Ю.С. Гецэвіч, С.І. Лысы (Мінск, АІПІ НАН Беларусі)

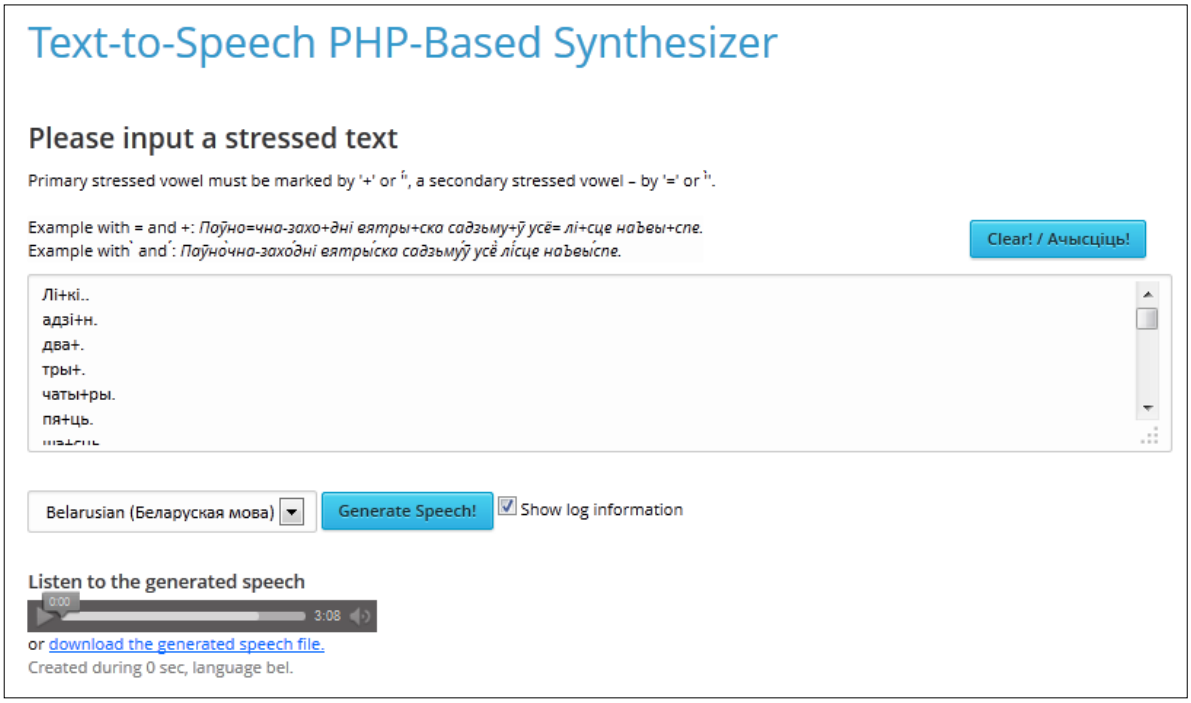
РАШЭННЕ ПРЫКЛАДНЫХ ЛІНГВІСТЫЧНЫХ ЗАДАЧ ПРЫ ДАПАМОЗЕ СЭРВІСАЎ РЭСУРСУ WWW.CORPUS.BY

Распрацоўка сістэмы сінтэзу маўлення па тэксце (ССМТ) з'яўляецца складанай задачай, якая ахоплівае вялікую колькасць неабходных падзадач, а таксама сутыкаецца з шэрагам пабочных прыкладных лінгвістычных задач, якія патрабуюць рашэння. Задачы, якія паўстаюць пры стварэнні ССМТ, можна ўмоўна падзяліць на тры групы:

- перадапрацоўка ўваходных тэкстаў;
- непасрэдна сінтэз маўлення;
- апрацоўка выніковых дадзеных.

Інтэрнэт-рэсурс www.corpus.by [1] мае сваёй мэтай рашэнне як непасрэднай задачы сінтэзавання маўлення, так і шэрагу пабочных задач перадапрацоўкі тэксту і апрацоўкі вынікаў працы сінтэзатара маўлення. Дзеля дасягнення гэтай мэты было прынята рашэнне распрацаваць сістэму невялікіх інтэрнэт-сэрвісаў, кожны з якіх вырашаў бы пэўную ўласную задачу, а ў спалучэнні з іншымі сэрвісамі — неабходныя падзадачы ССМТ. Але перш чым перайсці да разгляду інтэрнэт-сэрвісаў, створаных у падтрымку да сінтэзатара маўлення, звернемся да кароткага агляду самога сінтэзатара маўлення па тэксце (СМТ) [2], які ён ёсць на сённяшні дзень.

