

РАСПРАЦОЎКА ЛІНГВІСТЫЧНЫХ РЭСУРСАЎ ДЛЯ ТЭСТАВАННЯ ФАНЕМНА-АЛАФОННАГА ПРАЦЭСАРА СІНТЭЗАТАРА БЕЛАРУСКАГА МАЎЛЕННЯ

На сённяшні дзень задача сінтэза маўлення з’яўляецца не новай, аднак не губляе сваёй актуальнасці з-за імкнення палепшыць працу існуючых сістэм генеравання чалавечага маўлення машынай.

Сінтэз вуснага маўлення па тэксце ажыццяўляецца на аснове лексіка-граматычнага аналізу ўваходнага тэксту. Арфаграфічны тэкст дакумента паступае на ўваход сінтэзатара і далей падвяргаецца паслядоўнай апрацоўцы спецыялізаваных працэсараў у адпаведнасці з агульнай структурай сінтэзатара маўлення па тэксце. Сінтэзатар уключае чатыры асноўныя модулі: тэкставы працэсар, прасадны працэсар, фанетычны працэсар і акустычны працэсар [1, с. 111].

Фанетычны працэсар пераўтварае арфаграфічны тэкст у фанемны і генеруе пазіцыйныя і камбінаторныя алафоны. Нагадаем, што алафон – гэта рэалізацыя фанемы, яе варыянт, абумоўлены пэўным фанетычным асяроддзем, які з’яўляецца канкрэтным маўленчым гукам. Нягледзячы на шырокі дыяпазон алафонаў адной фанемы, носьбіт мовы заўсёды ў стане іх распазнаць. Менавіта таму так важна распрацаваць дакладную базу ведаў для тэставання фанемна-алафоннага працэсара сінтэзатара беларускага маўлення. Гэта дазволіць згенераваць усе пазіцыі алафонаў для далейшага тэставання і выпраўлення магчымых памылак у працы сінтэзатара.

Раней ужо былі распрацаваны адзіныя табліцы правілаў пераўтварэння “Фанема-Алафон” для рускай мовы [1, с. 253-258], у якіх алафон пазначаецца імем фанемы, магчымай мяккасцю і наступнымі за ім трыма лікавымі індэксамі, якія абазначаюць: для галосных – націскны, часткова націскны ці ненаціскны гук, а таксама левае і правае асяроддзе; для зычных – адзіночны ці падвоены алафон, левае і правае асяроддзе. Улічваючы гэта, а таксама пэўныя асаблівасці беларускай мовы, былі распрацаваныя адпаведныя табліцы правілаў пераўтварэння “Фанема-Алафон” для беларускай мовы. Гэтыя табліцы для беларускай мовы налічваюць 761 алафон, але падлічаная колькасць алафонаў паводле табліц не з’яўляецца дакладнай. Гэта тлумачыцца тым, што ў табліцах прысутнічаюць тэарэтычна спрагназаваныя алафоны, г.зн., што алафоны не былі правераныя на практычнае ўжыванне ў рэальных словах, дзе яны сустракаюцца хаця б адзін раз. Таму на аснове гэтых табліц была распачата праца над стварэннем лінгвістычнага рэсурса для тэставання фанемна-алафоннага працэсара сінтэзатара беларускага маўлення і выяўлення недахопаў і недакладнасцей у вышэй пазначаных табліцах. Дадзены рэсурс уяўляе сабой спіс алафонаў для беларускай мовы колькасцю 761 адзінка, на кожны з якіх неабходна знайсці прыклад слова ці адзначыць, што такога прыкладу не знайшлося.

Базай для пакрыцця дадзеных алафонаў з’яўляліся спецыяльна створаныя тэматыкі слоў («прафесіі», «ежа і напоі», «геаграфічныя назвы» і

Service of Searching and Sorting Allophones

Please input a text with allophones

Калі ласка, увядзіце тэкст з алафонамі

Clear! / Ачысціць!

Please input stop-allophones

Калі ласка, увядзіце стоп-алафоны

D004,A322,K003,U233,M'002,E042,N003,T002,Y323,/,J'012,I141,
/,M004,A312,T002,E322,R002,Y323,J'011,A041,L004,Y312,/,A323,DZ'004,I142,N000,/,#P4,
K004,N'002,I042,GH004,A332,/,D004,A323,S'002,L'002,E042,D004,A321,V012,A313,N'104,E342,
/,A321,P102,R012,Y322,R002,O023,DZ'004,E343,/,J'012,I141,
/,P002,R012,Y322,CH001,Y022,N004,A322,GH004,/,B004,A312,GH002,A033,C'102,A342,
/,N004,A322,R002,O022,D004,A321,W013,/,B004,Y311,L002,A012,
/,A321,P002,U311,B001,L'004,I342,K004,A331,V011,A012,N004,A321,
/,W012,A112,N001,GH004,L'004,I143,I141,/,W013,T002,Y123,S'002,A342,CH002,A320,/,#C3,

U131
I332
Y120
U110
E140
A000
A001
A002

Get allophones! / Знайсці алафоны!

