OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

**УДК [004.522+004.934+004.91]:004.89**

# МЕТАД ПАБУДОВЫ КАМПАНЕНТАЎ СІНТЭЗУ МАЎЛЕННЯ ПА ТЭКСЦЕ ДЛЯ НАТУРАЛЬНА-МАЎЛЕНЧАГА ІНТЭРФЕЙСА ПРЫ ДАПАМОЗЕ NOOJ

Гецэвіч Ю.С.[*], Скопінава А.М.[*], Окрут Т.І.[*]

[*] *Аб'яднаны інстытут праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі, г. Мінск, Беларусь*

**{yury.hetsevich, skelena777, tatberrie}@gmail.com**

У дадзеным артыкуле акрэсліваецца падыход да пабудовы адасобленых кампанентаў для натуральна-маўленчага інтэрфейса інтэлектуальных сістэм праз вырашэнне камп'ютэрна-лінгвістычных задач сінтэзу маўлення, у прыватнасці праз праграмны сродак NooJ. Аўтарамі паэтапна апісаныя і алгарытмічна прадстаўленыя рашэнні задачы пошуку і класіфікавання колькасных выразаў з адзінкамі вымярэння, а таксама задачы аўтаматызаванага афармлення дыялогаў.

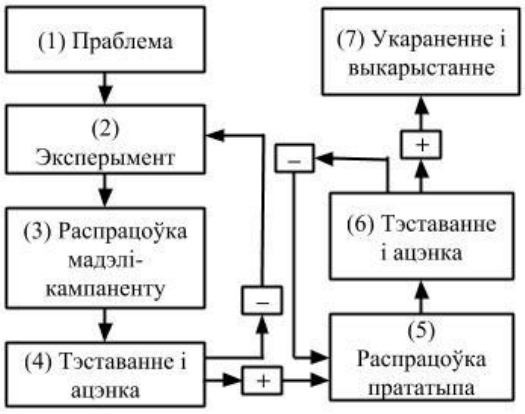
**Ключавыя словы:** кампанент; канчатковы аўтамат; NooJ; колькасны выраз з адзінкай вымярэння; ідэнтыфікацыя выразаў; агучванне дыялогаў.

## Уводзіны

Натуральна-маўленчы інтэрфейс інтэлектуальных сістэм патрабуе інтэграцыі складаных кампанентаў і модуляў для забеспячэння якаснага ўзаемадзеяння між карыстальнікам і сістэмай. Карыстальнік хоча чуць ад сістэмы агучаны тэкст самых шырокіх тэматычных даменаў у форме звычайнага маўлення. Для вырашэння гэтай агульнай задачы патрэбна звярнуцца да сінтэзатара маўлення па тэксце, які мае добра распрацаваны лінгвістычны працэсар тэксту. Аўтары артыкула прапануюць выкарыстаць праграмны сродак NooJ [NooJ, 2002] і апісаць метад для распрацоўкі якасных кампанентаў для паляпшэння лінгвістычнага працэсара сінтэзатара маўлення па тэксце праз вырашэнне розных камп'ютэрна-лінгвістычных задач.

Пад камп'ютэрна-лінгвістычнай задачай па электронным тэксце будзем разумець такую задачу (праблему), якая, па-першае, ставіцца адносна электроннага тэксту; па-другое, закранае пытанні канкрэтнага пошуку, класіфікацыі ці перапрацоўкі паслядоўнасцяў знакаў і сімвалаў жаданага электроннага тэксту; па-трэцяе, яе канчатковым рашэннем павінна быць камп'ютэрная праграма для папярэдняй апрацоўкі тэксту, праца якой можа быць праверана карыстальнікам на неабмежаванай колькасці іншых электронных тэкстаў [Гецэвіч і інш., 2013a]. Такое вызначэнне з'яўляецца абагульненым падыходам да шэрагу задач, якія фармуляваліся і вырашаліся ў працах [Гецэвіч, 2011], [Гецэвіч і інш., 2012], [Гецэвіч і інш., 2013b], [Skopinava et al., 2013].

