ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ И РЕЧИ

УДК [004.522+004.934+004.91]:004.89

Ю.С. Гецэвіч, Т.І. Окрут, Б.М. Лабанаў

АЛГАРЫТМЫ ІДЭНТЫФІКАЦЫІ РЭПЛІК СА СЛОВАМІ АЎТАРА Ў ЭЛЕКТРОННЫХ ТЭКСТАХ НА БЕЛАРУСКАЙ МОВЕ

Разглядаюцца асноўныя этапы стварэння аўтаматызаваных алгарытмаў для ідэнтыфікацыі рэплік з устаўкамі слоў аўтара, прапаноўваецца іх дапрацоўка ў мованезалежным напрамку. Прыводзяцца вынікі ацэнкі працы распрацаваных мадэляў на трэніровачным і тэставым тэкстах з дакладнасцю ў тэрмінах сярэдняй гарманічнай меры больш за 90 %.

Уводзіны

У цяперашні час інфармацыя выконвае вельмі значную ролю ў жыцці чалавека, пранізвае ўсе сферы чалавечай дзейнасці. Яе шырокае прымяненне складае неабходную аснову функцыянавання сучаснага грамадства. Пры гэтым асабліва цэняцца аб'ём і хуткасць яе атрымання, чаму значна садзейнічае з'яўленне інтэрнэта, з дапамогай якога даступна вялікая колькасць электронных кніг, у тым ліку і ў агучаным выглядзе. Аднак чытанне тэксту з экрана можа сапсаваць зрок, а дыктарская (акторская) агучка кнігі патрабуе вялікіх выдаткаў часу і сродкаў для стварэння гукавага запісу і яго карэктыроўкі.

Альтэрнатыўным спосабам агучвання тэксту кнігі выступае выкарыстанне сістэм сінтэзу маўлення па тэксце (ССМТ). За кароткі час такія сістэмы здольны стварыць электронны гукавы файл па ўваходным электронным тэксце. Для таго каб сінтэзаванае маўленне было разборлівым, прымяняюцца сістэмы перадапрацоўкі літарна-сімвальных канструкцый, пошуку невядомых слоў і зняцця шматзначнасці слоў у электронных тэкстах [1].

Тэкст кнігі можа ўключаць не толькі словы аўтара, але і словы двух ці больш моўцаў (персанажаў). Гэта абгрунтоўвае матываванае жаданне стваральніка аўдыёкніг выкарыстаць розных дыктараў ці сінтэзаваныя галасы, каб аўдыёкніга была больш набліжанай да адлюстравання ўнікальных характарыстык маўлення персанажаў. Разам з тым час разметкі рэплік для чытання іх рознымі дыктарамі ці сінтэзатарамі маўлення застаецца вялікім з-за патрэбы рэдактарскай прачыткі ўсяго тэксту перад агучваннем для абазначэння слоў персанажаў.

Калі рабіць агляд, то на сённяшні дзень ужо існуюць некаторыя напрацоўкі, звязаныя з гэтым напрамкам аналізу тэксту. Так, напрыклад, анлайн-сістэма Text Analysis Demo дазваляе ідэнтыфікаваць персанажаў у тэксце і іх выказванні, што ў адносінах да ССМТ з'яўляецца важным момантам у вызначэнні роду моўцы [2, 3]. Таксама групай еўрапейскіх навукоўцаў былі распрацаваны алгарытмы па ідэнтыфікацыі персанажаў і аўтаматычнаму вызначэнню іх ролі ў творы з дапамогай сінтаксічных граматык NooJ [4]. Што датычыцца славянскіх моў, можна адзначыць працу харвацкіх навукоўцаў па вызначэнні простай мовы тэксту [5], але ў ёй не разглядаецца праблема вызначэння роду моўцаў. У гэтым напрамку распрацоўваюцца такія сродкі стварэння аўдыёкніг, як праграмы MP3book2005 і AUDIOBOOK [6, 7]. У іх убудаваны спецыяльныя блокі лагічнага аналізу дыялогаў, якія рэалізуюць разметку слоў персанажа і слоў аўтара ў дыялагічным тэксце. У AUDIOBOOK былі ажыццёўлены крокі ў бок рэалізацыі чытання па ролях, але праграма ахоплівае не ўсе выпадкі. У ёй ігнаруецца структура афармлення простай мовы, калі ёсць словы моўцы з некалькімі аўтарскімі ўстаўкамі. Таксама не разглядаецца магчымасць вызначаць род моўцы па іншых вызначальніках, напрыклад па спалучэннях тыпу «дзеяслоў + назоўнік у мужчынскім родзе» ва ўстаўках аўтара:

– Надо написать «ять», – *отвечает ученик*.

Варта адзначыць, што праграма AUDIOBOOK лепш працуе з рускамоўнымі і англамоўнымі маўленчымі рухавічкамі, а блокі лагічнага аналізу дыялогаў увогуле не прадугледжваюць працу з іншымі мовамі, акрамя рускай.

