XII Международная конференция

РАЗВИТИЕ ИНФОРМАТИЗАЦИИ И ГОСУДАРСТВЕННОЙ СИСТЕМЫ НАУЧНО-ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ

РИНТИ-2013

20 ноября 2013 года, Минск

ДОКЛАДЫ

УДК 002; 004

Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013) : доклады XII Международной конференции, Минск, 20 ноября 2013 г. – Минск : ОИПИ НАН Беларуси, 2013. – 412 с. – ISBN 978-985-6744-78-8.

Представлены доклады XII Международной конференции «Развитие информатизации и государственной системы научно-технической информации» (РИНТИ-2013), Минск, 20 ноября 2013 г., в которых рассматриваются вопросы научно-методического, информационного и технологического обеспечения развития информатизации, создания и развития автоматизированных систем научно-технической информации, корпоративных библиотечно-информационных систем и технологий, а также направления использования информационно-коммуникационных технологий и результаты научных исследований в данных областях.

Материалы конференции будут полезны специалистам в области информационно-коммуникационных технологий, занимающимся разработкой и внедрением автоматизированных информационных систем управления, развитием информационной инфраструктуры Беларуси, реализацией проектов государственных программ в сфере информатизации и систем научно-технической информации.

Одобрены программным комитетом и печатаются по решению редакционной коллегии Объединенного института проблем информатики Национальной академии наук Беларуси в виде, представленном авторами.

Научные редакторы:

доктор физико-математических наук, профессор А.В. Тузиков;
кандидат технических наук, доцент Р.Б. Григянец;
кандидат технических наук В.Н. Венгеров

ISBN 978-985-6744-78-8

УДК [004.522+004.934+004.91] : 004.89

МАДЭЛЯВАННЕ І РАСПРАЦОЎКА СІСТЭМ ПОШУКУ КОЛЬКАСНЫХ ВЫРАЗАЎ З АДЗІНКАМІ ВЫМЯРЭННЯ Ў ЭЛЕКТРОННЫХ ТЭКСТАХ НА БЕЛАРУСКАЙ І РУСКАЙ МОВАХ

Ю.С. Гецэвіч[1], А.М. Скопінава[1], А.Ф. Есіс[2]
[1]Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск;
[2]Беларускі дзяржаўны ўніверсітэт інфарматыкі і радыёэлектронікі, Мінск

Разгледжаны працэс пабудовы алгарытмічных мадэляў і эксперыментальна-праграмнага комплексу для вырашэння задачы пошуку і класіфікацыі колькасных выразаў з адзінкамі вымярэння на матэрыяле беларуска- і рускамоўных электронных тэкстаў.

Уводзіны

Электронныя тэксты навукова-тэхнічнай тэматыкі часта ўтрымліваюць колькасныя выразы з адзінкамі вымярэння (КВАВ), напрыклад 27 %, 110 м/с, 15,25 кг і інш. Яны павінны быць дакладна пабудаваны і ўжыты, каб карыстальнік тэксту мог правільна і адназначна зразумець колькасны бок пэўнага якаснага паказчыка.

Падчас апрацоўкі колькасных выразаў з адзінкамі вымярэння ўзнікаюць разнастайныя задачы і праблемы ў наступных сферах [1–3]: *выдавецкія ўстановы* (праблема аўтаматызаванай лакалізацыі канкрэтнага спісу выразаў з адзінкамі вымярэння і задача хуткай праверкі правільнасці ўжыванняў разгорнутых формаў назваў адзінак вымярэння ў тэкстах); *сістэмы сінтэзу маўлення па тэксце* (праблема правільнай генерацыі арфаграфічнага тэксту па ўваходным тэксце); *сістэмы пошуку і апрацоўкі інфармацыі для бібліятэк, Інтэрнэту, патэнтавых базаў дадзеных* (задача фармавання дакладных пошукавых запытаў; задачы аўтаматычнага рэферавання і анатавання інфармацыі).

