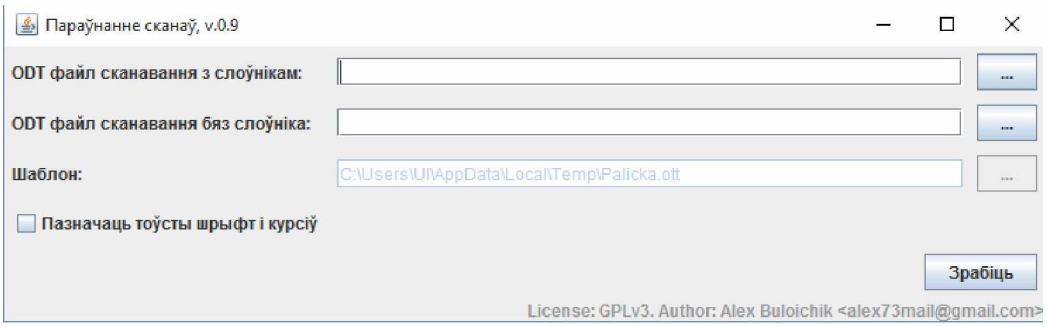
Уладзімір Кошчанка, Антон Смалюк

ПРАГРАМНАЕ ЗАБЕСПЯЧЭННЕ І ПРАЦЭС ПАДРЫХТОЎКІ ТЭКСТАЎ ДЛЯ НАЦЫЯНАЛЬНАГА КОРПУСУ БЕЛАРУСКАЙ МОВЫ

Падчас працы над Нацыянальным корпусам беларускай (НКБ) мовы быў створаны шэраг праграм для працы з лексіка-граматычнай базай, якая ляжыць у аснове граматычнай разметкі корпусу, і ўласна Нацыянальным корпусам беларускай мовы: рэдактар метатэкставай разметкі (пашпарту) корпусу, рэдактар корпусу (дазваляе захоўваць тэкст у фармаце корпусу, здымаць рознаўроўневую аманімію ў паўаўтаматычным рэжыме, папаўняць лексіка-граматычную базу і г.д.), праграма праверкі лексіка-граматычнай базы на паслядоўнасць, праграма параўнання сканаваных тэкстаў, праграма канвертацыі тэкстаў у папярэдні фармат корпусу.

Папаўненне корусу. Адсканаваны тэкст распазнаецца ў праграме для аптычнага распазнавання тэкстаў *ABBYY FineReader*. Для аптымізацыі працы на аснове лексіка-граматычнай базы быў створаны слоўнік для FineReader і шэраг праграмных сродкаў, якія дазволілі аўтаматызаваць працэс.

Тэкст праходзіць распазнаванне два разы: першы раз распазнаецца аптычна без слоўніка, а другі раз – са слоўнікам. Атрыманыя два файлы параўноўваюцца пры дапамозе ўласнарапрацаванай праграмы “*Параўнанне сканаў*”:



Праграма візуальна параўноўвае файл, адсканаваны без слоўніка, і файл, адсканаваны са слоўнікам, сінхранізуе гэтыя файлы, прымяняе шаблон і вынікі захоўвае ў асобны файл з пашырэннем .odt.

Выніковы тэкст у файле .odt мае наступны выгляд:

I

Восень заўсёды падкрадваецца неўзаметку. Янічэ {Яшчэ }прыгравае сонца, усё вакол зялёнае, яшчэ цвітуць гуркі, на доўгай гарбузовай касе смела ўсміхаецца вялізная пяцікутная жоўта-залатая кветка. Лета не хоча здавацца. І {І }раптам прачнешся раніцаю {ранідаю }— на вуліцы туман, густы, белы, як малако. А калі ён развеецца, на бярозе ярчэй засвецяцца жоўтыя лісцікі, бульбоўнік пачарнее, бы абвараны, і раса на траве халодная, аж коле ў босыя пяткі.

Звычайна на трэцім перапынку хлапчукі выбягалі на ўзгор'е і скакалі з абрыву ў пясчаную ямку. На дне яе жвірысты пясок быў халаднаваты, хоць зверху трохі прагравaўся, ды неўзабаве парэпаныя ад цыпак ногі перамешвалі яго, ператаўкалі, як проса ў ступе.