Такім чынам, дадзены артыкул працягвае распачатую ў артыкуле [8] тэму распрацоўкі алгарытмаў ідэнтыфікацыі простай мовы і вызначэння роду моўцаў па ўстаўках слоў аўтара ў электронным тэксце. Перад аўтарамі ставіцца задача распрацоўкі моўанезалежных мадэляў і стварэння асобных лінгвістычных рэсурсаў для іх з мэтаю фармалізацыі як можна большай колькасці сінтаксічных структур дыялагічных рэплік і вызначэння роду моўцаў (персанажаў) па ўстаўленых у простую мову словах аўтара. Разглядаюцца варыянты выкарыстання распрацаваных алгарытмаў у CCMT, для стварэння аўдыёкніг ці ў якасці дапаможнага прыкладання ў працы дыктараў (рэдактараў).

1. Збор і сістэматызацыя матэрыялу для вызначэння роду моўцаў

Для распрацоўкі алгарытмаў экспертамі быў выбраны твор Уладзіміра Караткевіча «Каласы пад сярпом тваім». Спачатку ў якасці трэніровачнага матэрыялу выкарыстоўваліся першыя 12 раздзелаў (падрабязную інфармацыю можна знайсці ў артыкуле [8]), далей былі дабаўлены астатнія раздзелы – усяго 32).

Усе алгарытмы будаваліся з дапамогай беларускага модуля міжнароднай лінгвістычнай праграмы NooJ [9]. Гэта праграма дазваляе распрацоўваць сінтаксічныя і марфалагічныя граматыкі і тэставаць іх на вялікай колькасці тэкстаў. З іх дапамогай можна потым ствараць сінтаксічныя анатацыі і экспартаваць размечаны тэкст як файл XML для далейшай апрацоўкі [10].

Заўважым, што NooJ апрацоўвае кожны наступны абзац тэксту асобна ад папярэдняга. Таму кожны абзац (усяго іх 5748) трэніровачнага тэкставага корпусу быў асобна прааналізаваны экспертам у наступным парадку:

– пазначэнне абзацаў з простай мовай;

– пазначэнне рэплік з устаўкамі слоў аўтара;

– пазначэнне рэплік моўцаў (персанажаў) мужчынскага і жаночага роду (адпаведна падкрэслены шрыфт для мужчынскага роду і курсіўны шрыфт для жаночага роду (табл. 1)).

Табліца 1

Фрагмент трэніровачнага матэрыялу з ручной разметкай для абзацаў з простай мовай

Паметы простай мовы	Простая мова з устаўкамі слоў аўтара	Тэкст твору «Каласы пад сярпом тваім» па абзацах
1	1	*– Бацька вады, – шэптам сказала Майка.*
1	1	– Бацька вод, – паправіў Алесь. – Вось так і Дняпро пачынаецца недзе.
1	1	*– Жывая вада, – сказала Яня.*
0	0	І яна апусцілася на калені і зламала пальчыкамі крыштальную паверхню.
1	0	*– Піце. Будзеце жыць сто год…*

Ручная апрацоўка паказала, што тэкст налічвае 2805 абзацаў з простай мовай (з іх 2580 простая мова, якая пачынаецца з працяжніка, 225 – без працяжніка) і 1281 рэплік з устаўкамі слоў аўтара (сярод іх 1075 рэплік моўцаў мужчынскага роду, 197 – жаночага роду і 9 – невыразных моўцаў).

На пачатковым этапе на аснове аналізу дыялагічных рэплік былі выведзены ніжэй прыведзеныя структуры афармлення простай мовы дыялогаў з наступнымі абазначэннямі: М – словы моўцы; А – словы аўтара; дужкі (,) – пачатак і завяршэнне набору варыяцый знакаў прыпынку; сімвал вертыкальнай лініі | – выкарыстоўваецца для раздзялення магчымых варыяцый знакаў прыпынку (аператар *«або»*) .

1. Словы моўцы без слоў аўтара:

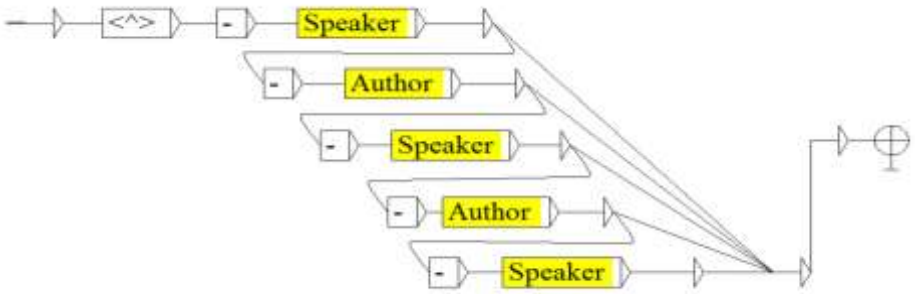
– М (! | !! | !!! | ? | ?! | … | .).