Разгледзім паэтапна працэс ад пастаноўкі пэўнай задачы да яе рашэння з улікам камп'ютэрных сродкаў і ўмоў. На малюнку 1 відаць, што дадзены працэс прадугледжвае ажыццяўленне сямі этапаў ад вызначэння задачы да стварэння прадукта (яе рашэння) для карыстальніка. Такім чынам, спачатку даследчыкам фармулюецца праблема (1). Пасля гэтага эксперыментальнымі шляхамі (пастаноўка і аспрэчанне гіпотэз) эксперт знаходзіць рашэнні дадзенай праблемы адносна невялікага фрагмента тэксту (2); для гэтых рашэнняў мадэлююцца кампаненты, якія прадугледжваюць наяўнасць пэўных рэсурсаў і алгарытмаў у выглядзе канчатковых аўтаматаў NooJ (3) [NooJ, 2002]. Затым атрыманыя канчатковыя аўтаматы тэстуюцца на адносна вялікай колькасці тэкстаў (4). Калі вынікі тэставання паводле адзнак дакладнасці і паўнаты не здавальняючыя, то мадэлі-кампаненты вяртаюцца для дапрацоўкі на стадыю (2); у процілеглым выпадку яны перадаюцца для стварэння на іх базе эксперыментальна-праграмных комплексаў (5). Пры гэтым тэставыя дадзеныя, распрацаваныя на этапе (4), выкарыстоўваюцца і на этапах (5) і (6) для дакладнай распрацоўкі, тэставання і ацэнкі праграмнага прадукта. Пры станоўчай ацэнцы (у межах дапушчальнай памылкі) праграма трапляе ў рукі карыстальніка (7) і набывае статус канчатковага прадукта, а пры адмоўнай – адбываецца вяртанне на стадыю (5).




Малюнак 1 – Абагульненая схема працэсу вырашэння камп'ютэрна-лінгвістычнай задачы

Пад карыстальнікам будзем разумець або экспертаў-лінгвістаў, або прамых карыстальнікаў, якія могуць ужываць атрыманыя праграмныя сродкі для папярэдняй апрацоўкі тэксту. У пацверджанне выкарыстальнасці метаду, які прыводзіцца, разгледзім рашэнні дзвюх розных камп'ютэрна-лінгвістычных задач у прыкладанні да сінтэзу маўлення па тэксце.

## 1. Апрацоўка колькасных выразаў з адзінкамі вымярэння

Важна, каб натуральна-маўленчы інтэрфейс інтэлектуальнай сістэмы мог агучыць наступныя літарна-знакавыя выразы: *8,024129(3)•10^23 моль^-1, 44 мА, -95 °С, 3 тыс. гадоў назад* і г.д. Колькасныя апісанні ўласцівыя і агульнай навуковай карціне свету, і побытавай сферы жыцця. Апроч складанаструктураванасці, усюдыіснасці, яны характарызуюцца і варыятыўнасцю формаў напісання і пазначэння. Так і ўзнікае неабходнасць спецыяльна распрацаваць паўнавартасныя кампаненты алгарытмаў з рэсурсамі для апрацоўкі КВАВ у прыкладанні да сінтэзу маўлення па тэксце.

Перавагай сістэмы NooJ датычна задачы пошуку і апрацоўкі КВАВ з'яўляецца тое, што распрацаваныя з дапамогай убудаванага візуальнага рэдактара і ў выглядзе канчатковых аўтаматаў кампаненты валодаюць навочнасцю, дзякуючы чаму іх можна адносна лёгка карэктаваць і папаўняць, што выключна важна ў сілу вышэй пералічаных уласцівасцяў КВАВ.

Вынікаючы з вышэй апісанай у частцы 1 схемы вырашэння камп'ютэрна-лінгвістычнай праблемы, для пачатку выразна сфармулюем задачу (1): знайсці, класіфікаваць і апрацаваць у электронных тэкстах колькасныя выразы з адзінкамі вымярэння. На этапе (2) экспертам праводзіліся назіранні і аналіз КВАВ адносна іх будовы і выкарыстання на матэрыяле электронных тэкстаў навукова-тэхнічнага і прававога тэматычнага дамена на беларускай і рускай мовах. На дадзены момант на этапе (3) аўтарамі змадэляваныя тры ўзаемадапаўняльныя складаныя кампаненты з алгарытмамі і рэсурсамі для пошуку колькасных выразаў з адзінкамі вымярэння ў вялікіх карпусах тэкстаў, якія дазваляюць: знаходзіць КВАВ і класіфікаваць іх па трох тыпах паводле міжнароднай сістэмы адзінак СІ (асноўныя, вытворныя, пазасістэмныя); знаходзіць КВАВ з метралагічнымі прыстаўкамі (кратнымі ці дольнымі, скарочанымі ці ў поўнай форме) і класіфікаваць іх паводле словаўтваральных асаблівасцяў; пераўтвараць КВАВ у арфаграфічныя словы (малюнак 2).