Скакалі хлапчукі не проста так сабе. Тут ішло зацятае спаборніцтва — хто далей скокне. Не раз быў чэмпіёнам Андрэйка Сахута. Але нечакана для ўсіх уперад вырваўся Парасчын Данілка. Андрэйка гэтага не мог перажыць. Калі б Паўлік, ягоны даўні сябрук, было б не так крыўдна, а то смаркаты Данька раптам сігануў далей за ўсіх. Мабыць, і Паўліка гэта раздражніла, ён шмаргануў носам і са здзіўленнем у голасе сказаў:

— Смелы ёты {еты }байструк. У матку ўдаўся. Яна ваўка забіла...

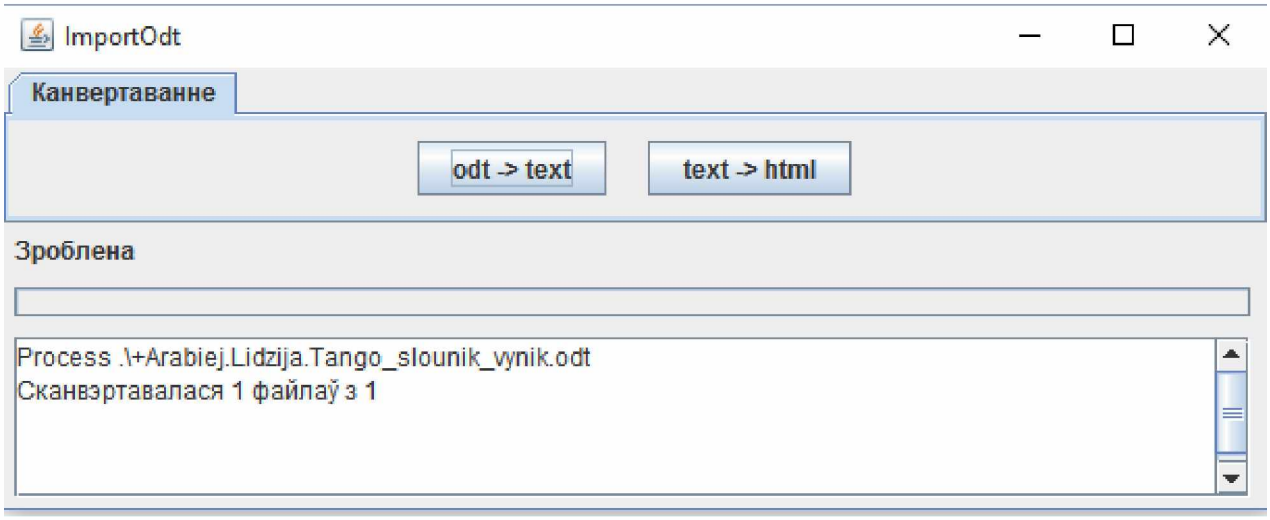
— Хто байструк? А то ў нос як дам зараз! — замахнуўся брудным кулачком Данька.

— Ну, паспрабуй! — натапырыўся Паўлік.— Як дам выспятка, дык і пакоцішся прама ў Бесядзь.

Колерам у тэксце вылучаны спрэчныя месцы, на якія трэба асабліва звяртаць увагу. У аснову кладзецца тэкст, расчытаны пры дапамозе слоўніка, але распазнаванне са слоўнікам мае адзін мінус: праграма падбірае найбліжэйшае слова, якое, паводле алгарытму, найбольш пасуе ў гэтым месцы і здараюцца выпадкі, калі слова падабрана, але яно памылковае, таму побач для праверкі ў фігурных дужках пакідаецца слова, распазнанае без слоўніка. У нашым выпадку відаць, што формы “*Янічэ*” (*Яшчэ*) і “*ёты*” (*еты*) былі падабраны няўдала, на што ўказвае варыянт у фігурных дужках, а слова “раніцаю” – удала. Такі падыход дазволіў істотна сэканоміць час пры вычытцы тэкстаў.

Адначасова з вычыткай тэксту правяралася і пазначалася яго структура, закладзеная ў шаблоне: іерархія загалоўкаў, эпіграфы, подпісы, паэтычныя ўстаўкі ў празаічных тэкстах.