2. Словы моўцы са словамі аўтара ў канцы:
– M (, | ! | !! | !!! | ? | ?! | … | .) – A (… | .).
3. Словы моўцы з некалькімі аўтарскімі ўстаўкамі:
– M (, | ! | !! | !!! | ? | ?! | … | .) – A (, | … | . | : | .) – M (, | ! | !! | !!! | ? | ?! | … | .)
(– A (, | … | . | : | .) – M (, | ! | !! | !!! | ? | ?! | … | .)).

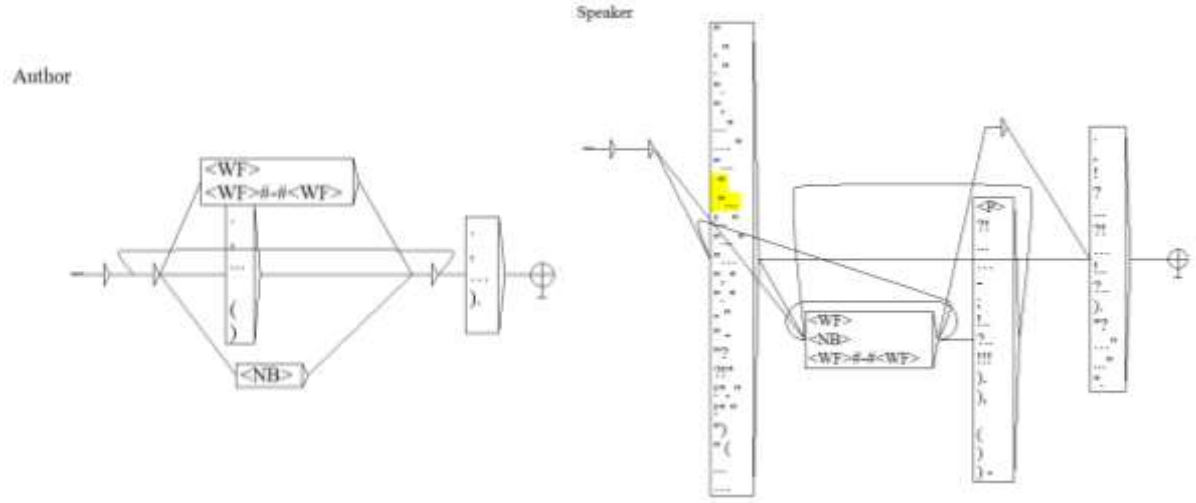
Варта адзначыць, што падчас трэніроўкі і тэсціроўкі распрацаваных алгарытмаў дадзеныя структуры амаль не падвергліся зменам, былі толькі дададзены некалькі спалучэнняў з двукоссем і правай круглай дужкай (унутраныя знакі прыпынку ў словах аўтара або моўцы разглядаюцца асобна).

2. Распрацоўка аўтаматызаваных алгарытмаў выяўлення слоў моўцаў і слоў аўтара

На аснове табл. 1 і фармалізаваных структур афармлення простай мовы была распрацавана сінтаксічная граматыка NooJ з назвай DS_All для аўтаматызаванага вызначэння ўсіх абзацаў з простай мовай (мал. 1). Яе галоўныя часткі-падграфы Speaker і Author ідэнтыфікуюць адпаведна словы моўцы (персанажа) і словы аўтара (мал. 2).



Мал. 1. Агульны выгляд сінтаксічнай граматыкі DS_All



Мал. 2. Падграфы Speaker і Author граматыкі DS_All,
дзе WF – любáя словаформа, NB – любáя паслядоўнасць лічбаў

Спрошчана апішам працу граматыкі для ідэнтыфікацыі першай і другой фармалізаваных структур простай мовы, а пасля абагульнім яе працу для трэцяй структуры. Паводле граматыкі DS_All простая мова пачынаецца з працяжніка, далей ідуць словы моўцы Speaker (лікі, розныя формы слоў са знакамі прыпынку), якія, у сваю чаргу, могуць скончыцца на кропку, коску, клічнік, пытальнік або іх камбінацыю (таксама былі дабаўлены некаторыя спалучэнні з двукоссем). Калі пасля слоў моўцы не ідуць словы аўтара, то граматыка завяршае працу. Так будзе ідэнтыфікавана

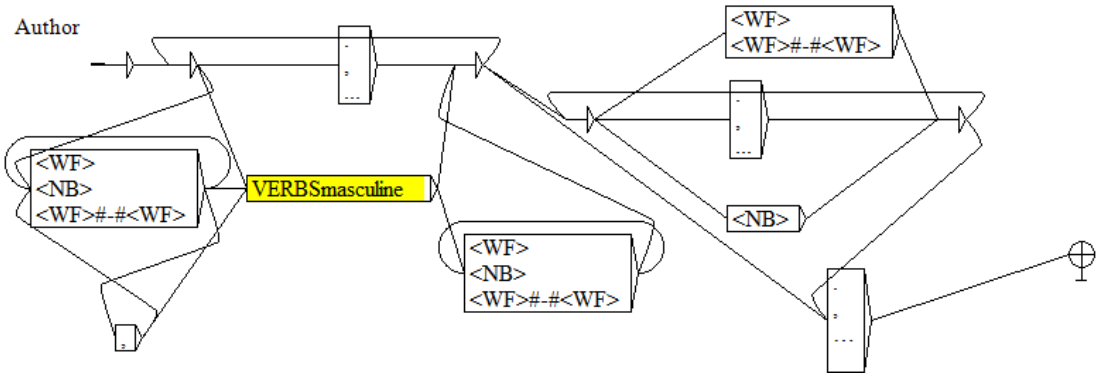
структура першага тыпу. Калі пасля іх зноў ідзе працяжнік, то граматыка задзейнічае падграф Author (таксама ўключае розныя формы слоў, такія знакі прыпынку, як коска, кропка, шматкроп'е і дужкі). Так будзе ідэнтыфікавана структура другога тыпу.