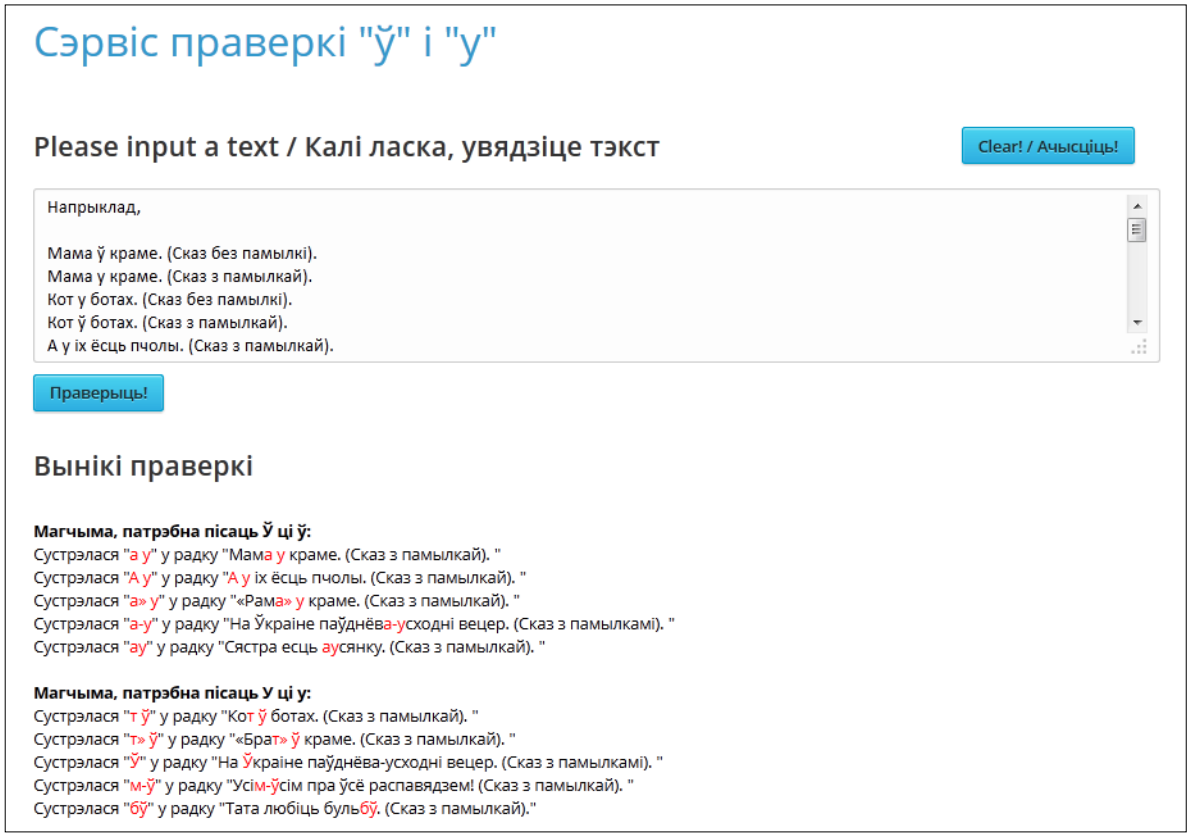
Сінтэзатар маўлення па тэксце з'яўляецца прыладай перапрацоўкі тэксту, пададзенага ў звычайным (графічным) выглядзе, у тэкст у гукавым выглядзе. Знешні інтэрфейс распрацаванага СМТ паказаны на малюнку 1. Каб атрымаць тэкст у гукавым выглядзе, карыстальнік павінен увесці гэты тэкст у адпаведнае вакно, абраць мову і націснуць кнопку "Generate Speech!». Для лепшага, больш зразумелага гучання тэксту, карыстальнік мусіць пазначыць у кожным слове асноўны націск знакам "+" (плю+с) пасля націскной галоснай, альбо знакам — "◌́» (акýт) над націскной галоснай. Адпаведна дадатковыя націскі — знакам "=" альбо "◌̀» (гравіс). Гэтая неабходнасць выклікана тым, што, на сённяшні дзень, да сінтэзатара не далучаны слоўнік. Націснуўшы кнопку "Generate Speech!", карыстальнік атрымлівае гукавы файлік у фармаце wav, які можа праслухаць не сыходзячы са старонкі, спампаваць альбо падзяліцца спасылкай, па якой гукавы файлік будзе ўвесь час даступны. Таксама карыстальнік мае магчымасць атрымаць прамежкавую інфармацыю аб выніках працы тэкставага, фанемнага, алафоннага і інш. блокаў СМТ, папярэдне пазначыўшы птушачкай "Show log information».



Мал. 1. Знешні інтэрфейс сінтэзатара маўлення па тэксце ў інтэрнэце.

Цяпер пяройдзем да апісання тых інтэрнэт-сэрвісаў, якія былі распрацаваны ў прылажэнне да СМТ.