Этап тэставання (4) паказаў, што, напрыклад, першы кампанент дае пошукавыя вынікі з дакладнасцю ў 72 %. Гэты адносна высокі паказчык даў падставу для распрацоўкі эксперыментальнага праграмнага комплексу (5), які б знаходзіў КВАВ і класіфікаваў іх па трох тыпах адносна міжнароднай сістэмы адзінак СІ (асноўныя, вытворныя, пазасістэмныя). Праграма атрымала назву "QEMU Identificator", г.зн. "Quantitative Expressions with Measurement Units Identificator" ці "Ідэнтыфікатар колькасных выразаў з адзінкамі вымярэння".

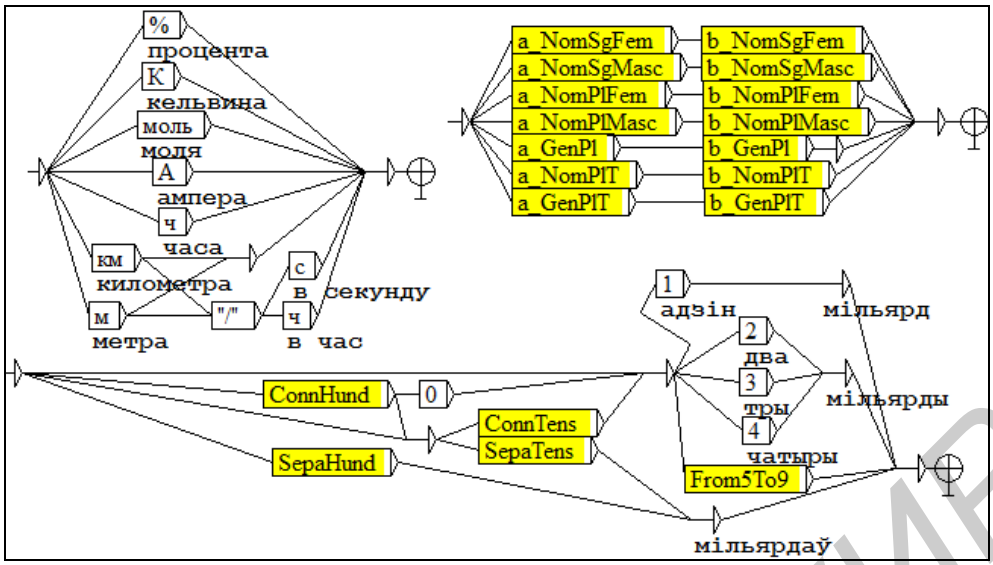
Разгледзім працу QEMU Identificator на прыкладзе аналізу фрагментаў беларуска- і рускамоўных тэкстаў з адпаведных навукова-тэхнічных тэкставых карпусоў. Тэксты для аналізу можна ўводзіць адвольна ў верхняе поле або загружаць тэкставыя файлы (у фармаце txt) праз адпаведную каманду ва ўкладцы меню.

Каманда *"Пераўтварыць" (Transform)* ажыццяўляе пошук КВАВ і прысвойвае ім пэўныя маркеры. Да прыкладу, на малюнку 3а было ідэнтыфікавана *40°С* як *<MEAS+Temperature in Celsius scale+D>*. Гэта азначае, што дадзены літарна-знакавы набор з'яўляецца колькасным выразам з адзінкай вымярэння тэмпературы па шкале Цэльсія, прычым па стандартах сістэмы СІ дадзеная адзінка з'яўляецца вытворнай.

