Ю.С. Гецэвіч (Мінск, АІПІ НАН Беларусі), Ю.С. Барадзіна (Мінск, БДУ)

КЛАСІФІКАЦЫЯ ФРАЗАЎ ДЫЯЛОГАЎ ПА ЭМАТЫЎНЫХ ПРЫКМЕТАХ НА МАТЭРЫЯЛЕ РУСКІХ І БЕЛАРУСКІХ МАСТАЦКІХ ТВОРАЎ

Сінтэз маўлення знаходзіць прымяненне ў розных сферах, напрыклад, пры стварэнні аўтаадказчыкаў, натуральна-моўных інтэрфейсаў, агучванні інфармацыйных паведамленняў у транспарце, на вакзале, у аэрапорце і г.д. Акрамя таго, тэхналогіі сінтэзу маўлення могуць выкарыстацца для машыннага стварэння аўдыёкніг. Звычайна гэта вымагае сур'езнай працы дыктараў і актораў, але з дапамогай існуючых напрацовак у сферы сінтэзу маўлення працэс можа быць аўтаматызаваны.

У лабараторыі распазнавання і сінтэза маўлення Аб'яднанага інстытута праблем інфарматыкі НАН Беларусі быў распрацаваны сінтэзатар, які тэхнічна ўжо можа "чытаць кнігі» [1, 269]. Тым не меньш, працэс якаснага сінтэзу маўлення яшчэ не скончаны, і застаецца некалькі істотных праблемаў, якія дагэтуль не былі вырашаны, і адна з іх — гэта праблема інтанацыі.

У п'есах аўтары спрашчаюць працу актораў з дапамогай рэмарак, якія падказваюць неабходную інтанацыю. У дыялогах празаічных тэкстаў прысутнічаюць словы аўтара, здольныя выконваць гэтую ж функцыю. Так, прыведзены ніжэй тэкст, калі будзе агучаны акторам у аўдыёкнізе, ніколі не будзе прачытаны манатонна: эмоцыі моўцы бачны праз знакі прыпынку, сутнасць самой фразы, і, не ў апошнюю чаргу, праз «падказкі» ў словах аўтара.

*– Ад нараджэньня вольныя! — **люта роў** ён. — Вось вам вашы вольнасьці! Усіх іх выразаць!*

Так, для якаснага стварэння аўдыёкніг з дапамогай сінтэзатара маўлення трэба ўлічваць эмоцыі, закладзеныя ў рэпліках герояў і словах аўтара.

Такім чынам, наступны артыкул знаёміць з першаснымі вынікамі працы, мэта якой была ў тым, каб знайсці ў фразах дыялогаў ідэнтыфікатары эматыўнай палярнасці (пазітыўны, негатыўны, нейтральны) і прапанаваць магчымыя сродкі іх фармалізацыі, а таксама пратэсціраваць іх на невялікім працоўным корпусе.

Даследванне вядзецца адначасова для беларускай і рускай моў. Для беларускай мовы ў якасці матэрыяла быў выбраны твор Алеся Рукаля «На службе князя Радзівіла», для рускай — аповесць Сяргея Даўлатава «Компромисс».

Раней супрацоўнікамі лабараторыі распазнавання і сінтэза маўлення быў распрацаваны комплекс граматык па выдзяленню з тэкста простай мовы разам са словамі аўтара і па вызначэнню ў іх пола моўцы, што таксама накіравана на аптымізацыю працэса стварэння аўдыёкніг з дапамогай сінтэзатара маўлення. З дапамогай адной з такіх граматык з корпусаў былі

выдзелены ўсе фразы простай мовы разам са словамі аўтара. Так, агульны памер выніковых канкардансаў склаў 174 рэплікі для беларускай мовы і 551 рэпліка для рускай мовы.

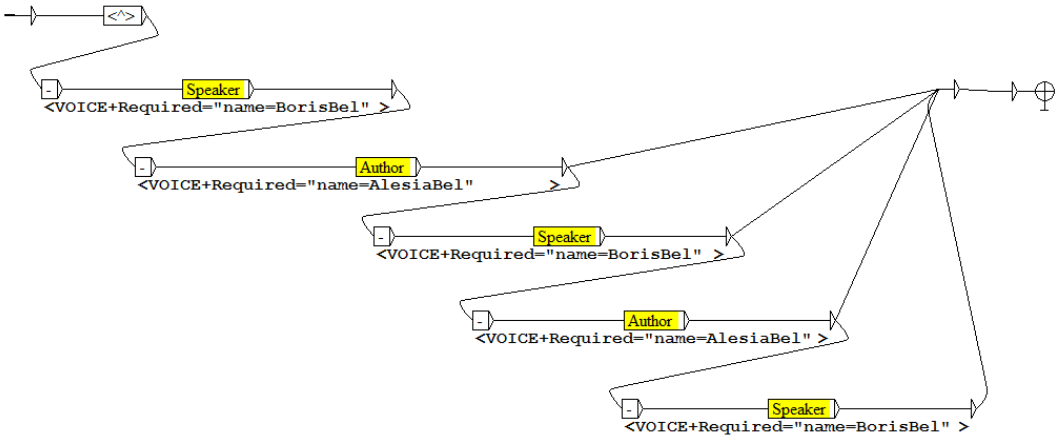
Для кожнай з рэплік рускай і беларускай моў была пазначана эматыўная палярнасць, ідэнтыфікатар, па якім яна вызначаецца і рэкамендацыі для алгарытма.

