OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК [004.522+004.934+004.91]:004.89

КАМПАНЕНТЫ ІДЭНТЫФІКАЦЫІ КОЛЬКАСНЫХ ВЫРАЗАЎ З АДЗІНКАМІ ВЫМЯРЭННЯ Ў ТЭКСТАХ НА БЕЛАРУСКАЙ І РУСКАЙ МОВАХ

Гецэвіч Ю.С.[*], Скопінава А.М.[*]

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

yury.hetsevich@gmail.com

skelena777@gmail.com

Разгледжана актуальнасць і акрэслены складанасці праблемы ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння на прыкладзе беларускай і рускай моў на матэрыяле навукова-тэхнічных і прававых тэкставых карпусоў. Апісаныя і алгарытмічна прадстаўленыя адасобленыя семантычныя кампаненты для пошуку колькасных выразаў з улікам варыятыўных спосабаў запісу іх складнікаў: лічбавых дэскрыптараў і непасрэдна адзінак вымярэння. Для атрыманых практычных вынікаў азначаныя магчымыя вобласці прымянення.
Ключавыя словы: кампанент; ідэнтыфікацыя выразаў; колькасны выраз; адзінка вымярэння; канчатковы аўтамат, NooJ.

Уводзіны

Калі заходзіць гаворка пра распрацоўку інтэлектуальных сістэм, заўсёды ўзнікае пытанне пра спосабы распазнавання (расшыфроўкі) структураванай інфармацыі. Вядома, што тэксты часта ўтрымліваюць складаныя літарна-сімвальныя канструкцыі, якія былі ўжытыя стваральнікамі тэкстаў для апісання розных падзей ці сфер жыцця. Такімі канструкцыямі з'яўляюцца колькасныя выразы ў спалучэнні з мернымі сістэмамі або з сістэмамі адзінак вымярэння, якія з'яўляюцца важнымі для метралогіі, матэматыкі, інфарматыкі, фізікі, тэорыі кадавання, прамысловасці, эканомікі і гандлю і інш. Колькасныя апісанні ўласцівыя агульнай навуковай карціне свету, і, канешне, бытавой сферы жыцця. У якасці прыкладаў можна назваць наступныя спалучэнні колькасцяў з часта выкарыстальнымі адзінкамі вымярэння: *сіла току 59 мА, даўжыня 400 м, маса трыццаць пяць кілаграм, цеплыня 200 кДж, 225 ккал, 35 руб., пяць галоўнаў паліва* і г. д.

Увогуле тэксты, якія ўтрымліваюць колькасныя выразы з адзінкамі вымярэння (КВАВ), патрабуюць спецыяльных прыкладных алгарытмаў для рашэння актуальных задач у наступных сферах:

- *інтэлектуальныя пытальна-адказныя сістэмы* (для фармавання разнастайных запытаў кампанента інтэлектуальнага інтэрфейса; для хуткай класіфікацыі выразаў з адзінкамі вымярэння па спецыяльных класах у базах ведаў; для ўдакладнення параметраў вызначаных аб'ектаў у блоку лагічных аперацый);
- *сістэмы сінтэзу маўлення па тэксце* (для правільнай генерацыі арфаграфічнага тэксту па ўваходным тэксце; для правільнага ўтварэння недзялімых паслядоўнасцяў слоў і сінтагмаў);
- *сістэмы пошуку і апрацоўкі інфармацыі, каталогі і бібліятэкі* (для фармавання пашыраных пошукавых запытаў, якія змогуць знайсці канкрэтныя мерныя выразы у Інтэрнэце ці ў лакальнай базе тэкстаў; для аўтаматычнага рэферавання і анатавання);
- *выдавецкія установы* (для аўтаматызаванай лакалізацыі канкрэтнага спісу выразаў з адзінкамі вымярэння і для хуткага размеркавання знойдзеных выразаў па класах; для хуткай праверкі правільнасці ўжыванняў разгорнутых формаў назваў адзінак вымярэння ў тэкстах).

Пералічым асноўныя складанасці, якія ўзнікаюць пры распрацоўцы алгарытмаў ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў тэкстах:

1. *Выразы з колькасцю і адзінакамі вымярэння маюць шырокую варыятыўнасць як па напісанні, так і па ўтварэнні.* Менавіта з-за гэтага адразу запісаць правілы лакалізацыі выразаў для ўсіх выпадкаў практычна немагчыма. Для спрашчэння гэтага працэсу патрэбна выкарыстоўваць прыстасаванні, якія дазваляюць зручна карэктаваць ужо распрацаваныя правілы і дадаваць новыя.

2. *Выраз з колькасцю і адзінкай вымярэння складана ідэнтыфікаваць і выдзеліць ў ім колькасць (лік ці лічэбнік) і назву адзінкі (скарочаную ці поўную назвы) без добра падрыхтаваных лінгвістычных слоўнікаў.* Так адбываецца па той прычыне, што гэтыя слоўнікі павінныя ўтрымліваць апісанне усіх словаформаў, скарачэнняў і правілаў пабудовы вытворных формаў колькасцяў і назваў адзінак вымярэння. Напрыклад, гэта патрэбна для правільнай лакалізацыі і аналізу наступных выразаў з варыятыўнымі спосабамі запісу адзінак вымярэння даўжыні: *25 метраў, 21 метр, 100 м, тры метры*.

3. *Выразы з колькасцю і адзінкай з'яўляюцца мовазалежнымі*. Напрыклад, у англійскай мове скарачэннем слова *метр* з'яўляецца сімвал "*m"*, а ў беларускай – "*м"*. Назвы адзінак вымярэння адрозніваюцца па напісанні ў беларускай і рускай мовах (напрыклад, «гадзі́на», «час»). Таму для кожнай мовы патрэбна рабіць адмысловыя ўдакладненні пры распрацоўцы алгарытмаў ідэнтыфікацыі КВАВ.