Аналагічна далей граматыка можа ідэнтыфікаваць ад адной да двух уставак слоў аўтара ў простую мову персанажа праз спрацоўванне паслядоўных падграфаў Speaker –> Author –> Speaker. Так будзе ідэнтыфікавана структура трэцяга тыпу.

3. Распрацоўка аўтаматызаваных алгарытмаў ідэнтыфікацыі роду моўцаў па словах аўтара

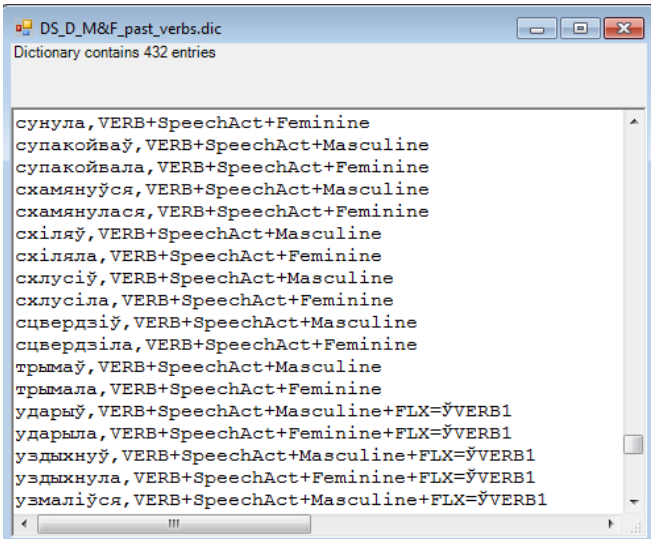
Найбольш прыдатнымі для ідэнтыфікацыі роду моўцы апынуліся дзеясловы дзеяння мінулага часу (паправіў, сказала) персанажаў.

На аснове родазалежнай разметкі тэксту (гл. табл. 1) былі пабудаваны граматыкі для абазначэння моўцаў мужчынскага і жаночага роду. Для гэтага граф Author граматыкі DS_All быў дапрацаваны. Спачатку ў яго быў даданы родазалежны падграф адпаведна для мужчынскага роду – VERBSmasculine (мал. 3), аналагічна для жаночага – VERBSfemenine. У выніку на аснове графа DirectSpeech былі пабудаваны дзве асобныя граматыкі: DirectSpeechMasculine і DirectSpeechFemenine.



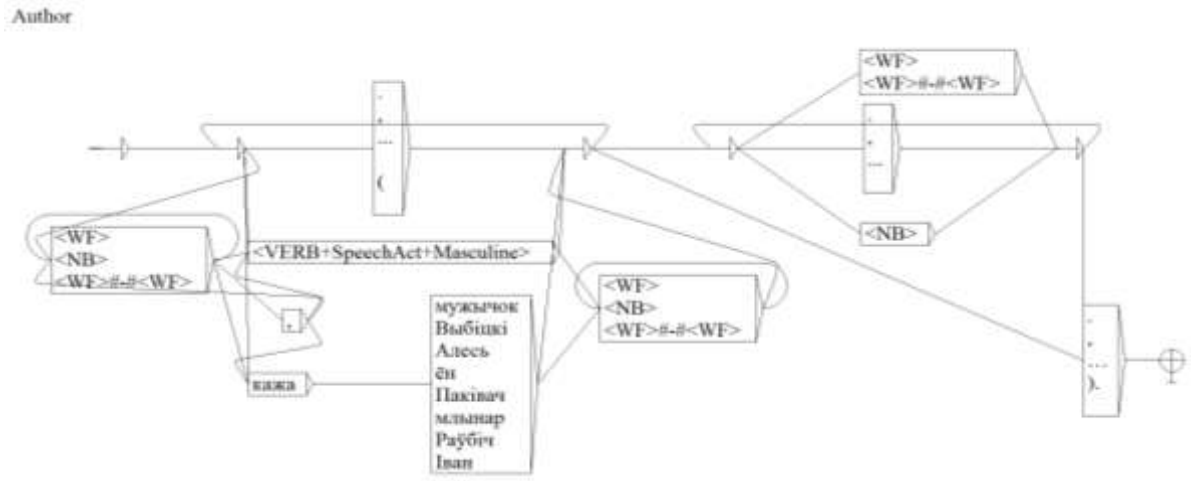
Мал. 3. Падграф Author граматыкі DirectSpeechMasculine

На далейшым этапе па меры папаўнення спісу дзеясловаў – паказчыкаў роду быў створаны асобны слоўнік. У ім парамі прадстаўлены дзеясловы мінулага часу ў формах для жаночага і мужчынскага роду. Для дзеясловаў, якія пачынаюцца з «у», была створана спецыяльная парадыгма ЎVERB1, яна ўлічвае пераход на «ў» пасля галосных (мал. 4).