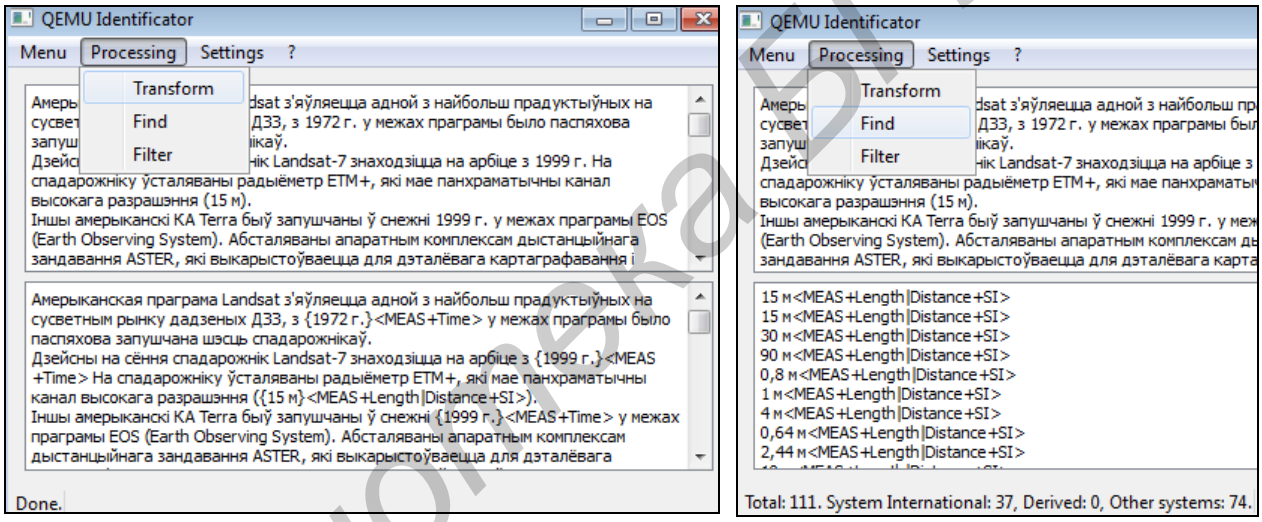
Каманда *"Знайсці" (Find)* на выйсці выдае спіс усіх знойдзеных КВАВ і падае колькасныя звесткі (малюнак 3b). У дадзеным тэставым тэксце знайшлося 111 КВАВ, з якіх 37 выразаў уключаюць асноўныя мерныя адзінкі СІ, 0 выразаў з вытворнымі ад СІ адзінкамі, а 74 колькасныя выразы змяшчаюць пазасістэмныя адзінкі.

Выкананне каманды *"Фільтраванне" (Filter)* дае магчымасць здзяйсняць разнастайныя пошукавыя запыты праз фармальную мову рэгулярных выразаў. Напрыклад, пасля ўводу *Voltage|Frequency* спіс знойдзеных КВАВ абмяжоўваецца толькі тымі, якія ўтрымоўваюць адзінкі вымярэння або электрычнай напругі, або чашчыні току (малюнак 3с): *1 Гц, 50 В, 150 В*.

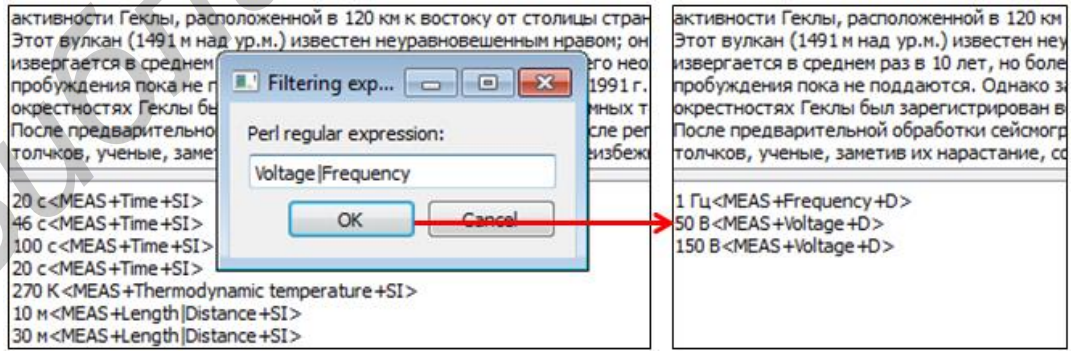
Зараз праграма QEMU Identificator тэстуецца распрацоўшчыкамі і карыстальнікамі. У далейшым плануецца палепшыць гэты праграмны прадукт, а таксама праграмна рэалізаваць астатнія кампаненты-мадэлі для пошуку колькасных выразаў з адзінкамі вымярэння.