Варта адзначыць, што некаторыя крокі для рашэння вышэй абазначаных праблем былі ажыццёўлены ў 2009 г. групай харвацкіх лінгвістаў, якія пабудавалі алгарытмы для вызначэння мерных выразаў даўжыні, плошчы і лікавых дыяпазонаў для ангельскай і харвацкай моў [Bekavac, 2009]. Мэтавая прадметная вобласць атрымала некаторае асвятленне ў працах шэрагу іншых еўрапейскіх лінгвістаў. Аднак іх даследаванні былі зробленыя больш тэарэтычна, чым практычна. Яны маюць апісальны характар і сканцэнтраваны не канкрэтна на адзінках вымярэння як на асобных паняццях, а як на «вызначаных выпадках ужыванняў слоў і выразаў, якія складаюць спецыяльную катэгорыю пайменаваных еднасцяў» [Cunningham, 1999] [пераклад тут і далей наш]. Таму яны апісаныя толькі паверхнева. Супрацоўнікі Балгарскай акадэміі навук і аддзялення Шэфілдскага ўніверсітэта адзначаюць, што іх «назіранні пра лінгвістычную сутнасць славянскіх пайменаваных еднасцяў заснаваны толькі на агульных характарыстыках і высновах згодна з асаблівасцямі іх ужывання ў тэкстах» [Paskaleva, 2002]. На практыцы аказваецца, што за тэрмінам «пайменаваная еднасць» хаваецца велізарны набор іншых складаных катэгорый: геаграфічных найменняў, асабістых прозвішчаў, імёнаў, мянушак людзей, назваў арганізацый, пазначэнняў дат, часу, грашовых і працэнтных адносін [Paskaleva, 2002; Mykowiecka, 2007]. Такім чынам, каб атрымаць эфектыўныя алгарытмы ідэнтыфікацыі, да кожнай з гэтых катэгорый варта падысці асобна. Балгара-сербскай камандай паняцце выразу з мернымі адзінкамі было разгледжана больш вузка і фармальна прадстаўлена ў выглядзе графа як «структура, якая складаецца са спалучэння ліку, запісанага словамі ці лічбамі, і індыкатара мернай адзінкі (*кіламетр*, *градус ухілу*, *міля*, *фут*, і да т.п.)» [Duško, 2007]. Тым не менш, атрыманы вынік прывязаны да вызначаных моўных сістэм (балгарскай і сербскай), у той час як астатнія на практыцы могуць у большай ці меншай ступені выкарыстоўвацца для харвацкай, ангельскай, польскай і ніводная – для беларускай ці рускай моў.

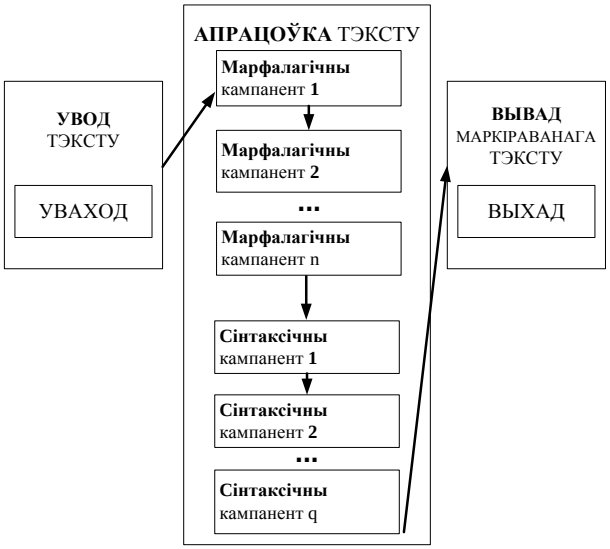
Таксама на сённяшні час ніводная пошукавая сістэма не прадстаўляе інтэрфейс для пабудовы пошукавага запыту з функцыяй лакалізацыі пэўных адзінак вымярэння.

Аўтарамі гэтага дакладу ўжо былі рэалізаваныя некаторыя крокі ў вызначаным напрамку. На XI міжнароднай канферэнцыі РІНТІ'2013 (г. Мінск, 15 лістапада 2012 г.) былі прадстаўлены алгарытмы, якія знаходзяць колькасныя выразы з адзінкамі вымярэння і класіфікуюць іх па трох тыпах (СІ, вытворныя ад СІ, не Сі) з дакладнасцю 72 % [Гецэвіч, 2012]. На практыцы кожны з гэтых алгарытмаў быў рэалізаваны ў выглядзе завершанага кампанента пры дапамозе камп'ютэрна-лінгвістычнага сродка NooJ [Silberztein, 2003].

У гэтым дакладзе ставіцца задача далей паляпшаць распрацаваныя алгарытмы. А менавіта на базе тэкстаў навукова-тэхнічнай і прававой тэматыкаў распрацаваць комплекс асобных сінтаксічных і марфалагічных кампанентаў, якія будуць здольныя ідэнтыфікаваць колькасныя выразы з *адзінкамі* вымярэння, якія перададзеныя пры дапамозе розных цэлых і скарочаных, дольных і кратных прэфікскаў (*фемтаграмы*, *кБайт, дм, гПа*), а таксама пры дапамозе цэлых і скарочаных асноў (*нанафарад, мЗв*).

1. Агульны агляд працэсу ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння

У кантэксце інтэлектуальных сістэм увесь шлях ад увода першапачатковага тэкставага фрагменту на ўваходзе да вывада апрацаванага тэксту з ідэнтыфікаванымі КВАВ можна прадставіць праз паслядоўнасць марфалагічных і сінтаксічных кампанентаў-апрацоўшчыкаў (малюнак 1).



Малюнак 1 – Універсальная схема працэсу ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння праз марфалагічныя і сінтаксічныя кампаненты

Спачатку ўведзены тэкст на любой мове апрацоўваецца з дапамогай слоўнікаў – кожны токен абазначаецца лексікаграфічнымі пазнакамі (напрыклад, род, лік, склон, клас канчаткаў), а далей незалежнымі марфалагічнымі (M1, M2,…, Mn) і сінтаксічнымі (S1, S2,.., Sq) кампанентамі. Кожны з кампанентаў з'яўляецца выканальным канчатковым аўтаматам, які выяўляецца графам з адным уваходам і адным выхадам, злучанымі шляхамі-канэктарамі з умовамі пераходу з аднаго вузла-стана ў іншы вузел-стан. Марфалагічныя кампаненты паслядоўна прымяняюцца да словаформаў-токенаў, каб даследаваць іх склад і абазначыць па жаданні распрацоўшчыка іх асаблівасці праз спецыяльныя адвольныя карыстальніцкія маркеры. Сінтаксічныя кампаненты таксама дазваляюць ужываць маркеры, але ўжо для *спалучэнняў* слоў, лічбаў, знакаў пунктуацыі і інш. сімвалаў.

З пункту погляду рэалізацыі і выкарыстання марфалагічныя і сінтаксічныя кампаненты з'яўляюцца спецыяльнымі файламі, якія могуць быць выкліканыя незалежна адзін ад аднаго староннімі праграмамі праз унутраныя сродкі NooJ.

Тэарэтычна рашэнне вышэй пастаўленай задачы таксама можна рэалізаваць з дапамогай рэгулярных выразаў. Але, у адрозненне ад іх, візуальныя графы (ці канчатковыя аўтаматы) дазваляюць хутчэй і зручней мадэфікаваць і правяраць алгарытмы кампанентаў падчас распрацоўкі. Гэта сапраўды важна, бо КВАВ4 уласцівы варыятыўны характар. Фармальна пералічыць і апісаць усе выпадкі іх ужыванняў адразу практычна не магчыма.