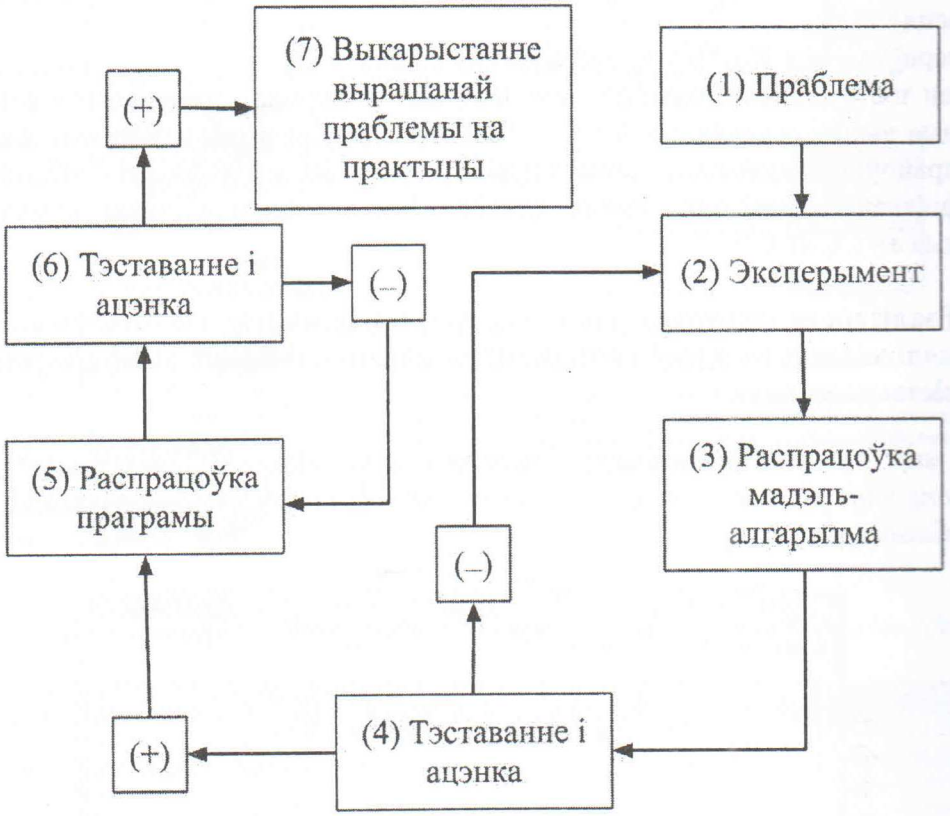
У дакладзе разгледжаны абагульнены падыход да рашэння камп'ютарна-лінгвістычных задач на электронных тэкстах на прыкладзе задачы пошуку колькасных выразаў з адзінкамі вымярэння ў навукова-тэхнічных тэкстах.

1. Агульны падыход да рашэння камп'ютарна-лінгвістычных задач па электронных тэкстах

Пад *камп'ютарна-лінгвістычнай задачай па электронным тэксце* будзем разумець такую задачу (праблему), якая, па-першае, ставіцца адносна электроннага тэксту, па-другое, тычыцца пытанняў канкрэтнага пошуку, класіфікацыі ці перапрацоўкі паслядоўнасцяў электронных сімвалаў мэтавага электроннага тэксту, па-трэцяе, яе канчатковым рашэннем павінна быць камп'ютарная праграма, праца якой можа быць праверана карыстальнікам на неабмежаванай колькасці іншых электронных тэкстаў. Такое азначэнне з'яўляецца абагульненым падыходам да шэрагу задач, якія ставіліся і вырашаліся ў працах [1–3, 5]. Прагледзім працэс ад пастаноўкі канкрэтнай задачы да яе вырашэння з улікам камп'ютарных сродкаў і ўмоваў. На мал. 1 бачна, што дадзены працэс прадугледжвае ажыццяўленне шасці этапаў ад вызначэння задачы да стварэння прадукту (яе вырашэння) для карыстальніка.

Такім чынам, спачатку перад даследчыкам ставіцца праблема (1). Пасля эксперыментальным шляхам эксперт знаходзіць рашэнні праблемы адносна невялікага

фрагменту тэксту (2), для гэтых рашэнняў распрацоўваецца дзейсны мадэль-алгарытм з дапамогай канчатковых аўтаматаў NooJ (3) [4]. Далей распрацаваныя канчатковыя аўтаматы тэстуюцца на большай колькасці тэкстаў (4).