Малюнак 2 – Кампанент для пераўтварэння колькасных выразаў з адзінкамі вымярэння ў арфаграфічныя паслядоўнасці слоў




*(a)* *(b)*




*(c)*

Малюнак 3 – Выкананне каманд *“Пераўтварыць” (a)*, *“Знайсці” (b)*, *“Фільтраванне” (c)* для ідэнтыфікацыі КВАВ у электронных тэкстах на беларускай *(a, b)* і рускай *(c)* мовах праз убудаваныя ў праграму QEMU Identificator марфалагічныя і сінтаксічныя кампаненты NooJ

## 2. Ідэнтыфікацыя і шматгалосая агучка дыялогаў у электронных тэкстах

Дзякуючы ўбудаванай у NooJ магчымасці анатаваць з дапамогай сінтаксічных кампанентаў электронны тэкст стала магчымым рашэнне яшчэ адной камп'ютэрна-лінгвістычнай задачы – ідэнтыфікаваць і агучваць рознымі сінтэзаванымі галасамі дыялогі розных персанажаў. Гэта дазволіць зрабіць натуральна-маўленчы інтэрфейс больш адпаведным семантычнаму сэнсу, які ён перадае.

Першым крокам для вырашэння дадзенай праблемы была распрацоўка кампанентаў для аўтаматычнай ідэнтыфікацыі простай мовы ў тэксце. Эксперты прааналізавалі тэкставы матэрыял на прадмет выяўлення структур афармлення простай мовы, якія пасля закладваліся ў сінтаксічны кампанент NooJ [Гецэвіч і інш., 2013с]. Атрыманыя вынікі ляглі ў аснову кампанента *DS_All* (малюнак 4), які спрацоўвае тады, калі простая мова пачынаецца з працяжніка, затым ідуць словы персанажа; пераход ад слоў аўтара і наадварот пазначаецца праз камбінацыю працяжніка з коскай, кропкай, клічнікам, пытальнікам ці іх спалучэннямі:

1. Словы персанажа (М) без слоў аўтара (А): *M (! | !! | !!! | ? | ?! | ... | .).*
2. Словы персанажа са словамі аўтара ў канцы: *M (, | ! | !! | !!! | ? | ?! | ... | . ) – A ( ... | . ).*
3. Словы персанажа з адной ці некалькімі аўтарскімі ўстаўкамі: *M (, | ! | !! | !!! | ? | ?! | ... | . ) – A (, | ... | . | : | . ) – M (, | ! | !! | !!! | ? | ?! | ... | . ) (– A (, | ... | . | : | . ) – M (, | ! | !! | !!! | ? | ?! | ... | . )).*

Калі пасля слоў персанажа ідуць словы аўтару, алгарытм працягвае пошук завяршэння фразы дыялогу. У падграфах *Speaker* і *Author* асобна разглядаюцца ўнутраныя знакі прыпынку, прычым аўтарскі тэкст мае іх меншую варыянтнасць: коска, кропка, шматкроп'е і дужкі.

Наступным крокам стала ідэнтыфікацыя роду персанажаў пасродкам аналізу слоў аўтара, якія характарызуюцца наяўнасцю такіх індыкатараў, як дзеясловы мінулага часу адзіночнага ліку (*казаў, казала*); імёны ўласныя (*Алесь, Майка*); назоўнікі, якія абазначаюць размоўцу (*бацька, дзяўчынка*). З трэніравальнага тэксту былі вынятыя рэплікі з аўтарскімі ўстаўкамі і размечаныя па родзе. Затым з улікам знойдзеных індыкатараў быў створаны канчатковы аўтамат для ідэнтыфікацыі роду. Для гэтага ў падграф *Author* былі дададзеныя родазалежныя падграфы *VERBSfeminine* (малюнак 5) і *VERBSmasculine* для вызначэння жаночага і мужчынскага роду. Пашырэнне спісу дзеясловаў-індыкатараў прывяло да стварэння цэлага рэсурсу ў выглядзе лінгвістычнага слоўніку (малюнак 6). У ім парамі прадстаўленыя дзеясловы мінулага часу ў формах жаночага і мужчынскага роду.

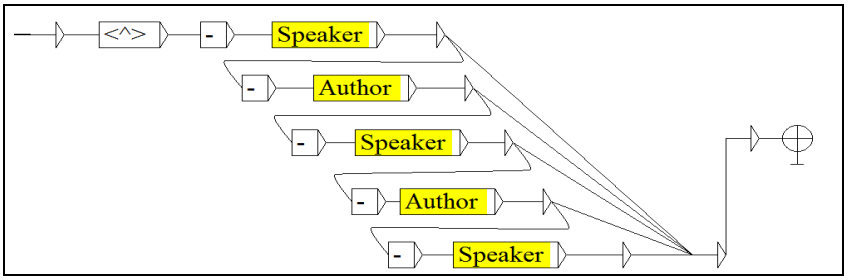
Падчас распрацоўкі алгарытмаў даследавалася і праблема граматычных амонімаў (амаформаў). Так, дзеяслоў *кажа* можа належаць і мужчынскаму, і жаночаму роду. У выніку праз такія формы паасобку нельга высветліць катэгорыю роду персанажа. Таму быў створаны дадатковы звязак графаў "дзеяслоў-назоўнік", дзе першы граф уключае граматычныя амаформы маўленчых дзеясловаў, а другі граф падае спіс выражаных назоўнікамі паказчыкаў роду. Атрыманыя графы прымяняюцца паслядоўна да мэтавага тэксту праз NooJ, прычым пазнакі аднаго захоўваюцца падчас спрацоўвання другога. Такім чынам, на выйсці тэкст атрымлівае разметку па рэпліках для мужчынскага і жаночага галасоў.