Такім чынам, вычытаны і структураваны тэкст далей апрацоўваецца пры дапамозе праграмы ImportODT (уласная распрацоўка):

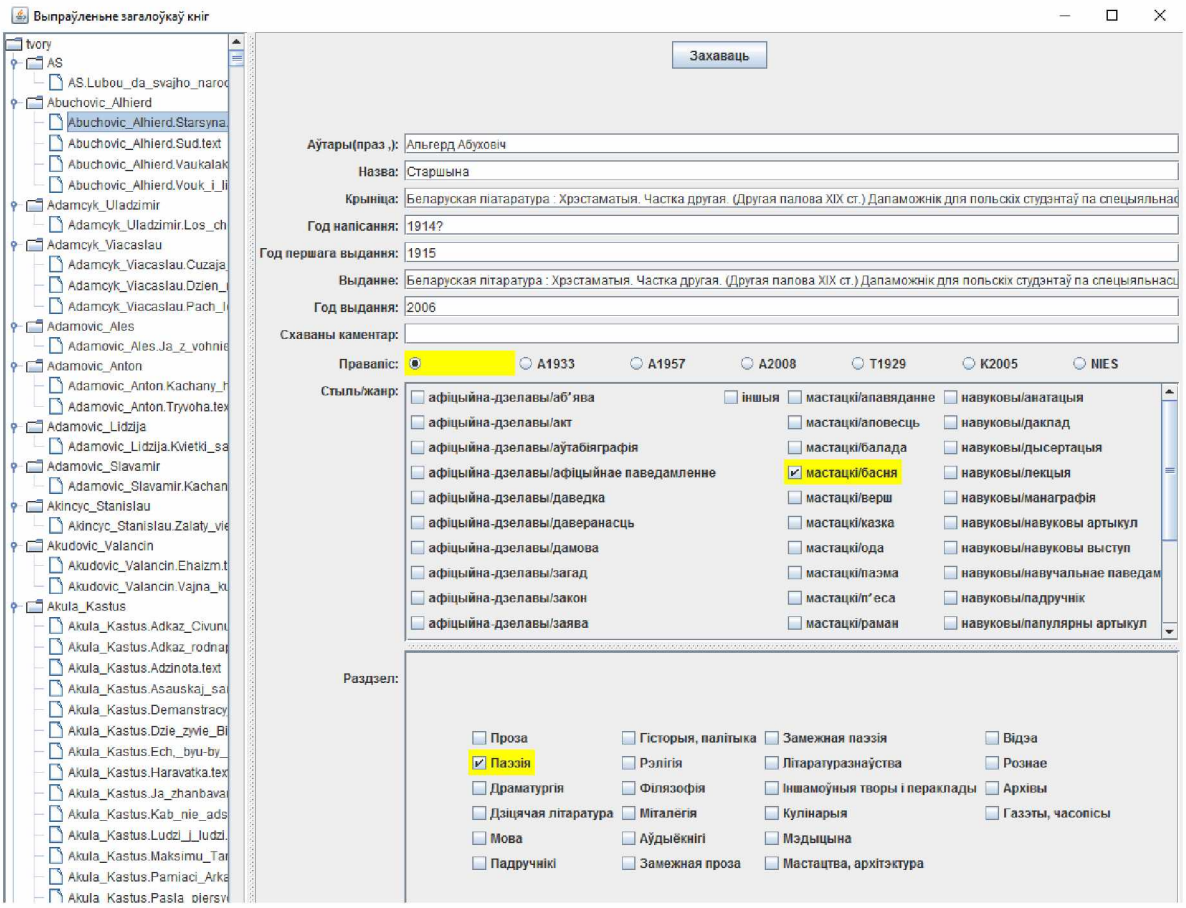


Праграма правярае правільнасць разметкі структуры тэксту і канвертуе файл *odt* у фармат *txt* і потым у фармат *html* з структурнай разметкай:

```
<!--·BOOK_BEGIN·-->¶
<p> </p>¶
<div·class="nootnote-block">¶
</div>¶
<h3><a·id="chapter1"></a>«ТАНЦУЙ·ТАНГО...»</h3>¶
<p>Гарадок·быў·нібыта·незнаёмы·чалавек,·пра·якога·думаеш·—·вельмі·ён·на·
некага·падобны,·толькі·не·ўспомніш·—·на·каго,·а·потым·пачынаеш·разумець,·што·
гэта·проста·такое·аблічча,·блізкае·табе·аблічча,·можа,·нават·земляка,·
чалавека·з·тваёй·вёскі.</p>¶
<p>Так·і·гэты·гарадок.·Колькі·іх,·вось·гэтакіх·—·з·зялёным·скверыкам·у·
цэнтры,·з·Домам·культуры,·у·архітэктуры·якога·можна·пазнаць·былую·царкву·
```