Мал. 1. Працэс вырашэння камп'ютарна-лінгвістычнай задачы

Калі вынікі тэставання паводле ацэнак дакладнасці і паўнаты не задавальняюць, то мадэль-алгарытм вяртаецца для дапрацоўкі на стадыю (2), у супрацьлеглым выпадку ён перадаецца для стварэння на яго базе эксперыментальна-праграмнага комплексу (5). Прычым тэставыя дадзеныя, якія былі распрацаваны на этапе (4), перадаюцца на этапы (5) і (6) для дакладнай распрацоўкі і тэставання праграмнага прадукта. Пасля распрацоўкі праграмы (5) адбываюцца яе тэставанне і ацэнка дакладнасці (6). Пры станоўчай ацэнцы (ў межах дапушчальнай памылкі) праграма трапляе ў рукі да карыстальніка (7) і набывае статус канчатковага прадукту, а пры адмоўнай – адбываецца вяртанне на стадыю (5).

2. Пошук колькасных выразаў з адзінкамі вымярэння як камп'ютарна-лінгвістычная задача

Спачатку паводле схемы на мал. 1 была вызначана канкрэтная задача (1): знайсці і класіфікаваць у электронных тэкстах КВАВ. На этапе (2) экспертам праводзіліся назіранні і аналіз колькасных выразаў з адзінкамі вымярэння адносна іх будовы і выкарыстання на мэтавым матэрыяле электронных тэкстаў на беларускай і рускай мовах. На дадзены момант паводле этапу (3) аўтарамі артыкулу змадэляваны тры ўзаемадапаўняючыя алгарытмы для пошуку КВАВ у вялікіх корпусах тэкстаў, якія дазваляюць:

– знаходзіць КВАВ і класіфікаваць іх па трох тыпах паводле Міжнароднай сістэмы адзінак СІ (СІ, вытворныя ад СІ, не СІ) [2];

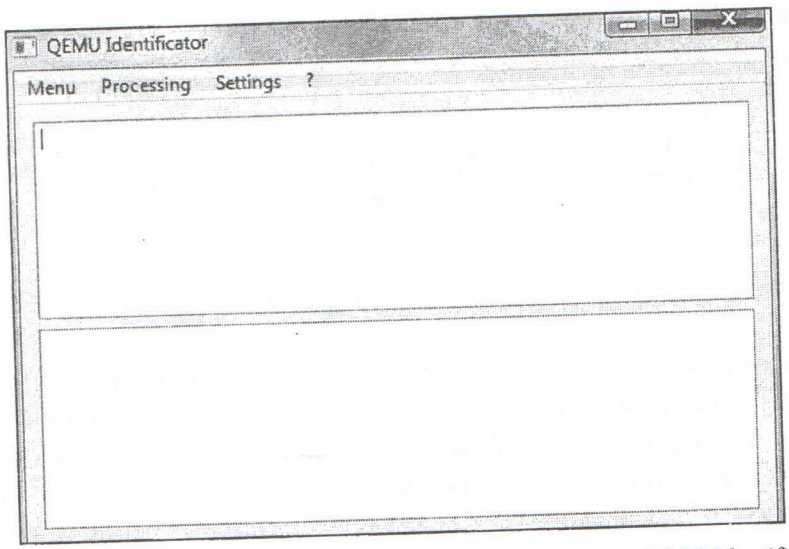
– знаходзіць КВАВ з метралагічнымі прыстаўкамі (кратнымі ці дольнымі, скарочанымі ці ў поўнай форме) і класіфікаваць іх паводле словаўтваральных асаблівасцяў [3];

– пераўтвараць КВАВ у арфаграфічныя словы [1, 3].

Этап тэставання (4) паказаў, што, напрыклад, першы алгарытмічны комплекс 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 (СІ, вытворныя ад СІ, не СІ).