Акрамя гэтага, NooJ дазваляе ствараць тэставыя корпусы тэкстаў, неабходныя для адладкі працы алгарытмаў. Аўтарамі былі створаны 4 тэставыя корпусы для 2 розных тэматычных даменаў: навукова-тэхнічнага (астраномія, фізіка, геаграфія, хімія, авіяцыя, касманаўтыка, гісторыя, энергетыка, транспарт і сувязь) і прававога (збор правіл дарожнага руху для Беларусі) – па 2 корпусы для кожнай з моў [Гецэвіч, 2012]. Далей усе корпусы былі папоўненыя тэкстамі з КВАВ розных тыпаў.

Наступны раздзел апісвае назначэнне і асаблівасці рэалізацыі кожнага кампанента комплексу ідэнтыфікацыі КВАВ.

2. Комплекс кампанентаў для ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння па словаўтваральных прыкметах

Згодна са словаўтваральнымі асаблівасцямі, у корпусе сустракаюцца наступныя адзінкі вымярэння:

- з цэлай асновай і без прэфікса (*метр, Герц, Ом*);
- з цэлай асновай і з цэлым прэфіксам (*нанафарады, міліампер*);
- з цэлай асновай і са скарочаным прэфіксам (*кБайт*).
- са скарочанай асновай і без прэфікса (*Дж, га, Па);*
- са скарочанай асновай і са скарочаным прэфіксам *(км, дл, гПа)*;

Гэтая ўмоўная класіфікацыя была выкарыстана як галоўны прынцып пры распрацоўцы кампанентаў ідэнтыфікацыі КВАВ.

Спачатку былі створаныя лінгвістычныя слоўнікі S для беларускай (БМ) і рускай (РМ) моў (малюнак 2). Гэтыя слоўнікі ўтрымліваюць базавыя асновы адзінак вымярэння – назоўнікі і іх абрэвіятуры. Яны адпавядаюць класіфікацыі як поўныя і скарочаныя асновы адзінак вымярэння, прычым кожная аснова пазначана адпаведным атрыбутам «Base» або «Mbase». Да поўных асноў дадаецца пазнака флексійнага класа. Відавочна, што слоўнік S з'яўляецца мовазалежным лінгвістычным рэсурсам, у адрозненне ад кампанентаў-алгарытмаў ідэнтыфікацыі КВАВ, якія пабудаваныя як мованезалежныя модулі.

Далей былі пабудаваныя мовазалежныя лінгвістычныя рэсурсы (Fsubmultiple, Fmultiple, Ssubmultiple, Smultiple) з усімі дазволенымі прэфіксамі для ўтварэння адзінак вымярэння. Крыніцай для гэтага былі дадзеныя Міжнароднага бюро мер і вагаў [BIPM – SI brochure (8th edition), 2006]. Для ўтварэння адзінак вымярэння могуць выкарыстоўвацца кратныя (multiple) ці дольныя (submultiple) прэфіксы, прычым ў скарочанай (прэфікс S) або поўнай (прэфікс F) формах (адпаведна малюнак 3, малюнак 4).

Згодна з вышэй апісанай класіфікацыяй па словаўтваральным прынцыпе, былі распрацаваныя 4 марфалагічныя мованезалежныя самастойныя кампаненты, якія выкарыстоўваюць слоўнік S і лінгвістычныя рэсурсы з прэфіксамі Fsubmultiple, Fmultiple, Ssubmultiple, Smultiple. Прызначэнне

першага з іх – ідэнтыфікацыя базавых адзінак вымярэння ў поўнай або скарочанай слоўнікавай форме без прэфіксаў (малюнак 5). Будзем гаварыць, што граф ці падграф «спрацаваў» тады, калі быў знойдзены любы шлях ад яго ўваходу да выхаду пры ўмове выканання ўсіх праверак паміж яго ўваходам і выхадам.

Алгарытм працы марфалагічнага кампанента М1 наступны:

1. Для кожнага слова *T* выходнага тэксту *TT* зрабіць дзеянні 2-3, калі словы скончацца, то крок 4.
2. Калі *T* супадае з любой асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі А і дадаць маркер Mub, далей крок 1.
3. Калі *T* супадае з любой асновай *A* з пазнакай Mbase са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі *A* і дадаць маркер Mbase, далей крок 1.
4. Вывад – прамаркаваны *TT*. Канец алгарытму.

Трэба адзначыць, што на выхадзе марфалагічнага кампанента М1 знойдзеная адзінка вымярэння атрымае адпаведны маркер, з дапамогай якога пазней можна будзе яе ідэнтыфікаваць у тэксце. Напрыклад, слова *Гц* атрымае маркер *Mub*, паводле гэтага маркера па запыце <ABBREVIATION+Mub> можна будзе яго пабачыць у выніковым канкардансе (малюнак 6).

Далей пачалася праца над марфалагічным кампанентам М2. Ён ідэнтыфікуе адзінкі вымярэння, якія ўтвораныя з дапамогай кратных і/ці дольных цэлых прэфіксаў (малюнак 7). У выніку яго працы адзінкі вымярэння могуць быць абазначаныя адным з трох маркераў:

- Mump – значыць, што ідэнтыфікавана адзінка вымярэння з кратным прэфіксам;
- Musp – значыць, што ідэнтыфікавана адзінка вымярэння з дольным прэфіксам;
- Muhp – значыць, што ідэнтыфікавана адзінка вымярэння з некалькімі прэфіксамі (напрыклад, *мікрамегафарад*). Такім спосабам утвараць адзінкі вымярэння не прадугледжана ў СІ, таму дадзеныя словы трэба абазначыць у тэксце, каб пазней вывесці спіс памылкова ўтвораных адзінак вымярэння.

Алгарытм працы марфалагічнага кампанента М2 наступны:

1. Для кожнага слова *T* выходнага тэксту *TT* зрабіць дзеянні 2-5, калі словы скончацца, то крок 6.
2. Калі пачатковая частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі А і дадаць маркер Mump, далей крок 1.
3. Калі пачатковая частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fsubmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі А і дадаць маркер Musp, далей крок 1.
4. Калі пачатковая частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fmultiple, наступная частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fsubmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі А і дадаць маркер Muhp, далей крок 1.
5. Калі пачатковая частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fsubmultiple, наступная частка *T* супадае з любой колькасцю паўтораў любога прэфікса *P* з Fmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі А і дадаць маркер Muhp, далей крок 1.
6. Вывад – прамаркаваны *TT*. Канец алгарытму.