Пасля таго як дакладнасць кампанентаў на трэніравальным тэкставым матэрыяле (звыш 100 тыс. словаўжыванняў з твору У.С. Караткевіча "Каласы пад сярпом тваім") дасягнула больш 95%, быў створаны тэставы корпус з урыўкаў твораў мастацкай літаратуры аб'ёмам каля 24 тыс. словаўжыванняў. Па падліках эксперта тэставы корпус утрымоўвае 481 рэпліку (рэплікі з працяжнікам), 165 рэплік персанажаў мужчынскага роду і 68 рэплік персанажаў жаночага роду. Тэставанне на тэкставым корпусе паказала, што кампаненты ідэнтыфікуюць простую мову ў электронных тэкстах і вызначаюць род персанажаў з дакладнасцю вышэй за 90%, што дазваляе пачаць іх укараненне ў сістэму сінтэзу маўлення па тэксце (СМТ). Па-першае, каб выкарыстанне ў сістэме СМТ пад стандарт SAPI 5.1 стала магчымым, неабходна прывесці тэксты да выгляду SAPI TTS XML [XML TTS Tutorial, 2013]. Па-другое, для выбару сістэмай адпаведнага голасу сінтаксічная анатацыя, якую генеруе кампанент, павінна быць адаптаваная пад наступны код:

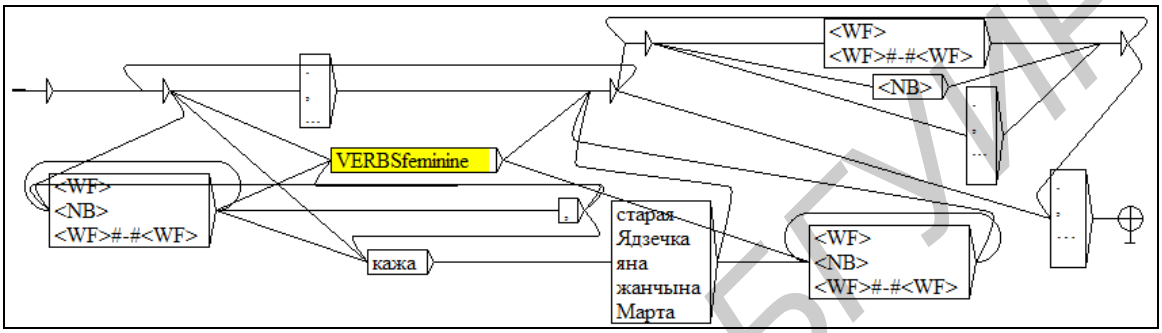
*<VOICE Required="name=[Назва голасу ў сістэме]">…Тэкст для агучвання…</VOICE>*.

Для гэтага ў падграфы *DS_M* і *DS_F* былі ўстаўленыя маркеры пазначэння шляхоў, па якіх спрацоўваюць кампаненты. Маркеры наладжаныя так, што неразмечаны тэкст (малюнак 7a) і словы аўтара агучваюцца голасам *AlesiaBel*, мужчынскія рэплікі – голасам *BorisBel*, а жаночыя – *ElenaBel*. Апрацаваныя кампанентамі сказы набываюць выгляд, які дэманструецца малюнкам 7b. Нарэшце ўжо размечаны тэкст можна падаваць на ўваход сістэмы СМТ. На малюнку 7c адлюстравана, як праграма SAPI5 TTSAPP аўтаматычна пераключае пастаўленыя ў сістэме галасы *AlesiaBel*, *BorisBel*, *ElenaBel* з дапамогай опцыі *Process XML*.

