалагічных навук С.А. Гецэвіч за кансультацыі ў філалагічных пытаннях, а таксама НАН Беларусі за выдзелены грант для распрацоўкі тэмы «Алгарытмы лінгвістычнай апрацоўкі тэкстаў на беларускай і рускай мовах».

Спіс літаратуры

1. Лобанов, Б.М. Компьютерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск : Беларуская наука, 2008. – 342 с.

2. Lobanov, B. Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian) / B. Lobanov, L. Tsirulnik // Speech and Computer : proc. of the 11th International conf. SPECOM'2006, St. Petersburg, Russia, 25–29 June, 2006 / Institute of Informatics and Automation of RAS, Speech Informatics Group. – SPb. : Anatolia, 2006. – P. 274–283.

3. Естественно-языковой интерфейс вопросно-ответных систем / В.А. Житко [и др.] // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2011 : материалы Междунар. науч.-техн. конф., Минск, 10–12 февраля 2011 г. – Минск, 2011. – С. 395–408.

4. Dutoit, T. An Introduction to Text-To-Speech Synthesis / T. Dutoit. – Netherlands : Kluwer Academic Publishers, 1997. – P. 63–64.

5. Uvarova, L.A. Mathematical Modeling. Problems. Methods. Applications / L.A. Uvarova, A.V. Latyshev. – N.Y. : Kluwer Academic/Plenum Publishers, 2000. – P. 123

6. Слоўнік беларускай мовы: Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / Ін-т мовазнаўства імя Я.Коласа АН БССР; пад рэд. М.В. Бірылы. – Мінск : БелСЭ, 1987. – 903 с.

7. Цирульник, Л.И. Проблема графической омонимии при синтезе русской и белорусской речи и статистические алгоритмы ее решения / Л.И. Цирульник, Ю.С. Гецевич, В.В. Веремей //

Белорусский язык в культурном и языковом пространстве славянских стран : тр. Междунар. конф., Минск, 24–25 ноября 2009 г. – Минск, 2009. – С. 323–333.

8. Гецевич, Ю.С. Система редактирования и пополнения словарей речевого интерфейса вопросно-ответной системы для беларуского и русского языков / Ю.С. Гецевич, В.Н. Вяльцев // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2011 : материалы Междунар. науч.-техн. конф., Минск, 10–12 февраля 2011 г. – Минск, 2011. – С. 413–424.

Паступіла 12.07.11

Аб'яднаны інстытут праблем
інфарматыкі НАН Беларусі,
Мінск, Сурганава, 6
e-mail: mix1122@gmail.com

Y.S. Hetsevich

AUTOMATED PROCESSING SYMBOL EXPRESSIONS IN THE TEXTS FOR BELARUSIAN SPEECH-TO-TEXT SYNTHESIS

The article is dedicated to the software tools permitting finding and processing new words, unknown symbol constructions, and homographs in the input text. The possibility of modification of lexical and grammatical categories and syllabic accents in words in the electronic dictionary for Belarusian speech-to-text synthesis applications is considered.