Адным з важных момантаў у сінтэзаванні маўлення з'яўляецца тое, што ўваходны тэкст мусіць не мець арфаграфічных памылак. Іначай, відавочна, вынікам працы СМТ будзе некарэктны гукавы тэкст, які рэжа слых. ***"Сэрвіс праверкі "у" і "ў"*** [3] распрацаваны для хуткай праверкі вялікіх электронных тэкстаў на беларускай мове з мэтай пошуку і выпраўлення адной з найбольш распаўсюджаных памылак — памылкі ў напісанні літар "у" і "ў». На ўваход сэрвісу падаецца тэкст на беларускай мове ў звычайным выглядзе. Калі карыстальнік націсне кнопку "Праверыць!», сэрвіс абазначыць магчымыя памылкі двух тыпаў: па-першае, калі напісана літара "у" ў пазіцыі, дзе верагодна павінна быць літара "ў», па-другое, у адваротным выпадку, калі напісана літара "ў", дзе верагодна павінна быць літара "у».



Мал. 2. Знешні інтэрфейс і прыклад працы “Сэрвісу праверкі “у” і “ў”.

У працэсе пошуку магчымых памылак сэрвіс ня толькі вызначае, галосная ці зычная літара знаходзіцца перад “у», але таксама аналізуе сімвалы, якія не з’яўляюцца літарамі, у выпадку, калі літара “у” знаходзіцца ў пачатку слова. Гэтыя сімвалы могуць мець непасрэдны ўплыў на напісанне слова. На сённяшні дзень сэрвіс правярае напісанне без уліку выключэнняў. Дадаць разгляд выключэнняў у сэрвісе плануецца.

Яшчэ адзін з сэрвісаў перадапрацоўкі тэксту — ***“Сэрвіс генеравання назваў сімвалаў”*** [4]. Напрыклад, карыстальнік надрукаваў слова “ма+ма», а СМТ пасля апрацоўкі зачытвае нешта незразумелае накшталт “мма». У чым праблема? Карыстальнік можа звярнуцца да сэрвісу і даведацца, з якіх сімвалаў складаецца яго запыт да сінтэзатара. У выніку апрацоўкі сэрвісам пазначанага вышэй слова, карыстальнік атрымае інфармацыю аб тым, што другі сімвал мае назву “Лацінская малая літара эй». Літара не кірылічная, і СМТ не можа яе апрацаваць.

Сэрвіс створаны дзеля таго, каб даць магчымасць карыстальніку атрымаць назвы сімвалаў кадыроўкі Windows-1251. На ўваход падаецца любáя паслядоўнасць сімвалаў (тэкст, асобныя словы, выпадковы набор сімвалаў) дадзенай кадыроўкі. Карыстальнік можа задаць адну з трох моваў (беларускую, рускую ці англійскую), а таксама абраць форму выводу інфармацыі (назвы сімвалаў у слупок альбо праз коску). Мэтай дадзенага сэрвісу з’яўляецца вырашэнне праблемы агучвання тэксту, у якім сустракаюцца незнаёмыя сінтэзатару сімвалы.

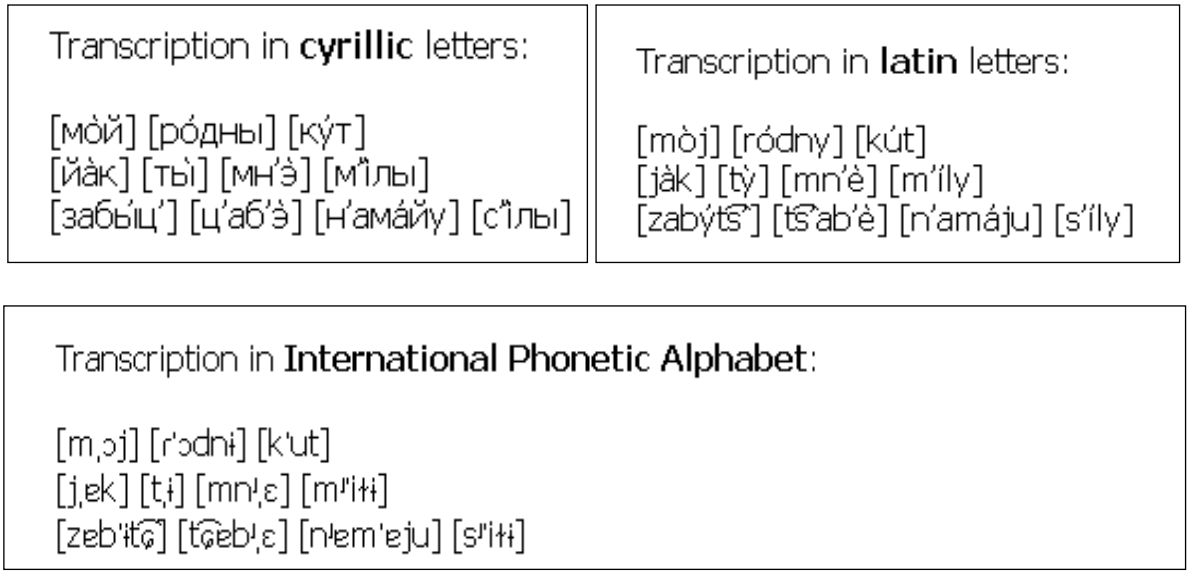
Адным з сэрвісаў апрацоўкі выніковых дадзеных СМТ з’яўляецца ***“Сэрвіс канвертавання алафоннага тэксту ў розныя транскрыпцыі”*** [5].