Напрыклад, фраза: — *Ну вось, — зямянін* ***дабрадушна ўсьміхнуўся*** *скрозь вусы, — нарэшце пазналі адзін аднаго. Давай абдымемся.* Эматыўная палярнасць для яе была вызначана як пазітыўная, паводле ідэнтыфікатараў *дабрадушна* і *ўсьміхнуўся*, якія ў сваю чаргу з'яўляюцца прыслоўем і дзеясловам. Прапанова для алгарытма апісваецца як ADVERB_pos+VERB_pos і ўяўляе сабой камбінацыю слоў са спісаў ADVERB_pos (пазітыўна танальныя прыслоўі) і VERB_pos (пазітыўна танальныя дзеясловы).

Пасля апрацоўкі ўсяго матэрыяла вызначылася, што ёсць два асноўных спосаба ўтварэння пазітыўнай ці негатыўнай эматыўнай танальнасці ў словах аўтара: (1) маркіраваныя дзеясловы, якія самі па сабе ствараюць неабходную танальнасць, напрыклад *усміхнуўся, працадзіў, выціснуў, роў*, і (2) ужыванне нейтральных па танальнасці дзеясловаў з эматыўна маркіраванымі прыслоўямі (*беззлобно, счастливо, грустно, сьцюдзена, суха, дабрадушна*) ці спалучэннямі слоў — назоўніка з прыназоўнікам (*з усмешкай, з цяжкасцю*), дзеепрыслоўнымі зваротамі (*цяжка дыхаючы, улыбаясь*), і іншымі акалічнасцямі спосабу дзеяння (*ненавидящим шепотом, исполненным муки голосом*). Таксама сустракаюцца выпадкі камбінавання першага і другога спосабаў.

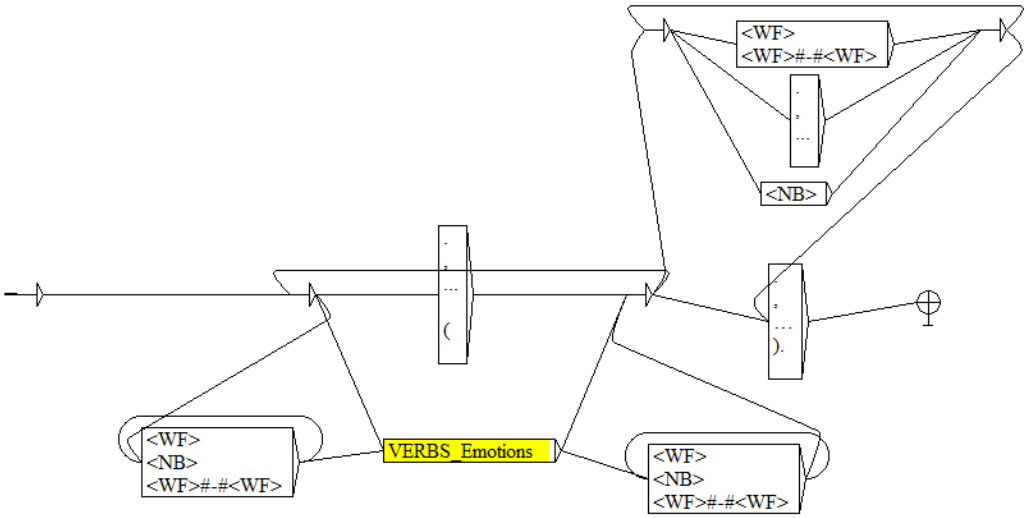
Паводле апісаных спосабаў былі складзеныя наступныя спісы ідэнтыфікатараў: VERBS_neg, VERBS_pos, VERBS_neut, ADVERBS_neg, ADVERBS_pos, COLLOCATION_neg, COLLOCATION_pos.