Get allophones only (Толькі алафоны)

The total number of elements is 20.
Different elements are 6:

Allophones sorted by alphabet

A220, occurs 8 time(s) in

U302,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
V002,O313,S'002,E341,M003,S002,O122,T002,/,P'002,A343,DZ'104,E343,S'002,A142,T002,/,S'002,E141,M003,/,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
U302,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
J'002,I142,R001,V'002,I142,N003,/,F'002,I142,SH002,E122,R003,/,T002,Y123,S'002,A142,CH002,**A220**,/,#C3,
V002,O313,S'002,E341,M003,S002,O122,T002,/,SH002,E322,S002,DZ'104,E343,S'002,A142,T002,/,S'002,E141,M003,/,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
V002,O313,S'002,E341,M003,S002,O122,D001,/,V002,O113,S'002,E341,M001,DZ'004,E343,S'002,A342,T002,/,T002,R022,Y122,
/,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
V002,O313,S'002,E341,M003,S002,O122,D001,/,V002,O113,S'002,E341,M001,DZ'004,E343,S'002,A342,T002,/,T002,R022,Y122,
/,T002,Y123,S'002,A342,CH002,**A220**,/,#C3,
GH004,E132,T002,**A220**,/,#P1,

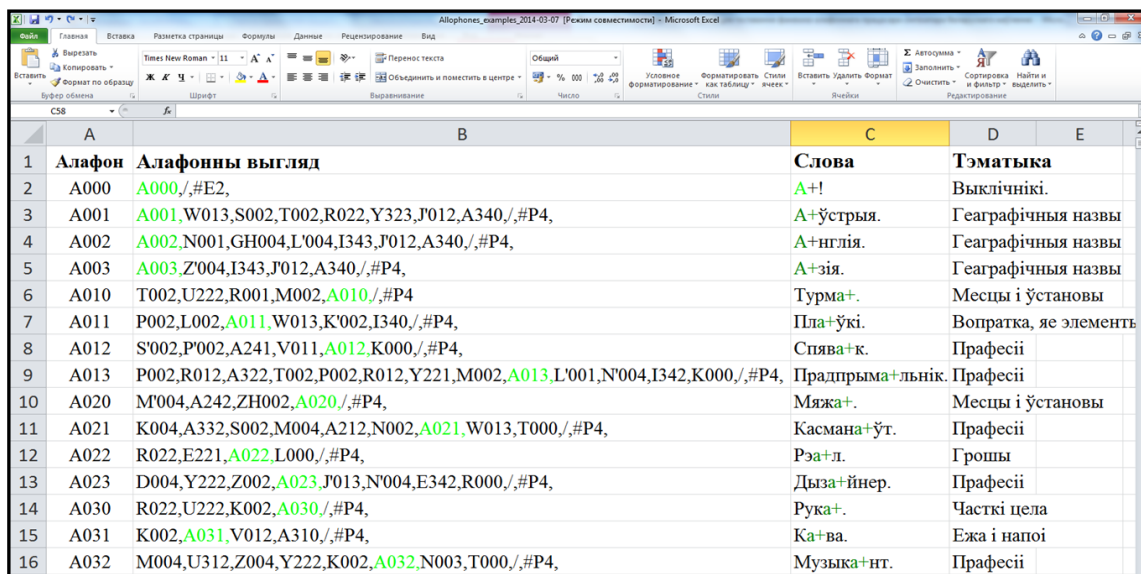
Мал. 2. Прыклад работы сэрвіса пошуку і сартыроўкі алафонаў
<http://corpus.by/allophenes-get/>

Пасля апрацоўкі сабраных тэматык стала зразумела, што дадзенага матэрыялу недастаткова для поўнага пакрыцця алафонаў беларускай мовы. Таму з электроннай версіі арфаграфічнага слоўніка беларускай мовы былі ўзяты ўсе пачатковыя формы назоўнікаў і апрацаваны пры дапамозе таго ж сэрвіса. Але і гэтага матэрыялу не хапіла. Многія пазіцыі алафонаў сустракаюцца на сутыку слоў, таму неабходна было ўлічваць не толькі адзінкавыя словы, але і розныя спалучэнні слоў. З гэтай мэтай для апрацоўкі быў узяты падручнік «Грамадазнаўства» за 10 клас, які папярэдне быў апрацаваны праграмай VocEditor. Гэта праграма дазволіла знайсці і апрацаваць усе словы з невядомай пазіцыяй націску, а таксама словы-амографы.

Для складання прыкладаў таксама былі выкарыстаныя дадзеныя пры выкарыстанні сэрвісаў «Text-to-speech synthesis», «Service of converting allophonic texts into different transcriptions» і «Orthoepic Dictionary Generator» карыстальнікамі. Вынікі працы алгарытмаў гэтых рэсурсаў выдаюцца карыстальніку ў выглядзе мэтавых транскрыпцый, а эксперту-лінгвісту і інжынеру-праграмісту — ў выглядзе ўсіх уведзеных уваходных, выходных і аналітычных (велічыня тэксту, IP-адрас, час і т.д.) дадзеных для аператыўнага выпраўлення памылак. Пасля аналізу згенераваных транскрыпцый некаторыя алафоны атрымлівалі прыклад выкарыстання.