3. Рэалізацыя эксперыментальнай праграмы пошуку і ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння паводле Міжнароднай сістэмы адзінак

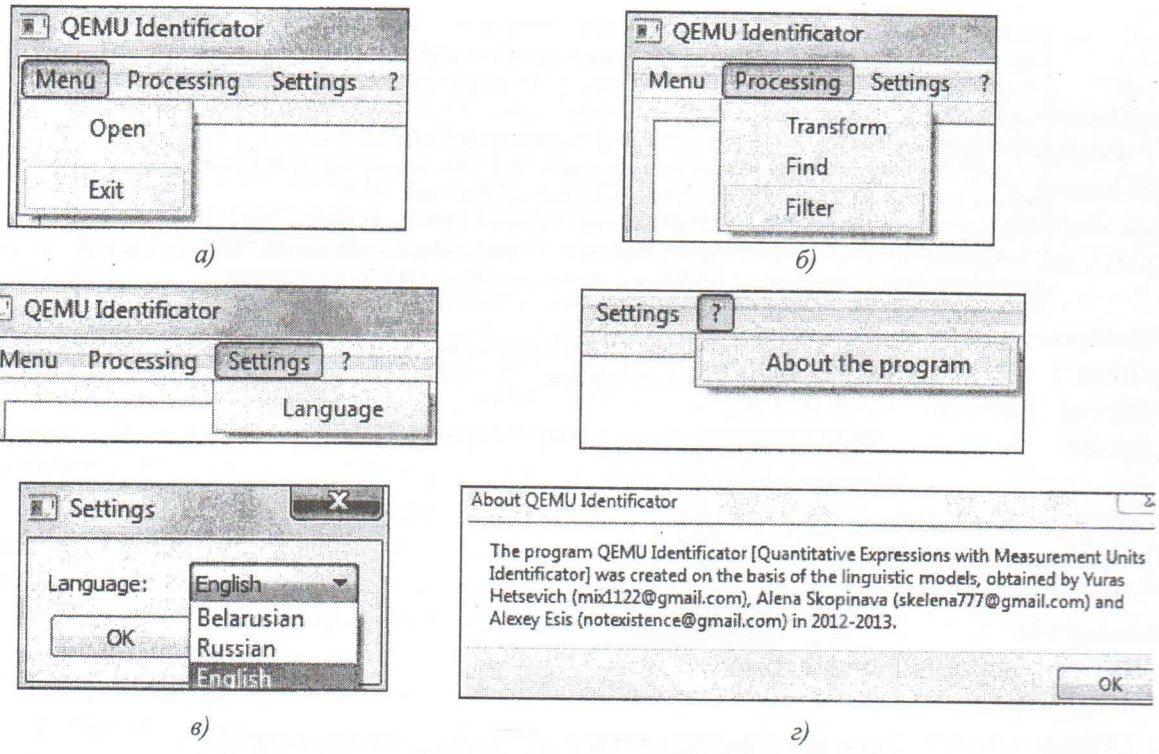
На мал. 2 дэманструецца запуск праграмы з назвай *QEMU Identificator* (*Quantative Expressions with Measurement Units Identificator* – Ідэнтыфікатар колькасных выразаў з адзінкамі вымярэння).



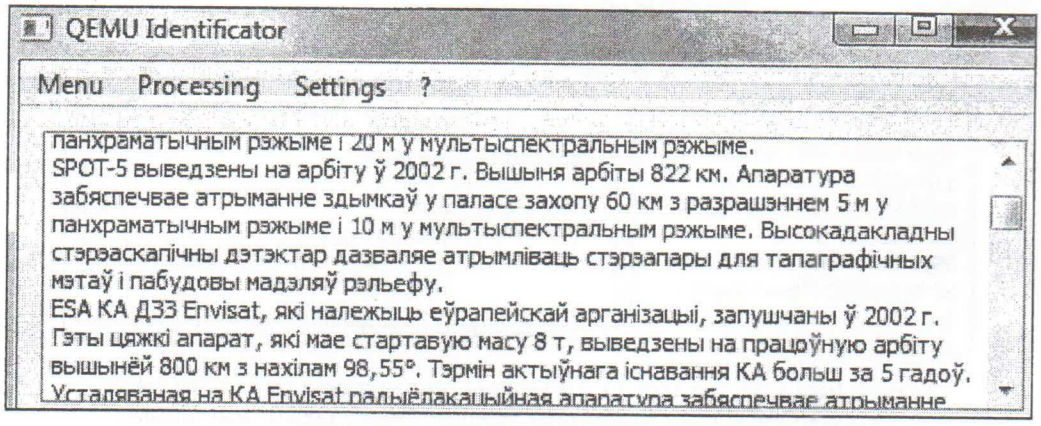
Мал. 2. Запуск эксперыментальна-праграмнага комплексу *QEMU Identificator*