Каб пратэсціраваць здольнасць кампутара па спісах вызначаць эматыўную танальнасць простай мовы са словамі аўтара, былі створаны тэставыя сінтаксічныя граматыкі на базе раней распрацаваных у лабараторыі для вызначэння простай мовы з дапамогай міжнароднай лінгвістычнай праграмы NooJ [2, с. 29; 3]. На малюнке 1 прадстаўлены агульны выгляд граматыкі, дзе блокі Speaker апісваюць простую мову, а Author — словы аўтара.

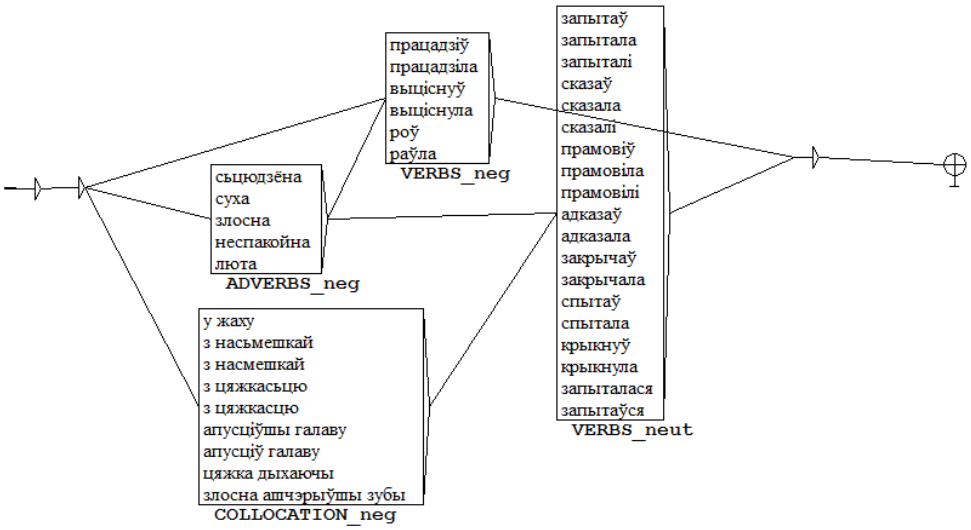


Мал. 1. Граматыка па вызначэньню простай мовы, агульны выгляд.

У падграфе Author (гл. малюнак 2) захоўваецца яшчэ адна граматыка, якая апісвае структуру слоў аўтара і ў якой у блоку VERBS_Emotions размяшчаюцца вышэйапісаныя спісы ідэнтыфікатараў, разам са знакамі прыпынку, уласцівымі для простай мовы (малюнак 3).



Мал. 2. Падграф Author, апісанне слоў аўтара.



Мал. 3. Выгляд падграфа VERBS_Emotions для негатыўнай танальнасці беларускай мовы.

Граматыкі для рускай і беларускай моў, пазітыўнай і негатыўнай эматыўных танальнасцяў былі прыменены да матэрыяла, і праграма знайшла ўсе вызначаныя экспертам сказы, напрыклад, — *Спортсменом будет, — улыбается главный врач Михкель Теппе; — Довлатов, — исполненным муки голосом произнес Туронок, — Довлатов, я вас уволю; — Што? — у жаху закрычалі ваяры* і г.д.

Такім чынам, былі сабраны калекцыі ідэнтыфікатараў эматыўнай палярнасці тэкста для рускай і беларускай моў, вынесены рэкамендацыі па іх фармалізацыі і распрацаваны тэставыя граматыкі па знаходжанні ў корпусе эматыўна палярных фразаў простай мовы са словамі аўтара.

Тэма патрабуе далейшай распрацоўкі, таму будуць павялічаныя тэкставыя корпусы і, адпаведна, спісы ідэнтыфікатараў дзеля паляпшэння якасці працы граматык. Акрамя таго, вызначэнне эмоцыяў будзе весціся не толькі па словах аўтара, але і па простай мове.

ЛІТАРАТУРА

1. Гецэвіч, Ю.С. Аўтаматызацыя шматгаласавога стварэння аўдыёкніг на беларускай мове з дапамогай сінтэзатараў маўлення па тэксце / Ю.С. Гецэвіч, Т.І. Окрут, Б.М. Лабанаў // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013): доклады XII Международной конференции (Минск, 20 ноября 2013 г.). – Минск: ОИПИ НАН Беларуси, 2013. – С. 269–276.

2. Hetsevich, Y. Overview of Belarusian And Russian dictionaries and their adaptation for NooJ //Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern. Conf. / eds. Vučković Kristina, Bekavac Božo, Silberztein Max. – Newcastle: Cambridge Scholars Publishing, 2012. – P. 29–40.

3. Лінгвістычны працэсар NooJ [Электронны рэсурс]. – 2002. – Рэжым доступу: http://nooj4nlp.net/pages/nooj.html. – Дата доступу: 10.03.2014