Прыклады працы марфалагічнага кампанента М2 прадстаўлены на малюнках 8 і 9.

Заўважым, што распрацаваныя марфалагічныя кампанеты з-за прастаўленых флексійных класаў у заходным слоўніку асноў ідэнтыфікуюць асновы адзінак вымярэння з любой варыятыўнасцю канчаткаў, напрыклад, *километров* ці *километр*. Таму здымаецца неабходнасць першапачатковай нармалізацыі слова да пачатковай формы, каб яго правільна ідэнтыфікаваць. Таксама на выхадзе марфалагічнага кампанента адзінка вымярэння атрымлівае ўсе ўласцівасці, якія былі замацаваныя за яе базавай асновай ў слоўніку S. Дзякуючы гэтаму, словы, напрыклад, *дэкалітрамі*, *наносекундами*, застаюцца паўнавартасным назоўнікамі з усімі прыналежнымі ім канчаткамі і характарыстыкамі, нягледзячы на адсутнасць гэтых слоў-асноў у яўным выглядзе ў слоўніку S (малюнак 8).

Наступным этапам стала распрацоўка марфалагічных кампанентаў М3 і М4 для ідэнтыфікацыі адзінак вымярэння, якія адпаведна ўтвораныя пры дапамозе цэлай асновы і скарочанага прэфіксу (малюнак 10) альбо – скарочанай асновы і скарочанага прэфіксу (малюнак 11).

Алгарытм працы марфалагічнага кампанента М3 наступны:

1. Для кожнага слова *T* зыходнага тэксту *TT* зрабіць дзеянні 2-3, калі словы скончацца, то крок 4.
2. Калі пачатковая частка *T* супадае з любым прэфіксам *P* з Smultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі A і дадаць маркер Mump, далей крок 1.
3. Калі пачатковая частка *T* супадае з любым прэфіксам *P* з Ssubmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Base са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі A і дадаць маркер Mump, далей крок 1.
4. Вывад – прамаркаваны *TT*. Канец алгарытму.

Алгарытм працы марфалагічнага кампанента M4 наступны:

1. Для кожнага слова *T* зыходнага тэксту *TT* зрабіць дзеянні 2-3, калі словы скончацца, то крок 4.
2. Калі пачатковая частка *T* супадае з любым прэфіксам *P* з Smultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Mbase са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі A і дадаць маркер Mump, далей крок 1.
3. Калі пачатковая частка *T* супадае з любым прэфіксам *P* з Ssubmultiple, а астатняя рэшта *T* супадае з асновай *A* з пазнакай Mbase са слоўніка *S*, то перанесці на *T* у зыходны тэкст *TT* усе граматычныя характарыстыкі A і дадаць маркер Mump, далей крок 1.
4. Вывад – прамаркаваны *TT*. Канец алгарытму.

На малюнках 12 і 13 прыведзены адпаведныя прыклады выніковых канкардансаў, якія атрымліваюцца пасля адпаведных пошукавых запытаў да прамаркаваныга *TT* кампанентамі M3 і M4:

- <NOUN+Mump> накіраваны на пошук назоўнікаў з цэлымі асновамі са скарочанымі кратнымі прыстаўкамі;
- <ABBREVIATION+Musp> накіраваны на пошук абрэвіятурных адзінак вымярэння са скарочанымі асновамі і скарочанымі дольнымі прэфіксамі.

Пасля рэалізацыі слоўніка SS і кампанентаў M1-M4 застаецца вызначыць колькасны дэскрыптар, які, часцей за ўсё, стаіць перад адзінкай вымярэння, і разам з ёй утварае KBAB. Колькасны дэскрыптар можа выражацца матэматычна (лічбамі, знакамі ці лічбавымі выразамі) альбо лінгвістычна (пры дапамозе лічэбнікаў, колькасных займеннікаў, прыслоўяў і іх спалучэнняў), напрыклад: *450 нанафарад, 15•10^(-25) Тэсла, тры молі, шмат градусаў, некалькі секунд* і да т.п.

На дадзеным этапе ажыццёўлены алгарытм ідэнтыфікацыі колькасных дэскрыптараў, якія перададзеныя сродкамі матэматыкі. Для гэтага быў распрацаваны самастойны сінтаксічны кампанент S1 (малюнак 14), які спрацоўвае не толькі на простыя, дзесятковыя і дробавыя лічбы ў розных варыяцыях пісьмовага запісу, але і на лічбавыя выразы з экспаненцыяльнымі часткамі (малюнак 15). Зазначым, што дадзены кампанент з'яўляецца моваенезалежным.

Для збору усіх маркераў, якія былі расстаўленыя ў тэксце *TT* праз слоўнік S і марфалагічныя кампаненты M1-M4, быў распрацаваны сінтаксічны кампанент S2 (малюнак 16). Ён спрацоўвае толькі на выразы з адзінкамі вымярэння, перад якімі стаіць колькасна-лічбавы дэскрыптар, які ідэнтыфікуецца праз убудаваны асобным падграфам сінтаксічны кампанент S1. У выніку кожнаму KBAB прысвойваецца маркер <MUEXPR>, паводле яго будуюцца выніковыя мэтавыя канкардансы колькасных выразаў з адзінкамі вымярэння (малюнак 17).

Такім чынам, каб у любым уваходным тэксце ідэнтыфікаваліся колькасныя выразы з адзінкамі вымярэння, якія ўтвораныя з дапамогай кратных ці дольных прыставак з цэлымі ці скарочанымі асновамі, патрэбна напоўніць універсальную схему з малюнка 1 канкрэтнымі кампанентамі. У выніку будзе пабудаваны шматкампанентны комплекс ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння (малюнак 18), які складаецца са слоўніка S, марфалагічных M1-M4 і сінтаксічных S1-S2 кампанентаў.

Заключэнне

Такім чынам, была пастаўлена і вырашана задача распрацоўкі кампанентаў для ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння з улікам варыятыўных спосабаў запісу іх складнікаў. Акрэслены магчымыя сферы ўжывання рашэнняў гэтай задачы: сістэмы сінтэзу маўлення па тэксце, сістэмы пошуку інфармацыі, бібліятэкі, выдавецкія установы, натуральна-маўленчыя інтэрфейсы і інтэлектуальныя сістэмы.

Распрацаваныя алгарытмы ідэнтыфікацыі з'яўляюцца самастойнымі незалежнымі кампанентамі, якія рэалізаваныя ў форме канчатковых аўтаматаў марфалагічных і сінтаксічных граматык у праграме NooJ. Кожны з гэтых кампанентаў паасобку ці ўсе кампаненты разам могуць быць убудаваныя ў іншыя сістэмы для дапамогі рашэння адпаведных задач.