Паколькі пасля ўсёй праведзенай працы пакрыццё алафонаў не перавышала мяжу 604 адзінкі, было вырашана шукаць алафоны, для якіх яшчэ не знойдзены адпаведнікі ўручную.

У выніку выкананай працы пакрыццё алафонаў склала 704 адзінкі, пры гэтым колькасць алафонаў для беларускай мовы будзе прапанавана паменшыць, бо створаныя адзіныя табліцы правілаў пераўтварэння «Фанема-Алафон» улічваюць абедзве мовы (рускую і беларускую), а таму ў іх сустракаюцца пэўныя недакладнасці. Напрыклад, паводле дадзеных табліц існуе пазіцыя г-фрыкатыўнага, калі яму папярэднічае любая фанема ці дапаможны сімвал, а пасля яго ідзе паўза, іншымі словамі г-фрыкатыўны на канцы слова. Па правілах жа беларускай мовы ён аглушаецца, таму гэты алафон можна выдаліць з агульнага спіса алафонаў для беларускай мовы.



	A	B	C	D	E
1	Алафон	Алафонны выгляд	Слова	Тэматыка	
2	A000	A000./, #E2,	A+!	Выклічнікі.	
3	A001	A001.W013.S002.T002.R022.Y323.J012.A340./, #P4,	A+ўстрыя.	Геаграфічныя назвы	
4	A002	A002.N001.GH004.L'004.I343.J012.A340./, #P4,	A+нглія.	Геаграфічныя назвы	
5	A003	A003.Z'004.I343.J012.A340./, #P4,	A+зія.	Геаграфічныя назвы	
6	A010	T002.U222.R001.M002.A010./, #P4	Турма+.	Месцы і ўстановы	
7	A011	P002.L002.A011.W013.K'002.I340./, #P4,	Пла+ўкі.	Вопратка, яе элементы	
8	A012	S'002.P'002.A241.V011.A012.K000./, #P4,	Спява+к.	Прафесіі	
9	A013	P002.R012.A322.T002.P002.R012.Y221.M002.A013.L'001.N'004.I342.K000./, #P4,	Прадпрыма+льнік.	Прафесіі	
10	A020	M'004.A242.ZH002.A020./, #P4,	Мяжа+.	Месцы і ўстановы	
11	A021	K004.A332.S002.M004.A212.N002.A021.W013.T000./, #P4,	Касмана+ўт.	Прафесіі	
12	A022	R022.E221.A022.L000./, #P4,	Рэа+л.	Грошы	
13	A023	D004.Y222.Z002.A023.J013.N'004.E342.R000./, #P4,	Дыза+йнер.	Прафесіі	
14	A030	R022.U222.K002.A030./, #P4,	Рука+.	Часткі цела	
15	A031	K002.A031.V012.A310./, #P4,	Ка+ва.	Ежа і напоі	
16	A032	M004.U312.Z004.Y222.K002.A032.N003.T000./, #P4,	Музыка+нт.	Прафесіі	

Мал. 3. Фрагмент лінгвістычнай рэсурсаў для пакрыцця алафонаў беларускай мовы ў прымяненні да сінтэзатара маўлення па тэксце

Другі лінгвістычны рэсурс з'яўляецца яшчэ адной базай ведаў, у якой пазначаецца арфаграфічны выгляд слова/спалучэння слоў, іх трансфармацыя ў алафонны выгляд стацыянарным і інтэрнэт-версіяй сінтэзатара маўлення, а таксама паметы лінгвіста на конт знойдзеных памылак у такой трансфармацыі. Для апісання памылак вынікі генерацыі сінтэзатараў правяраюцца, знойдзеныя памылкі пазначаюцца чырвоным колерам без унясення зменаў, а ў асобную графу ўносіцца правільны алафонны выгляд слова/спалучэння слоў. На аснове дадзенай базы ведаў праца сінтэзатараў будзе паляпшацца шляхам унясення ў іх змен з улікам пазначаных заўваг.

Такім чынам, распрацоўка лінгвістычных рэсурсаў для тэставання фанемна-алафоннага працэсара сінтэзатара беларускага маўлення з'яўляецца важнай і разам з тым складанай задачай. Магчымасць аўтаматычнай апрацоўкі некаторых дадзеных дазваляе значна палегчыць працу лінгвіста па вырашэнні дадзенай праблемы.

ЛІТАРАТУРА

1. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи. Минск: Беларуская навука, 2008.
2. Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – 2013. – Mode of access : <http://corpus.by/tts3/>. – Date of access: 09.03.2014.

3. Service of Searching and Sorting Allophones / Сэрвіс пошуку і сартыроўкі алафонаў [Electronic resource]. – 2013. – Mode of access : <http://corpus.by/allophenes-get/>. – Date of access: 08.12.2013.