Мал. 4. Слоўнік дзеясловаў-паказчыкаў

На мал. 5 можна заўважыць, як выкарыстоўваецца атрыманы слоўнік: замест падграфа VERBSmasculine цяпер прымяняюцца спецыяльныя тэгі (катэгорыі) SpeechAct (семантычная пазнака для дзеясловаў – каментарыяў простай мовы) і Masculine, адпаведна для падграфа VERBSfemenine – тэгі SpeechAct і Feminine.



Мал. 5. Падграф Author граматыкі DirectSpeechMasculine пасля дапаўнення

У ходзе распрацоўкі алгарытмаў таксама была даследавана праблема граматычных амафармаў. Так, дзеяслоўная форма «кажа» (форма цяперашняга часу для дзеяслова казаць) можа адносіцца і да мужчынскага і да жаночага роду, у выніку чаго такія формы не могуць асобна выкарыстоўвацца для ідэнтыфікацыі роду моўцы. Для вырашэння дадзенай задачы была створана дадатковая звязка графаў «дзеяслоў – назоўнік», дзе першы граф уключае граматычныя амафармы дзеясловаў маўлення, другі – спіс назоўнікаў – паказчыкаў роду моўцы (персанажа). Графы адлюстраваны на мал. 5.

Атрыманыя граматыкі могуць прымяняцца паслядоўна да мэтавага тэксту праз NooJ, прычым вынікі пазнак першай граматыкі захоўваюцца другой граматыкай. Такім чынам будзе атрымана разметка тэксту на рэплікі для мужчынскага і жаночага галасоў. Слоўнік дзеясловаў – паказчыкаў роду на дадзены момант змяшчае 432 запісы.

Функцыя праграмы NooJ *Locate Pattern* дазваляе праглядзець знойдзеныя абзацы простай мовы ў тэксце граматыкай DirectSpeech у выглядзе канкарданса (мал. 6).

Text	Before	Seq.	After
Kalasy_08.not	ў бацькі:	- Быў жа, здаецца, святы з мурынаў? Ці, можа, не?	- Быў, - сказаў
Kalasy_08.not	можа, не?	- Быў, - сказаў бацька. - Здаецца, Хведар-мурын.	- Ну вось
Kalasy_08.not	па сходах.	- Каго яшчэ няма, Georges? - спытала матухна.	- Раўбічаў няма
Kalasy_08.not	спытала матухна.	- Раўбічаў няма. Кроера няма. Старога Вежы няма.	- Добра. Хай
Kalasy_08.not	Вежы няма.	- Добра. Хай ідуць дзеці, - уздыхнула Загорская.	Алесь бег

Мал. 6. Простая мова, якую знайшла граматыка DS_All

Каб працу гэтых граматык можна было выкарыстоўваць у ССМТ пад стандарт SAPI 5.1, неабходна прывесці тэксты да выгляду SAPI TTS XML [11]. Таму для выбару неабходнага сінтэзатара маўлення сінтаксічная анатацыя, якую генерыруе граматыка, павінна быць адаптаваная пад наступны код:

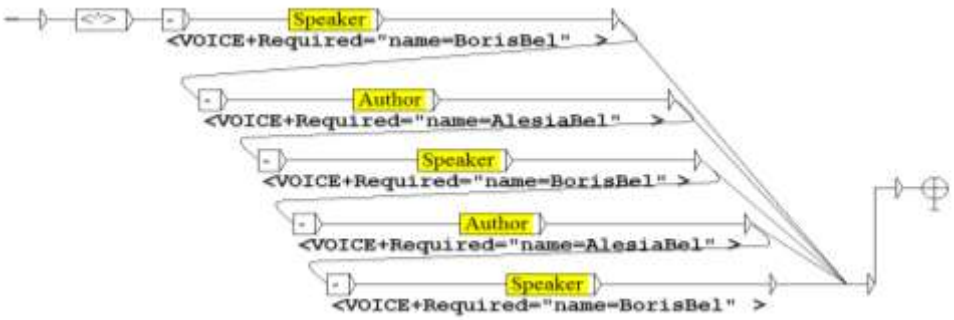
<VOICE Required="name=[Назва голасу ў сістэме]">

…Тэкст для агучкі…

</VOICE>.