Калі ў структурнай разметцы ёсць недакладнасці ці супярэчнасці, то праграма ўказвае на месца і тып памылкі. Такім чынам, тэкст падрыхтаваны для ўключэння ў корпус.

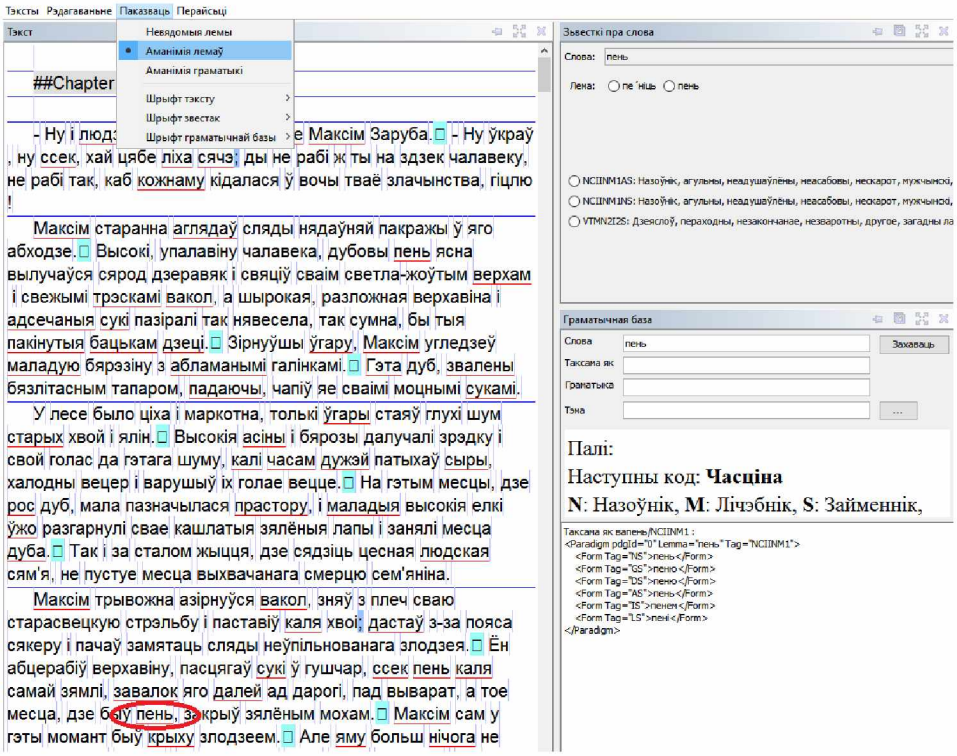
Наступны этап – пашпартызацыя тэксту. Для гэтых мэтаў была распрацавана асобная праграма (гл. малюнак ніжэй), якая дазваляе ў паўаўтаматычным рэжыме ствараць апісанне тэксту (метатэкставую інфармацыю):



Пры дапамозе праграмы можна пазначыць аўтара (-аў) тэксту, яго назву, крыніцу атрыманага тэксту, год яго стварэння, першага выдання, актуальнага выдання, зрабіць дадатковыя заўвагі, пазначыць правапіс, стыль і жанр. Адкрытая архітэктура праграмы дазваляе пры патрэбе дадаваць іншыя параметры тэксту. У выніку ў файл дадаецца інфармацыя наступнага кшталту:

##Authors: Альгерд Абуховіч
##Title: Старшына
##SectionsTheme: Паэзія
##Source: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці "Беларуская філалогія" / Уклад. і камент. М. Хаўстовіча. Мн., 2006. - 204 с.
##Description: упершыню апублікаваны ў Беларускі календар «Нашай нівы» на 1915 г. – Вільня, 1915
##Edition: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці “Беларуская філалогія” / Уклад. і камент. М. Хаўстовіча. Мн., 2006. – 204 с
##StyleGenre: мастацкі/басня
##CreationYear: 1914?
##PublicationYear: 2006
##FirstPublicationYear: 1915
##HiddenSource:

Пасля дадавання метатэкставай інфармацыі, тэкст апрацоўваецца праграмай рэдагавання корпусу:



Праграма працуе на камп'ютары карыстальніка, выкарыстоўвае толькі чытанне з лексіка-граматычнай базы. З праграмай могуць працаваць розная людзі адначасова (кожны на сваім камп'ютары), але кожны будзе апрацоўваць свой тэкст. Працэс нагадвае рэдагаванне звычайнага тэксту ў тэкставым рэдактары:

- адчыняецца тэкставы файл;
- праграма робіць аўтаматычнае сегментаванне і такенізацыю;
- спецыяліст рэдагуе тэкст, змяняе сегментаванне, такенізацыю, здымае аманімію і г.д.;
- тэкст захоўваецца ў фармаце корпусу і перадаецца ў сховішча тэкстаў корпусу.

Праграма дазваляе працаваць у некалькіх рэжымах: 1) *Невядомыя лемы* – выдзяляюцца словы, якія не трапілі ў ЛГБ. Праграма дазваляе разгарнуць парадыгму гэтых слоў, прысвоіць адпаведную граматыку і дадаць слова ў ЛГБ; 2) *Аманімія лемаў* – выдзяляюцца словы з часцінамоўнай аманіміяй (гл. прыклад на малюнку вышэй). Гэты рэжым дазваляе здымаць часцінамоўную аманімію ў паўаўтаматычным рэжыме; 3) *Аманімія граматыкі* – выдзяляюцца словы з граматычнай аманіміяй (аманімія ўнутры адной парадыгмы), зняцце граматычнай аманіміі таксама ажыццяўляецца ў паўаўтаматычным рэжыме.

Дадаткова праграма паказвае межы абзацаў, сказаў і слоў, што дазваляе кантраляваць правільнасць пазначэння структурных элементаў пры папярэдняй аўтаматычнай разметцы.

Пры захаванні, праграма канвертуе адпрацаваны тэкст ў фармат корпусу:

```
<XMLText>
    <Header>
        <Tag name="Authors">Якуб Колас</Tag>
        <Tag name="CreationYear">1913</Tag>
        <Tag name="FirstPublicationYear">1914</Tag>
        <Tag name="SectionAuthor">Апавяданьні</Tag>
        <Tag name="SectionsTheme">Проза</Tag>
        <Tag name="StyleGenre">мастацкі/апавяданне</Tag>
        <Tag name="Title">Малады дубок</Tag>
    </Header>
    <Content>
        <p/>
        <Tag name="Chapter">I</Tag>
        <p/>
        <p>
            <se>
                <z cat="KM1">-</z>
                <s char=" "/>
                <w cat="E_Y" lemma="ну">Ну</w>
                <s char=" "/>
                <w cat="CKX_E" lemma="і">і</w>
                <s char=" "/>
                <w cat="NCAPNM1NP" lemma="чалаве´к" manual="true">людзі</w>
                <z cat="KE">!</z>
            </se>
            <se>
                <s char=" "/>
                <z cat="KM1">-</z>
                <s char=" "/>
                <w cat="VDMN2PXSM" lemma="гавары´ць">гаварыў</w>
```

На сённяшні дзень, створаны электронны корпус тэкстаў аб’ёмам 80 000 000 словаўжыванняў са структурнай і граматычнай разметкай і паш-партызацыяй і распрацаваны эксперыментальны рухавік корпусу.

Папярэдняя версія корпусу аб’ёмам каля 50 000 000 словаўжыванняў размешчана з мэтай тэсціравання ў інтэрнэце па адрасе http://bnkorpus.info/. Корпус дазваляе шукаць дакладную форму слова, слова з усімі слова-формамі, паводле граматычных характарыстык. Пошук можна ажыццяўляць з улікам аўтара, году напісання, стылю, жанру і г.д. Створана магчымасць выбаркі па кластарах (спалучэннях суседніх слоў).