Варта зазначыць, што атрыманыя кампаненты граматык NooJ наглядна перадаюць зыходныя алгарытмы і дазваляюць іх хутка мадэфікаваць, маштабаваць і папаўняць лінгвістычнымі рэсурсамі, што важна для павелічэння дакладнасці і паўнаты ідэнтыфікацыі выразаў з адзінкамі вымярэння ў тэкстах.

У будучым плануецца паляпшаць кампаненты праз вырашэнне наступных задач:

- ідэнтыфікаваць не толькі матэматычныя, але і лінгвістычна запісаныя колькасныя дэскрыптары ў спалучэннях з назвамі адзінак вымярэння;
- падключыць да дадзенага комплексу прадстаўленыя на канферэнцыі РІНТІ'2013 кампаненты для ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння ў міжнароднай сістэме СІ, вытворнымі ад СІ і не з СІ;
- ідэнтыфікаваць знак адмоўнасці/дадатнасці перад лікавым дэскрыптарам;
- зменшыць колькасць памылак пры ідэнтыфікацыі шматзначных выразаў, напрыклад, у тых выпадках, калі алгарытм «блытае» адзінкі вымярэння між сабой (*г* для *год*, *грам*, *гадзіна*) ці адзінкі вымярэння з маркамі транспартных сродкаў (*МАЗ-4А 5*, а не *4 амперы*);
- папоўніць слоўнік базавых асноў адзінак вымярэння менш ужывальнымі асновамі, а да графаў дзесятковых прэфіксаў дадаць двойкавыя прэфіксы Міжнароднай электратэхнічнай камісіі.

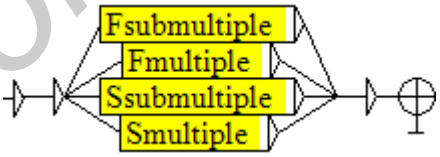
```
Б,ABBREVIATION+Mbase
байт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
бекерэль,NOUN+FLX=АВАЛЬ+s6+UNAMB+Base
біт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
В,ABBREVIATION+Mbase
Вт,ABBREVIATION+Mbase
ват,NOUN+FLX=БАЙТ+s2+UNAMB+Base
вольт,NOUN+FLX=БАЙТ+s2+UNAMB+Base
г,ABBREVIATION+Mbase
га,ABBREVIATION+Mbase
гг,ABBREVIATION+Mbase
гектар,NOUN+FLX=ГЕКТАР+s5+UNAMB+Base
герц,NOUN+FLX=АМПЕР+s2+UNAMB+Base
год,NOUN+FLX=ГОД+sN+UNAMB+Base
град,ABBREVIATION+Mbase
грам,NOUN+FLX=ГРАМ+s3+UNAMB+Base
Гц,ABBREVIATION+Mbase
гц,ABBREVIATION+Mbase
```

а)

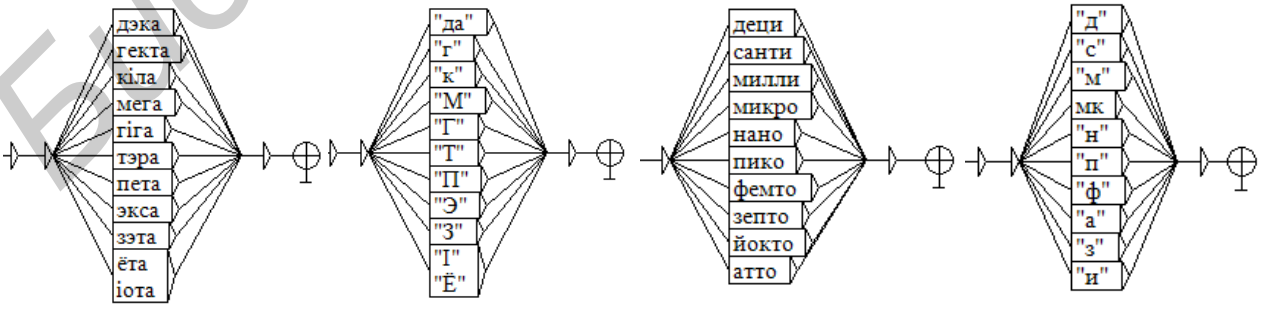
```
ампер,NOUN+FLX=АЛТЫН+s4+UNAMB+Base
А,ABBREVIATION+Mbase
байт,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
бит,NOUN+FLX=АБАЖУР+s2+UNAMB+Base
Б,ABBREVIATION+Mbase
ватт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
Вт,ABBREVIATION+Mbase
вольт,NOUN+FLX=АЛТЫН+s2+UNAMB+Base
В,ABBREVIATION+Mbase
гектар,NOUN+FLX=АБАЖУР+s5+UNAMB+Base
га,ABBREVIATION+Mbase
герц,NOUN+FLX=ГЕРЦ+s2+UNAMB+Base
Гц,ABBREVIATION+Mbase
год,NOUN+FLX=ГОД+sN+UNAMB+Base
г,ABBREVIATION+Mbase
гг,ABBREVIATION+Mbase
град,ABBREVIATION+Mbase
грамм,NOUN+FLX=АНГСТРЕМ+s3+UNAMB+Base
```

б)

Малюнак 2 – Слоўнікі-рэсурсы базавых асноў адзінак вымярэння для беларускай (а) і рускай (б) моў

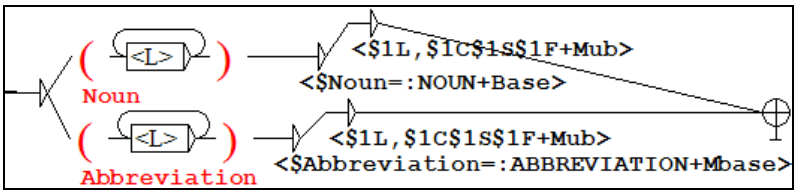


Малюнак 3 – Класіфікацыя прэфіксаў у выглядзе канчатковага аўтамата



а) б) в) г)