Для гэтага ў DirectSpeechMasculine і DirectSpeechFemenine былі дададзены маркеры абазначэння шляхоў, якія спрацавалі ў граматыках (мал. 7). Маркеры наладжаны так, што не

пазначаны тэкст твору і словы аўтара чытаюцца голасам AlesiaBel, мужчынскія рэплікі – голасам BorisBel, а жаночыя – ElenaBel.



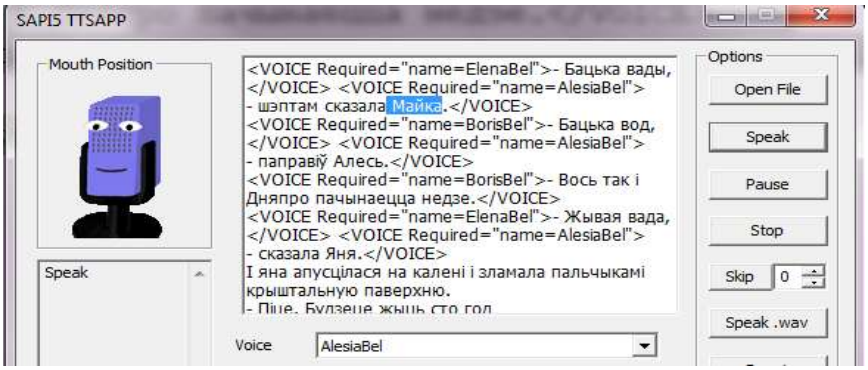
Мал. 7. Прыклад простай мовы, якую знайшла граматыка DS_All,
з абазначанымі адпаведнымі галасамі для словаў моўцы і аўтара

Для сказаў з табл. 1 афармленне тэксту твору пасля спрацоўвання граматык будзе выглядаць як на прыведзеным ніжэй прыкладзе. Словы аўтара і моўцаў змяшчаюцца ў тэгі VOICE з атрыбутам Required са значэннямі, якія адпавядаюць назвам галасоў сінтэзатараў. Як і планавалася, для слоў аўтара выкарыстоўваецца голас AlesiaBel, а для мужчынскіх і жаночых рэплік моўцаў адпаведна – галасы BorisBel і ElenaBel.

Анатаваны тэкст праз тэгі VoiceXML для аўтаматычнага пераключэння
сінтэзатараў у залежнасці ад роду моўцы і слоў аўтара

```
<VOICE Required="name=ElenaBel">- Бацька вады,</VOICE> <VOICE Required="name=AlesiaBel">
- шэптам сказала Майка.</VOICE>
<VOICE Required="name=BorisBel">- Бацька вод,</VOICE> <VOICE Required="name=AlesiaBel">
- паправіў Алесь.</VOICE>
<VOICE Required="name=BorisBel">- Вось так і Дняпро пачынаецца недзе.</VOICE>
<VOICE Required="name=ElenaBel">- Жывая вада,</VOICE> <VOICE Required="name=AlesiaBel">
- сказала Яня.</VOICE>
І яна апусцілася на калені і зламала пальчыкамі крыштальную паверхню.
- Піце. Будзеце жыць сто год…
```

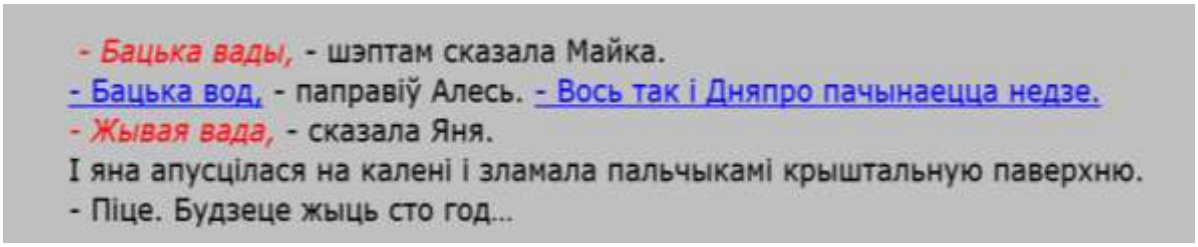
Пасля гэтага такі размечаны тэкст можна падаць на ўваход CCMT. На мал. 8 паказана, як праграма SAPI5 TTSAPP аўтаматычна пераключае пастаўленыя ў сістэме галасы AlesiaBel, BorisBel, ElenaBel пры выбранай опцыі Process XML.



Мал. 8. Агучванне тэксту трыма сінтэзатарамі маўлення, які папярэдне быў аўтаматычна размечаны граматыкамі DirectSpeechMasculine і DirectSpeechFemenine

Вынікі працы створаных граматык таксама магчыма выкарыстоўваць і ў мэтах візуальнага абазначэння слоў аўтара і слоў іншых персанажаў у залежнасці ад патрэбнага роду голасу. Гэта можа быць выкарыстана рэдактарам для хуткага прагляду колькасці персанажаў, якія гавораць у творы, каб падабраць неабходную колькасць сінтэзатараў маўлення ці дыктараў.