Панэль кіравання складаецца з наступных укладак: *Menu* (Меню), *Processing* (Апрацоўка), *Settings* (Налады) і (даведка) (мал. 3). Праз меню карыстальнік можа адкрыць файл (*Open*) і выйсці з праграмы (*Exit*) (мал. 3, *а*). Укладка апрацоўкі прадугледжвае аналіз тэксту праз каманды пераўтварыць (*Transform*), знайсці (*Find*) і фільтраваць (*Filter*) (мал. 3, *б*). Праз укладку налад даецца магчымасць выбраць мову інтэрфейсу (*Language*): беларускую (*Belarusian*), рускую (*Russian*) або англійскую (*English*) (мал. 3, *в*). Нарэшце ў даведцы падаюцца агульная інфармацыя аб праграме і кантактныя дадзеныя распрацоўшчыкаў (*About the program*) (мал. 3, *г*).

Разгледзім працу *QEMU Identificator* на прыкладзе апрацоўкі фрагменту тэксту з навукова-тэхнічнага беларускамоўнага тэкставага корпусу. Тэксты для аналізу можна ўводзіць адвольна ў верхняе поле (мал. 2), а можна адкрываць гатовыя тэкставыя фрагменты ў фармаце txt праз адпаведную каманду (мал. 3, *а*). У нашым выпадку возьмем гатовы (мал. 4).



Мал. 3. Знаёмства з інтэрфейсам эксперыментальна-праграмнага комплексу *QEMU Identificator*



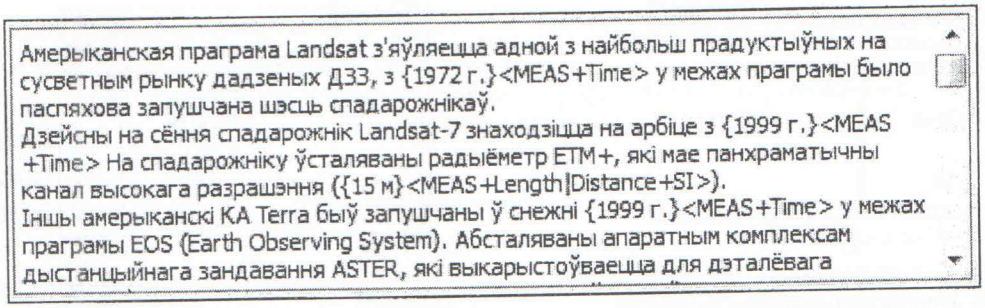
Мал. 4. Фрагмент навукова-тэхнічнага тэксту для аналізу праз *QEMU Identificator*

Прадэманструем, што адбываецца пры выкарыстанні каманд:

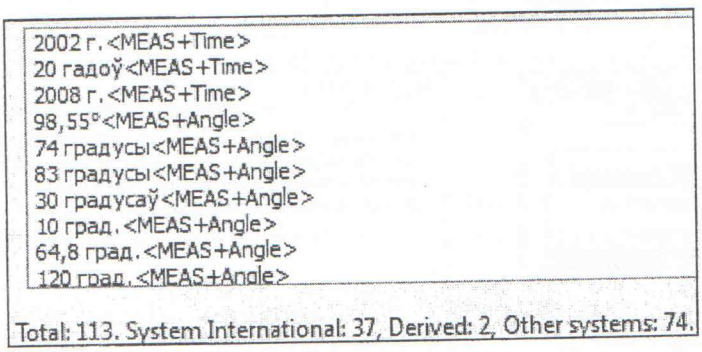
Transform – знойдзеным КВАВ надаюцца пэўныя маркеры, вынік адлюстроўваецца ў ніжнім полі інтэрфейсу; да прыкладу: у тэксце сустрэлася *15 м*, праграма яго пераўтварыла ў *{15 м}<MEAS+Length|Distance+SI>* – гэта азначае, што дадзены літарна-сімвальны набор з'яўляецца колькасным выразам з адзінкай вымярэння даўжыні або адлегласці і гэта адзінка належыць сістэме СІ (мал. 5);