Сэрвіс прызначаны для генеравання транскрыпцый па ўваходным тэксце. На ўваход падаецца тэкст у алафонным выглядзе. Ніжэй прыведзены першыя два радкі знакамітага беларускага верша ў алафонным запісе:

M004,O113,J'013,/,R032,O022,D001,N004,Y322,/,K001,U032,T000,/,#C3,
J'002,A142,K004,/,T002,Y121,/,M001,N'004,E143,/,M'002,I041,L004,Y310,/,#E2
,

(Алафонны запіс слова можна атрымаць пры дапамозе апісанага вышэй сінтэзатара маўлення, які знаходзіцца па адрасе www.corpus.by/tts3, пры ўключанай опцыі Show log information).

Увёўшы тэкст, карыстальнік можа абраць адзін альбо некалькі тыпаў транскрыпцыі з трох магчымых.



Мал. 3. Прыклад працы “Сэрвісу канвертавання алафоннага тэксту ў розныя транскрыпцыі”.

Першы тып транскрыпцыі — кірылічны, ён можа быць карысны школьнікам, настаўнікам, студэнтам і выкладчыкам ВНУ. Наступныя два — лацінскія: спрошчаная лацінская транскрыпцыя і транскрыпцыя ў Міжнародным фанетычным алфавіце **(**International Phonetic Alphabet — IPA**)** [6]. Канвертаванне тэксту ў спрошчаную лацінскую транскрыпцыю распрацавана паводле працы У.А. Кошчанкі [7]. Транскрыпцыя ў Міжнародным фанетычным алфавіце арыентаваная на інтэрналізацыю беларускай мовы. Яна можа быць карыснай як для навуковых працаў у сувязі з беларускай мовай (прыкладам, стварэнне арфаэпічнага слоўніка), так і для звычайных карыстальнікаў (асабліва замежнікаў), якія, дзякуючы гэтаму сэрвісу, атрымліваюць магчымасць убачыць гукавы склад беларускіх слоў паводле міжнароднага стандарту.

У выніку пастаўленай задачы была створаная сістэма інтэрнэт-сэрвісаў у прылажэнне да CMT, у тым ліку “Сэрвіс праверкі “у” і “ў””, “Сэрвіс генеравання назваў сімвалаў (кадыроўка Windows-1251)», “Сэрвіс канвертавання алафоннага тэксту ў розныя транскрыпцыі” і інш. Гэтыя

сэрвісы знаходзяцца ў вольным доступе рэсурсу www.corpus.by. У будучыні плануецца праца над удасканаленнем ужо існуючых інтэрнэт-сэрвісаў, інтэграцыя распрацаваных інтэрнэт-сэрвісаў у больш складаныя сістэмы, а таксама стварэнне новых інтэрнэт-сэрвісаў.

ЛІТАРАТУРА

1. Corpus.by [Electronic resource]. – 2013. – Mode of access: http://corpus.by/. – Date of access: 06.03.2014.

2. Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – 2013. – Mode of access: http://corpus.by/tts3/. – Date of access: 06.03.2014.

3. Сэрвіс праверкі “ў” і “у” [Electronic resource]. – 2013. – Mode of access: http://corpus.by/ u_check/. – Date of access: 06.03.2014.

4. Service of generating names of characters (encoding Windows-1251) / Сэрвіс генеравання назваў сімвалаў (кадыроўка Windows-1251) [Electronic resource]. – 2014. – Mode of access: http://corpus.by/getNamesOfCharacters /. – Date of access: 06.03.2014.

5. Service of converting allophonic texts into different transcriptions / Сэрвіс канвертавання алафоннага тэксту ў розныя транскрыпцыі [Electronic resource]. – 2013. – Mode of access: http://corpus.by/convertAllophToDifPhonemes/. – Date of access: 06.03.2014.

6. Wiktionary:International Phonetic Alphabet. [Electronic resource]. – 2008. – Mode of access: http://en.wiktionary.org/wiki/Wiktionary:International_Phonetic_Alphabet. – Date of access: 07.03.2014.

7. Беларуска-англійскі размоўнік / уклад. У. А. Кошчанка. – Мінск.: Артыя Груп, 2010.