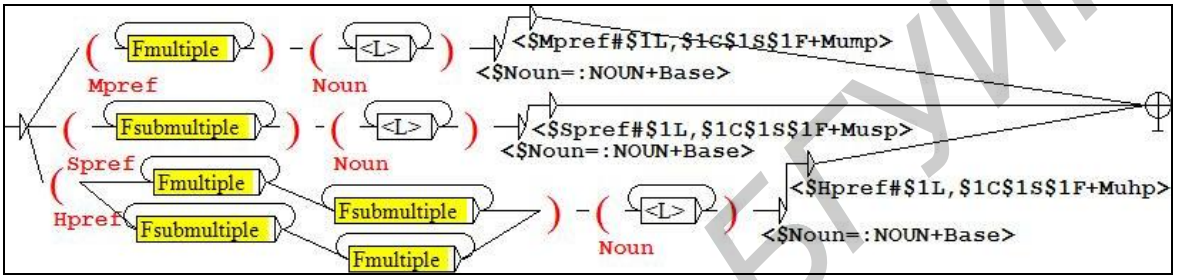
Малюнак 4 – Графы для ідэнтыфікацыі кратных прыставак у поўнай (а) і скарочанай (б) формах для беларускай мовы і дольных прыставак у поўнай (в) і скарочанай (г) формах для рускай мовы



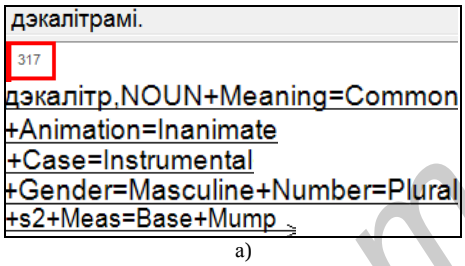
Малюнак 5 – Марфалагічны кампанент М1 для ідэнтыфікацыі базавых адзінак вымярэння з цэлай і скарочанай асновай без прэфіксаў

Before	Seq.	After
гніта на халадзільніку 1-2.4	Тл	- у зазоры магніта тыповага
зблізку 2,4 мЗв у год. 1	Н	ёсць сіла зямнога прыцягнення
00 км. Кулóн (пазначэнне:	Кл	, С) — адзінка вымярэння электрычнага
правадніка пры сіле току 1	А	за час 1 с. ... Комплекс
току 1 А за час 1	с	. ... Комплекс ASTER, які складаецца
е атрымліваць кожныя 30	хв	(у штатным рэжыме) здымкі
, 2-200 МЭВ, 2-30 кэв, 0,1	Гц	-300 кгц, 0-50 кгц; перыядычнасць абна
пэратуры раўняецца 0,025	эВ	. Энэргія электрона ў прамянёвай

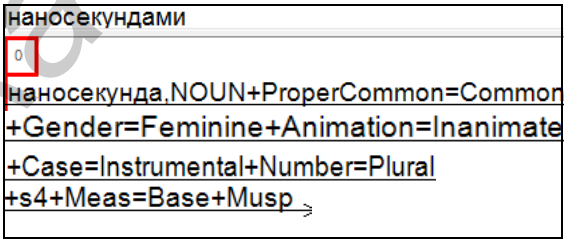
Малюнак 6 – Выніковы канкарданс базавых адзінак вымярэння па запыце <ABBREVIATION+Mub> для беларускай мовы



Малюнак 7 – Марфалагічны кампанент М2 для ідэнтыфікацыі адзінак вымярэння з цэлымі асновамі і цэлымі кратнымі і/ці дольнымі прэфіксамі



а)

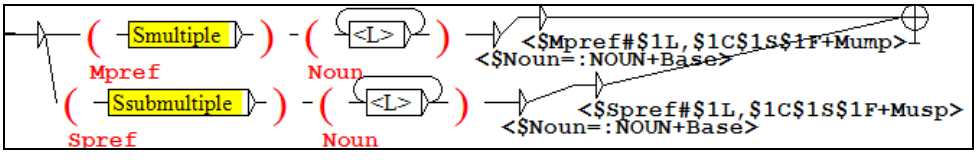


б)

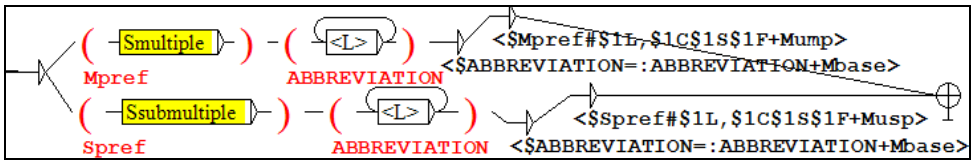
Малюнак 8 – Прыклад анатаванай словаформы для беларускай (а) і рускай (б) моў

Before	Seq.	After
на расстоянии несколько сотен	километров	. Первый вариант ударного
кусков может достигать нескольких	килограммов	. Куски брони поражают
излучения мощностью в сотни	мегаватт	. Проблема в том
составляет уже десятки тысяч	мегагерц	, что соответствует волнам
своих жестких дисках тысячи	гигабайт	информации, третьи подключаются
энергии лазерного излучения порядка	мегаджоуля	(106 Дж) и кпд

Малюнак 9 – Выніковы канкарданс адзінак вымярэння ў поўнай форме па запыце <NOUN+Mump> на прыкладзе рускай мовы



Малюнак 10 – Марфалагічны кампанент М3 для ідэнтыфікацыі адзінак вымярэння з цэлай асновай і скарочаным прэфіксам



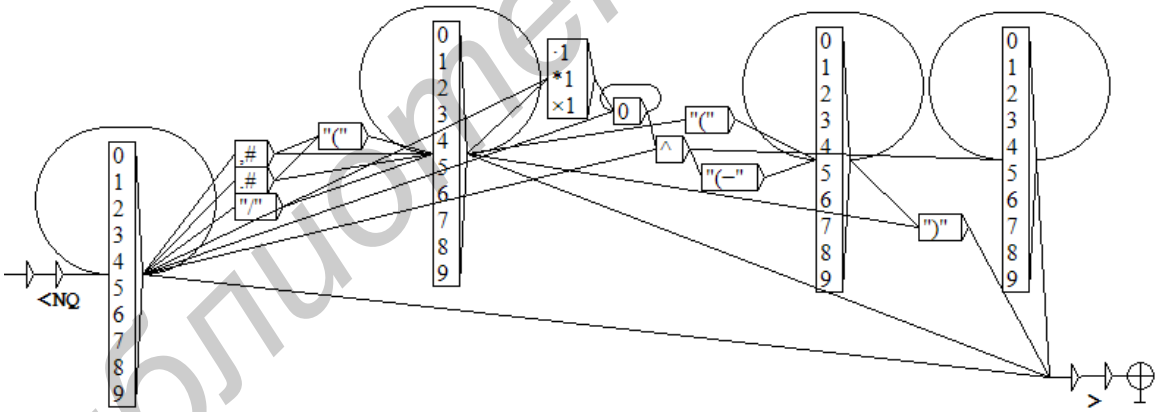
Малюнак 11 – Марфалагічны кампанент M4 для ідэнтыфікацыі адзінак вымярэння са скарочанай асновай і са скарочаным прэфіксам