Такім чынам, распрацаваныя мадэлі-алгарытмы паказалі станоўчыя вынікі ў спалучэнні з сістэмай СМТ і ў далейшым могуць выкарыстоўвацца для стварэння дадатковага блока аўтаматычнага выбару мужчынскага альбо жаночага голасу сістэмы СМТ. Шматгалосы сінтэзатар маўлення па тэксце дасць магчымасць натуральна-маўленчаму інтэрфейсу захоўваць індывідуальныя асаблівасці персанажаў, якія гавораць між сабой у паведамленні.




Малюнак 4 – Сінтаксічны кампанент для аўтаматычнай ідэнтыфікацыі простай мовы ў тэксце




Малюнак 5 – Сінтаксічны падкампанент для вызначэння рэплік персанажаў жаночага роду

```
трымаў,VERB+SpeechAct+Masculine
трымала,VERB+SpeechAct+Feminine
ударыў,VERB+SpeechAct+Masculine+FLX=ЎVERB1
ударыла,VERB+SpeechAct+Feminine+FLX=ЎVERB1
```

Малюнак 6 – Фрагмент слоўніка індыкатараў роду, выражаных дзеясловамі мінулага часу

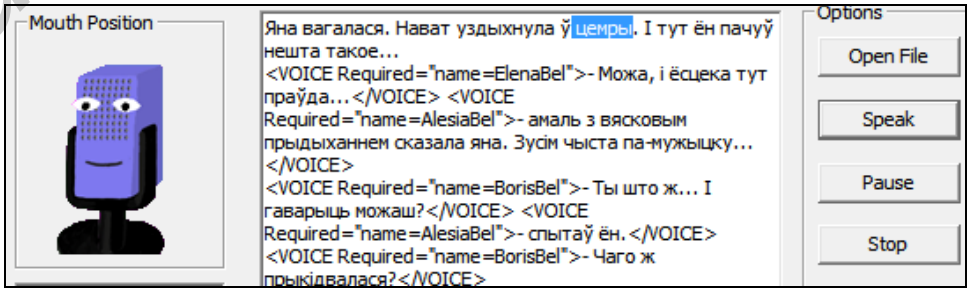
*(a)*

```
- Можа, і ёсцека тут праўда... - амаль з вясковым прыдыханнем сказала яна. Зусім чыста па-мужыцку
- Ты што ж... І гаварыць можаш? - спытаў ён. - Чаго ж прыкідвалася?
```

*(b)*

```
2 <VOICE Required="name=ElenaBel">- Можа, і ёсцека тут праўда...</VOICE> <VOICE Required="name=AlesiaBel">- амаль
  з вясковым прыдыханнем сказала яна. Зусім чыста па-мужыцку...</VOICE>
3 <VOICE Required="name=BorisBel">- Ты што ж... І гаварыць можаш?</VOICE> <VOICE Required="name=AlesiaBel">-
  спытаў ён.</VOICE> <VOICE Required="name=BorisBel">- Чаго ж прыкідвалася?</VOICE>
```

*(c)*




Малюнак 7 – Працэс ад анатавання электроннага тэксту *(a)* праз VoiceXML *(b)* для аўтаматычнага пераключэння галасоў да шматгалосага агучвання *(c)* аўтаматычна размечанага тэксту

## Заключэнне

Такім чынам, быў прапанаваны метад пабудовы адасобленых кампанентаў для натуральна-маўленчага інтэрфейса інтэлектуальных сістэм праз вырашэнне дзвюх камп'ютэрна-лінгвістычных задач сінтэзу маўлення. Для гэтага выкарыстоўваўся наладжвальны лінгвістычны працэсар NooJ.

Распрацаваныя мадэлі рашэнняў з'яўляюцца самастойнымі незалежнымі кампанентамі, кожны з якіх можа або выкарыстоўвацца паасобку, або ўсе кампаненты разам могуць быць убудаваныя ў іншыя сістэмы. У будучыні аўтарамі плануецца павышэнне дакладнасці працы кампанента ўжо генеравання арфаграфічнага тэксту па колькасных выразах з адзінкамі вымярэння, а таксама распрацоўка кампанента ідэнтыфікацыі роду персанажаў у рэпліках без уставак слоў аўтару.

## Бібліяграфічны спіс


**[Гецэвіч і інш., 2013a]** Гецэвіч, Ю.С. Мадэляванне і распрацоўка сістэм пошуку колькасных выразаў з адзінкамі вымярэння ў электронных тэкстах на беларускай і рускай мовах / Ю.С. Гецэвіч, А.М. Скопінава, А.Ф. Есіс // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013) : доклады XII Международной конференции (Минск, 20 ноября 2013 г.). – Минск : ОИПИ НАН Беларуси, 2013. – С. 282–287.