Для апрацоўкі файлаў VoiceXML была распрацавана спецыяльная праграма VoiceXmlToColorReplacer, якая дазваляе канвертаваць тэгі пазнак неабходнага голасу сінтэзатара ў тэгі HTML з рознымі стылямі выяўлення рэплік (мал. 9).



Мал. 9. Аўтаматычна размечаны тэкст, які зручна агучваць дыктарам

Так, словы аўтара друкуюцца простым шрыфтам чорнага колеру, мужчынскія рэплікі – сінім падкрэсленым, а жаночы голас – чырвоным курсіўным шрыфтам.

4. Ацэнка працы распрацаваных алгарытмаў

У выніку трэніроўкі граматык на 32 раздзелах твору «Каласы пад сярпом тваім» агульная граматыка DS_All знаходзіць 2574 з 2580 рэплік, з іх 2562 правільныя. Граматыка для вызначэння мужчынскага роду DirectSpeechMasculine знаходзіць 1045 з 1075 рэплік (рэплікі з устаўкамі слоў аўтара) і граматыка для вызначэння жаночага роду DirectSpeechFemenine – 183 з 197. Для больш дэталізаванай ацэнкі выкарыстоўваюцца значэнні дакладнасці (P), паўнаты (R) і іх сярэдняй гарманічнай велічыні (табл. 2). Пры гэтым M – правільна знойдзеныя граматыкай рэплікі, L – усе знойдзеныя граматыкай рэплікі і N – вызначаныя экспертам правільныя рэплікі ў тэксце.

Табліца 2

Ацэнка працы сінтаксічных граматык NooJ па вызначэнню сказаў з простай мовай і роду персанажаў па словах аўтара (трэніровачны корпус)

Назвы граматык	Дакладнасць $P = M/L$	Паўната $R = M/N$	Сярэдняя гарманічная велічыня (F1-measure), % $2*P*R*100 / (P+R)$
DirectSpeech	2562/2574 = 0,995	2562/2580 = 0,993	99,4
DirectSpeechMasculine	1029/1045 = 0,984	1029/1075 = 0,957	97
DirectSpeechFemenine	181/183 = 0,989	181/197 = 0,919	98,9

У якасці тэставага матэрыялу быў выкарыстаны корпус з мастацкай літаратурай, куды ўвайшлі некаторыя раздзелы твору Якуба Коласа «На ростанях», а таксама «У гарах дажджы» Івана Мележа, «Жалезная кнопка» Людмілы Рублеўскай і «Асеннія лісты» Цёткі. Тэксты падбіраліся метадам выпадковай выбаркі і разам налічваюць 23 867 словаўваходжанняў (першыя 32 раздзелы твору «Каласы пад сярпом тваім» налічваюць 106 217 словаўваходжанняў).

Па падліках эксперта, створаны корпус усяго ўтрымлівае 481 рэпліку (рэплікі з працяжніка), 165 рэплік моўцаў мужчынскага роду і 68 рэплік моўцаў жаночага роду.

Дэталёвыя вынікі тэсціроўкі створаных алгарытмаў на тэставым корпусе адлюстроўваюцца ў табл. 3.

Табліца 3

Ацэнка працы сінтаксічных граматык NooJ па вызначэнню сказаў з простай мовай і роду персанажаў па словах аўтара (тэставы корпус)

Назвы граматык	Дакладнасць $P = M/L$	Паўната $R = M/N$	Сярэдняя гарманічная велічыня (F1-measure), % $2*P*R*100 / (P+R)$
DirectSpeech	461/462 = 0,995	461/481 = 0,958	97,6
DirectSpeechMasculine	143/145 = 0,986	143/165 = 0,866	92,2
DirectSpeechFemenine	57/58 = 0,982	57/68 = 0,838	90,4

Заключэнне

Такім чынам, была пастаўлена і вырашана задача па паляпшэнню распрацаваных алгарытмаў і лінгвістычных рэсурсаў для ідэнтыфікацыі простай мовы ў тэксце і вызначэння роду моўцаў па ўстаўках слоў аўтара з дакладнасцю больш за 90 %. У выніку распрацоўкі слоўніка з паказчыкамі роду атрыманыя алгарытмы могуць разглядацца як рэсурсанезалежныя і выкарыстоўвацца для апрацоўкі тэкстаў на іншых славянскіх мовах, неабходна толькі папаўняць адпаведныя слоўнікі. У той жа час распрацаваныя мадэлі паказалі даволі добрыя вынікі ў спалучэнні з ССМТ і ў далейшым могуць быць выкарыстаны ў пабудове дадатковага блоку аўтаматычнага выбару мужчынскага ці жаночага голасу ССМТ. Распрацаваны ў такі спосаб шматгалосы сінтэзатар маўлення па тэксце можа быць добрым прыстасаваннем для хуткага стварэння аўдыёкніг, якія агучаны з захаваннем унікальных асаблівасцей персанажаў твораў.