Before	Seq.	After
подтвержденная величина кванта Go ~ 13	кОм	^(-1), позволили авторам интерпретиро
уникационных соединений, например, 100	Мбит	/с в стандарте 100BASE
именно 100 000 000 бит/с, а 10	Гбит	/с в стандарте 10GBASE
на которой указан объём 1,44	Мбайт	, на самом деле вмещает
самом деле вмещает лишь 1440	Кбайт	, то есть 1,38 Мбайт в
лишь 1440 Кбайт, то есть 1,38	Мбайт	в обычном понимании. Что
с ГМД порциями по 5	Кбайт	. По данным спутниковых наблюдений
формат, скорость передачи до 64	Мбит	/с) и 137,4 МГц (метровый

Малюнак 12 – Выніковы канкарданс адзінак вымярэння са скарочанымі прэфіксамі і цэлымі асновамі па запыце <NOUN+Mump> на прыкладзе рускай мовы

Before	Seq.	After
дакладнасцю да 0,2 % роўная масе 1	дм	³ хімічна чыстай воды пры
звычайна вар'іруецца зблізку 2,4	м3в	у год. 1 Н ёсць
адзінкі вымярэння: мікрон, роўны 1	мкм	, і ангстрэм (А*), роўны
адзінку масы - грам (0,001 кг). 31	мкТл	(3,1×10^(-5) Тл) - напружана
пф, а не 60 нф; 2 000	мкф	, а не 2 мф). Прыстаўкі
пры 0° шыраты (на экватары) 5	мТл	- сіла звычайнага магніта на
нф; 2 000 мкф, а не 2	мф	). Прыстаўкі, якія адпавядаю
Ч); спектральнае разрашэнне - 10-20	нм	(ІЧ); радыяметрычнае разра
(пішуць 60 000 пф, а не 60	нф	; 2 000 мкф, а не 2 мф
р ёмістасці = (1/9)·10^(-11) Ф = 1,11…	пф	. Дзесятковыя кратныя і дзе

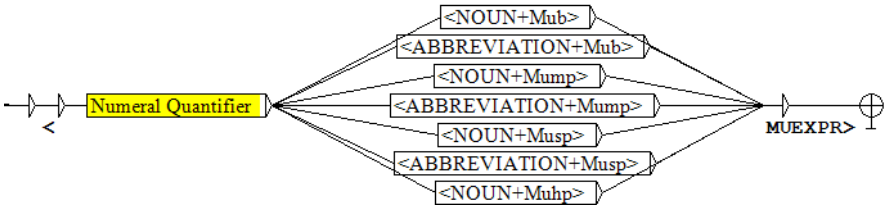
Малюнак 13 – Выніковы канкарданс адзінак вымярэння са скарочанымі асновамі і скарочанымі прэфіксамі па запыце <ABBREVIATION+Musp> на прыкладзе беларускай мовы



Малюнак 14 – Сінтаксічны моўнезалежны кампанент S1 для ідэнтыфікацыі колькасна-лічбавых дэскрыптараў

Before	Seq.	After
дзеяння сілы. Вага цела масай	102	г (т. е. сіла гравітацыі
для аўтамабіля, тралейбуса, прычэпа; · -	13,5	метра для аўтобуса з дзвюма
нергія фатона чырвонага бачнага святла:	2,61·10^(-19)	Дж.
магнітная індукцыя ў сярэднім складае	5·10^(-5)	Тл, а на экватары (шырата
пры атамным бамбаванні Хірасімы: каля	6·10^13	Дж. Энергія фатона чырвонага

Малюнак 15 – Прыклады вынікаў пошуку колькасна-лічбавых дэскрыптараў для беларускай мовы, якія атрыманыя пры дапамозе сінтаксічнага кампанента S1



Малюнак 16 – Галоўны сінтаксічны кампанент S2, які вызначае выразы з колькасна-лічбавымі дэскрыптарамі і токенамі з пазнакамі марфалагічных камнанетнаў M1-M4

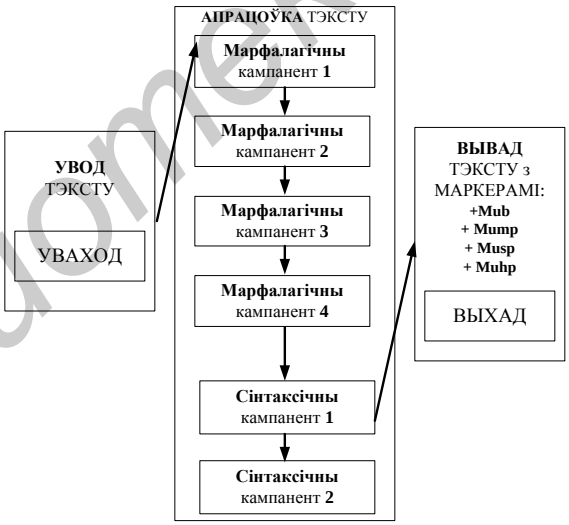
Before	Seq.	After
адзінку масы - грам (0,001 кг).	31 мкТл	(3,1×10^(-5) Тл) - напружанасць магні
звычайна вар'іруецца зблізку	2,4 мЗв	у год. 1 Н ёсць
на апору з сілай	9.81 Н	. Прыбліжэнне, што 1 кг адпавядае
дамі або нанафарадамі (пішуць	60 000 пф	, а не 60 нф; 2 000 мкф
ёмістасць шара з радыусам	1 сантыметр	, змешчанага ў вакуум. 1 сантыметр
Mbit(альбо проста Mb).	1 мегабіт	= 1000^2 біт = 10^6 біт = 1000000 біт.
святло ў вакууме за (	1 / 299 792 458) секунды	. Метр быў упершыню ўведзены
ны дыяпазон - 40 кэв-3 МЭВ, 2-	200 МЭВ	, 2-30 кэв, 0,1 Гц-300 кгц, 0-50 кгц
ай трубцы тэлевізара - парадку	20 кілаэлектронвольт	. Энэргіі касьмічных прамянёў - ад
энргіі касьмічных прамянёў - ад	1 мегаэлектронвольта	да1000 тэралектронвольтаў.