Find – выводзіцца спіс усіх знойдзеных КВАВ і ніжэй падаюцца статыстычныя звесткі: агульная колькасць КВАВ, іх разнастайнасць; да прыкладу: у тэставым тэксце знайшлося 113 КВАВ, з якіх 37 выразаў змяшчаюць сістэмныя мерныя адзінкі, 2 – вытворныя і 74 – пазасістэмныя (мал. 6);

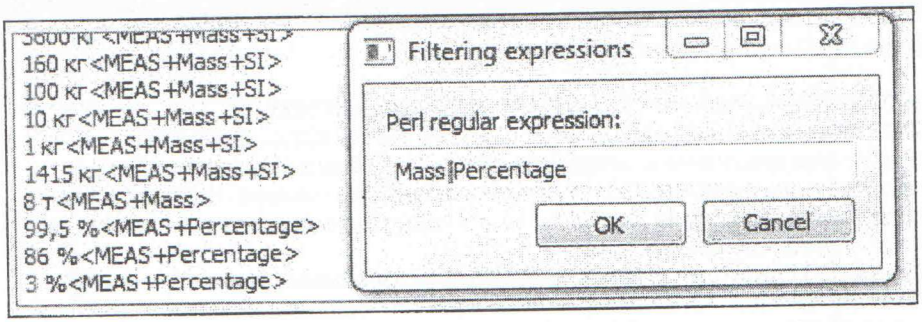
Filter – даецца магчымасць здзяйсняць разнастайныя пошукавыя запыты праз фармальную мову рэгулярных выразаў; напрыклад, праз выраз *Mass|Percentage* можна атрымаць адразу спіс КВАВ і масы, і працэнта (мал. 7).



Мал. 5. Маркіраванне КВАВ у навукова-тэхнічным тэксце праз каманду *Transform* у *QEMU Identificator*



Мал. 6. Вынікі пошуку КВАВ у навукова-тэхнічным тэксце праз каманду *Find* у *QEMU Identificator*



Мал. 7. Вынікі фільтрацыі КВАВ масы ці працэнта праз рэгулярны выраз у навукова-тэхнічным тэксце праз каманду *Filter* у *QEMU Identificator*

Заключэнне

Разгледжаны абагульнены падыход да рашэння камп'ютарна-лінгвістычных задач па электронных тэкстах на прыкладзе задачы пошуку колькасных выразаў з адзінкамі вымярэння ў навукова-тэхнічных тэкстах.

Распрацаваны праграмны эксперыментальны комплекс дазваляе зручна шукаць колькасныя выразы з адзінкамі вымярэння сістэмы СІ і вытворных ад СІ у электронных тэкстах на беларускай ці рускай мовах. Зараз гэтая рэалізацыя тэстуецца распрацоўшчыкамі і карыстальнікамі. На базе вынікаў, якія будуць атрыманы, плануецца ў далейшым палепшыць гэты праграмны прадукт, а таксама праграмна рэалізаваць астатнія алгарытмічныя мадэлі для пошуку колькасных выразаў з адзінкамі вымярэння.

Спіс літаратуры

1. Skopinava, A.M. Processing of quantitative expressions with units of measurement in scientific texts as applied to Belarusian and Russian text-to-speech synthesis / A.M. Skopinava, Yu.S. Hetsevich, 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.

2. 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 ноября 2012 г. – Минск : ОИПИ НАН Беларуси, 2012. – С. 260–265.

3. 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 февраля 2013 г. – Минск : БГУИР, 2013 г. – С. 319–328.

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

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