Нягледзячы на станоўчыя вынікі ацэнкі працы атрыманых алгарытмаў на трэніровачным тэксце, падчас апрацоўкі тэставага корпусу былі высветленыя наступныя праблемы:

– разнастайнасць сімвальных кадыровак выклікае парушэнні ў працэсе ідэнтыфікацыі знакаў прыпынку;

– уніфікацыя алфавітаў (калі літары розных алфавітаў пазначаюцца аднолькавымі сімваламі) не дазваляе граматыкам знаходзіць тыя словы-пазначальнікі, у якіх ёсць устаўкі лацінскіх сімвалаў замест кірылічных;

– спалучэнні сімвалаў, якія пазначаюць пераход ад слоў моўцы да слоў аўтара ці наадварот, часам выкарыстоўваюцца ў дыялагічным тэксце для іншых мэт;

– непаўната слоўнікавых рэсурсаў для вызначэння роду моўцаў памяншае прадукцыйнасць распрацаваных алгарытмаў.

Для вырашэння пералічаных вышэй праблем патрабуецца праца над папаўненнем базы знакаў прыпынку і даданнем слоўнікавых рэсурсаў. У далейшыя планы аўтараў таксама ўваходзіць распрацоўка алгарытмаў для ідэнтыфікацыі роду моўцаў непасрэдна па словах моўцаў і для вызначэння ўзросту моўцаў па словах аўтара і моўцаў.

Спіс літаратуры

1. Гецэвіч, Ю.С. Аўтаматызаваная апрацоўка сімвальных выразаў у тэкстах для сістэмы сінтэзу беларускага маўлення / Ю.С. Гецэвіч // Інфарматыка. – 2011. – № 4. – С. 82–93.

2. AlchemyAPI Interactive Text Analysis Demo // AlchemyAPI [Electronic resource]. – 2013. – Mode of access : http://www.alchemyapi.com/api/demo.html. – Date of access : 23.07.2013.

3. Quotations Extraction // AlchemyAPI [Electronic resource]. – 2013. – Mode of access : http://www.alchemyapi.com/api/entity/quotations.html. – Date of access : 23.07.2013.

4. Assignment of Character and Action Types in Folk Tales / P. Lendvai [et al.] // Formalising Natural Languages with NooJ : Selected Papers from the NooJ 2010 Intern. Conf. / eds. Z. Gavriilidou, E. Chatzipapa, L. Papadopoulou, M. Silberzstein. – Greece: Democritus University of Thrace, 2010. – P. 102–111.

5. Jurić, T. Direct Speech Recognition in Text / T. Jurić, M. Stupar, D. Boras // Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern.

Conf. / eds. K. Vučković, B. Bekavac, M. Silberztein. – Newcastle : Cambridge Scholars Publishing, 2012. – P. 122–127.

6. Праграма камп'ютарнага запісу аўдыёкніг [Электронны рэсурс]. – Рэжым доступу : http://mp3book2005.ru/1.htm. – Дата доступу : 25.10.2013

7. Праграма AUDIOBOOK [Электронны рэсурс]. – 2013. – Рэжым доступу : http://kom-pas.narod.ru/ audiobook_net.htm. – Дата доступу : 25.10.2013

8. Гецэвіч, Ю.С. Аўтаматызацыя шматгаласавога стварэння аўдыёкніг на беларускай мове з дапамогай сінтэзатараў маўлення па тэксце / Ю.С. Гецэвіч, Т.І. Окрут, Б.М. Лабанаў // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013) : доклады XII Междунар. конф. (Минск, 20 ноября 2013 г.). – Минск : ОИПИ НАН Беларуси, 2013. – С. 60–67.

9. Hetsevich, Y. Belarusian Module for NooJ / Y. Hetsevich, S. Hetsevich, Ya. Yakubovich // NooJ web-site [Electronic resource]. – 2012. – Mode of access : http://www.nooj4nlp.net/pages/belarusian.html. – Date of access : 16.03.2012.

10. Лінгвістычны працэсар NooJ [Электронны рэсурс]. – 2002. – Рэжым доступу : http://www.nooj4nlp.net/pages/nooj.html. – Дата доступу : 01.07.2013.

11. XML TTS Tutorial (SAPI 5.3) // Microsoft Developer Network [Electronic resource]. – 2013. – Mode of access : http://msdn.microsoft.com/en-us/library/ms717077%28v= vs.85%29.aspx. – Date of access : 29.07.2013.

Паступіла 12.11.2013

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

Y.S. Hetsevich, T.I. Okrut, B.M. Lobanov

ALGORITHMS FOR IDENTIFICATION OF CUES WITH AUTHORS' TEXT INSERTIONS IN BELARUSIAN ELECTRONIC BOOKS

The main stages of algorithms for characters' gender identification in Belarusian electronic texts are described. The algorithms are based on punctuation marking and gender indicators detection, such as past tense verbs and nouns with gender attributes. For indicators, special dictionaries are developed, thus making the algorithms more language-independent and allowing to create dictionaries for cognate languages. Testing showed the following results: the mean harmonic quantity for masculine gender detection makes up 92,2 %, and for feminine gender detection – 90,4%.