**[Гецэвіч, 2011]** Гецэвіч, Ю.С. Аўтаматызаваная апрацоўка сімвальных выразаў у тэкстах для сістэмы сінтэзу беларускага маўлення / Ю.С. Гецэвіч // Информатика. – 2011. – № 4. – С. 82–93.

**[Гецэвіч і інш., 2012]** Гецэвіч, Ю.С. Ідэнтыфікацыя выразаў з адзінкамі вымярэння ў навукова-тэхнічных і прававых тэкстах на беларускай і рускай мовах / Ю.С. Гецэвіч, А.М. Скопінава // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2012) : доклады XI Междунар. конф., Минск, 15 нояб. 2012 г. – Минск : ОИПИ НАН Беларуси, 2012. – С. 260–265.

**[Гецэвіч і інш., 2013b]** Гецэвіч, Ю.С. Кампаненты ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў тэкстах на беларускай і рускай мовах / Ю.С. Гецэвіч, А.М. Скопінава // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS–2013) : материалы III Междунар. науч.-техн. конф., Минск, 21–23 февр. 2013 г. – Минск : БГУИР, 2013 г. – С. 319–328.

**[Skopinava et al., 2013]** Skopinava, A.M. Processing of quantitative expressions with units of measurement in scientific texts as applied to Belarusian and Russian text-to-speech synthesis / Yu. S. Hetsevich, A.M. Skopinava, B.M. Lobanov // Компьютерная лингвистика и интеллектуальные технологии : материалы Междунар. конф. «Диалог», Московская обл., г. Бекасово, 29 мая – 2 июня 2013 г. – Вып. 12 (19). – В 2 т. – Т.1. – М. : Изд-во РГГУ, 2013. – С. 634–651.

**[NooJ, 2002]** Лінгвістычны працэсар NooJ [Электронны рэсурс]. – 2002. – Рэжым доступу : http://www.nooj4nlp.net /pages/nooj.html. – Дата доступу : 01.07.2012.

**[Гецэвіч і інш., 2013c]** 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 Международной конференции (Минск, 20 ноября 2013 г.). – Минск : ОИПИ НАН Беларуси, 2013. – С. 269–276.

**[XML TTS Tutorial, 2013]** 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.


## METHOD OF CONSTRUCTING TEXT-TO-SPEECH COMPONENTS FOR NATURAL LANGUAGE INTERFACES WITH THE HELP OF NOOJ


Hetsevich Yu.S.[*], Skopinava A.M.[*], Okrut T.I.[*]

[*]*United Institute of Informatics Problems of the National Academy of Sciences, Minsk, Republic of Belarus*

**{yury.hetsevich, skelena777, tatberrie}@gmail.com**



This article outlines an approach to construction of detached components for natural language interfaces through solution of computer-linguistic problems by means of the linguistic processor NooJ. The authors focus on two different problems and describe stage-by-stage solutions to them.


### INTRODUCTION

In order to make text interfaces more "natural", systems of human computer interaction should be able to voice electronic texts. High-quality text-to-speech synthesis cannot be achieved without solving various computer-linguistic problems. Under the term "a computer-linguistic problem", we mean a task, which refers to electronic texts; and concerns identifying, classifying, and processing of words and symbol sequences; to solve it means to develop a program for preliminary text processing.

### MAIN PART

The first problem consists in processing quantitative expressions with measurement units (QEMU). By the moment three complementary blocks of components and resources have been built for Belarusian and Russian. They allow: identifying and classifying QEMU according to the International Bureau of Weights and Measures (expressions with SI-basic, SI-derived, and off-system measurement units); classifying QEMU according to word formation peculiarities (full or shortened, multiple or submultiple prefixes); expanding QEMU into orthographic words.

The second problem lies in many-voiced scoring of dialogues in Belarusian and Russian. The authors have created a component with algorithms and resources, which choose and switch voices automatically for systems of text-to-speech synthesis. Multi-voiced scoring of dialogues allows voicing electronic books, and preserving sound distinction among personages.

### CONCLUSION

We have described a general approach to construction of components for natural language interfaces with the help of NooJ. In future it is planned to expand the resources in order to cover wider groups of measurement units, and to extract more gender indicators.