а)

Before	Seq.	After
организм ток не превышал	1 мА	. На человека токи статического
могут сказать «файл в	100 килобайт	»). При обозначении скоростей тел
ротивление величиной от 1 до	100 МОм	, чтобы протекающий через челове
икромегафарад пикотеравольт	13 йоттайоктограммов	Каждая строка содержит информа
до 64 Мбит/с) и	137,4 МГц	(метровый диапазон, формат АРТ
евонширский изумруд» массой	1383,95 каратов	. Изумруды выращивают искусстве
время жизни мюонов - около	2.2 мкс	- осложняет задачу создания мюон
сса которой оказалась равной	22 фемтограммам	(1 фг = 1•10^(-15) г). . Мюоны, как
то они оказались равными:	8.1·10^21 Дж	(уменьшение массы ледников на
: - высота 670 км - наклонение	98,00 град	. Срок активного существования 1 г

б)

Малюнак 17 – Фрагменты вынікаў ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння пасля апрацоўкі беларуска- (а) і руска- (б) -моўных тэкстаў праз комплекс марфалагічных і сінтаксічных кампанентаў



Малюнак 18 – Выніковая схема шматкампанентнага комплексу ідэнтыфікацыі колькасных выразаў з адзінкамі вымярэння

Бібліяграфічны спіс

[Bekavac, 2009] Bekavac, B. Units of Measurement Detection Module for NooJ/ B. Bekavac// Conference on NooJ 2009. – Tunisia. – 2009. – P.121-127.

[Cunningham, 1999] Cunningham, H. Information Extraction: a User Guide (revised version)/ H. Cunningham// Research Memorandum CS-99-07. – Department of Computer Science, University of Sheffield. – 1999.

[Paskaleva, 2002] Paskaleva E., Angelova, G., Jankova, M., Bontcheva, K., Cunningham, H., Wilks, Y. Slavonic Named Entities in Gate/ E. Paskaleva [et al.]// Research Memorandum CS-02-01. – Department of Computer Science, University of Sheffield. – 2002.

[Mykowiecka, 2007] Mykowiecka, A., Kupść, A., Marciniak, M., Piskorski, J. Resources for Information Ex-traction from Polish texts/A. Mykowiecka// Proceedings of 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. – Poznan. – 2007.

[Silberztein, 2003] Silberztein, M. NooJ Manual [Electronic resource]. – 2003. – Mode of access : http://www.nooj4nlp.net/NooJManual.pdf. – Date of access : 01.07.2012.

[Duško, 2007] Duško, V., Krstev, C., Koeva, S. Towards a Complex Model for Morpho-Syntactic Annotation/V. Duško[et al.]// Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages. – Borovets, Bulgaria. – 2007. – P. 65-71.

[Гецэвіч, 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) : даклады XI Міжнароднай канферэнцыі (Мінск, 15 лістапада 2012 г.). – Мінск : АІПІ НАН Беларусі, 2012. – С. 260–265.

[BIPM – SI brochure (8th edition), 2006] BIPM - SI brochure (8th edition) [Electronic resource]. – 2006. – Mode of access : http://www.bipm.org/en/si/si_brochure/. – Date of access : 10.11.2012.

COMPONENTS FOR IDENTIFICATION OF QUANTITIVE EXPRESSIONS WITH MEASUREMENT UNITS IN BELARUSIAN AND RUSSIAN TEXTS

Hetsevich Y.S. *, Skopinava A.M. *
**United Institute of Informatics Problems,*
National Academy of Sciences,
Minsk, Republic of Belarus

{yury.hetsevich, skelena777}@gmail.com

A study of an identifying process of quantitive expressions with measurements units (QEMU) in terms of word formation in thematically distinct texts for Belarusian and Russian is reported here. Models, isolated semantical components and resources for recognition of quantitive expressions with measurements units are described and algorithmically presented. Scopes of application for the obtained practical results are specified.

INTRODUCTION

The urgency of the problem is dictated by the ubiquity of quantitive expressions with measurements units and their enormous variety. Texts which contain QEMU require specific algorithms of identification and processing in such areas as corpora and database management systems, libraries, information retrieval systems, text-to-speech synthesizers, publishing institutions, intellectual systems, natural-language interfaces. Nowadays no search engine provides an interface for building up a search query with QEMU localization function. Observations of other European research workers are too general, their practical results are limited only to the definite language systems. Our research work consists in development of components and linguistic resources in order to identify and classify QEMU with them on the material of hand-crafted text corpora for the Belarusian and Russian languages.

MAIN PART

Dealing with QEMU implies many difficulties, conditioned by a great variety of numeral quantifiers, names of units, their language-dependent origin. It is extremely important to use tools that allow adjusting easily already developed rules and adding new ones. The international computer-linguistic program NooJ is one of such tools. It allows implementing sophisticated algorithms of searching for compound text fragments in Belarusian and Russian in the form of visual executable graphs, which later form the necessary components. The basic principle for the components is the word-formative classification of QEMU:

1) QEMU with full-form stems and without prefixes *(метр, Герц, Ом), (eng. meter, Hertz, Ohm)*;
2) QEMU with full-form stems and full-form prefixes (*нанафарады, міліампер*), *(eng. nanofarads, milliampere)*;
3) QEMU with full-form stems and shortened prefixes *(кБайт), (eng. Kbyte);*
4) QEMU with shortened stems and without prefixes (*Дж, га, Па*), *(eng. J, ha, Pa)*;
5) QEMU with shortened stems and shortened prefixes *(км, дл, гПа), (eng. km, dL, hPa).*

Each type is algorithmically described by one of the four morphological components and linguistic resource (figure 2): for QEMU types 1 and 2 see figure 5, for type 3 – figure 7, for type 4 – figure 11, for type 5 – figure 10. Some results are presented in the form of concordances in figures 6, 9, 12, 13. Besides, two syntactic components are developed: the subgraph for numeral quantifiers (figures 14, 15); and the main linking syntactical component that coordinates all the others (figures 17, 18). Thus, by means of the components any text fragment can be morphologically and syntactically analyzed. It is to note, the developed linguistic resources (electronic dictionaries with basic QEMU and subgraphs with SI-prefixes; figures 2-4) are language-dependent, while the morphological and syntactical components are universal.

CONCLUSION

The obtained components are created in the form of finite-state automata through a set of morphological and syntactic grammars within the powerful linguistic processor NooJ. The finite-state automaton demonstrates how the algorithms work and indicate how they can be further updated in order to improve the accuracy. Further improvements are being planned:

- developing components which will identify numeral quantifiers expressed not only by numbers (mathematics objects), but also by numerals (parts of speech);
- providing simultaneous identification of QEMU according to the word formation peculiarities and System International classification;
- disambiguating multiple-valued expressions, for example, in such cases when algorithms "confuse" some units with each other (the same initial letter ‘г’ for ‘год’ year, ‘грам’ gram, ‘гадзіна’ hour) (г for the year, grams per hour);
- identifying front plus and minus signs, disambiguating minus, hyphen and dash signs;
- updating the base of QEMU with rare ones.