УДК 811.161.3'373.423.1'322
https://doi.org/10.37661/1816-0301-2023-20-4-87-100

Арыгінальны артыкул
Original Paper

Мадэль аўтаматызаванай ідэнтыфікацыі амографаў для беларускай мовы

Ю. С. Гецэвіч, Я. С. Зяноўка✉, Д. І. Латышэвіч, А. А. Бакуновіч, А. Я. Драгун, М. А. Казлова

Аб'яднаны інстытут праблем інфарматыкі
Нацыянальнай акадэміі навук Беларусі,
вул. Сурганава, 6, Мінск, 220012, Беларусь
✉*E-mail: evgeniakacan@gmail.com*

Анатацыя
Мэты. Мэтай працы з'яўляецца апісанне прататыпнай сістэмы для аўтаматызаванага здымання аманіміі ў электронных тэкстах на беларускай і рускай мовах. Гэта звязана з актуальнай праблемай аўтаматычнай апрацоўкі тэкстаў на марфалагічным узроўні, працэс якой ускладняецца флектыўнасцю беларускай мовы з разнастайнай і багатай сістэмай марфалагічных характарыстык часцін мовы.
Метады. У працы выкарыстоўваюцца правілавыя метады ідэнтыфікацыі амаграфіі і метады, заснаваныя на ведах.
Вынікі. Прапанаваны метады і падыходы для праектавання сістэм аўтаматычнага вызначэння амографаў. Падрабязна прадстаўлены метад, заснаваны на ведах, на аснове якога распрацаваны пакрокавы алгарытм ідэнтыфікацыі амографаў і рэалізаваны эфектыўны і хуткадзейны прататып для іх здымання на рускай і беларускай мовах.
Заключэнне. Прадстаўлены працоўны прататып пошуку амографаў, які з'яўляецца першым рэсурсам па здыманні шматзначнасці для беларускай мовы ў адкрытым доступе.

Ключавыя словы: аманімія, здыманне аманіміі, шматзначнасць, аўтаматычная апрацоўка электронных тэкстаў, беларуская мова, слоўнік

Для цытавання. Мадэль аўтаматызаванай ідэнтыфікацыі амографаў для беларускай мовы / Ю. С. Гецэвіч [і інш.] // Информатика. – 2023. − Т. 20, № 4. – С. 87–100.
https://doi.org/10.37661/1816-0301-2023-20-4-87-100

Канфлікт інтарэсаў. Аўтары заяўляюць аб адсутнасці канфлікту інтарэсаў.

Паступіла ў рэдакцыю | Received 22.08.2023
Падпісана ў друк | Accepted 01.11.2023
Апублікавана | Published 29.12.2023

A model of homographs automatic identification for the Belarusian language

Yuras S. Hetsevich, Yauheniya S. Zianouka✉, David I. Latyshevich, Andrey A. Bakunovich, Anastasia Ya. Drahun, Margarita A. Kazlova

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus
✉*E-mail: evgeniakacan@gmail.com*

Abstract
Objectives. A prototype system for automated removal of homonyms in Belarusian and Russian electronic texts is described. This is due to the urgent problem of automatic text processing at the morphological level, the process of which is complicated by the inflection of the Belarusian language with a diverse and rich system of morphological characteristics of parts of speech.
Methods. The work uses regular homographs identification methods and knowledge-based methods.
Results. Methods and approaches for designing systems for automatic detection of homographs are proposed. An algorithm for identifying homographs on the basis of knowledge-based method has been developed. An effective and fast-acting prototype for their removal in Russian and Belarusian has been implemented.
Conclusion. A working prototype of the homograph search is presented, which is the first resource for removing ambiguity for the Belarusian language in open access.

Keywords: homonymy, removal of homonyms, ambiguity, automatic processing of electronic texts, the Belarusian language, dictionary

For citation. Hetsevich Yu. S., Zianouka Ya. S., Latyshevich D. I., Bakunovich A. A., Drahun A. Ya., Kazlova M. A. *A model of homographs automatic identification for the Belarusian language*. Informatika [*Informatics*], 2023, vol. 20, no. 4, pp. 87−100 (In Russ.). https://doi.org/10.37661/1816-0301-2023-20-4-87-100

Conflict of interest. The authors declare of no conflict of interest.

Уводзіны. Натуральная мова ўяўляе сабой вялікую адкрытую шматузроўневую сістэму знакаў, якая ўзнікла для абмену інфармацыяй у працэсе практычнай дзейнасці чалавека. Складанасць яе мадэлявання абумоўлена бесперапыннымі зменамі ў сувязі з гэтай дзейнасцю. Тэкст на натуральнай мове складаецца з асобных элементаў, якія ў сваім адзінстве аб'яднаны рознымі тыпамі лексічнай, граматычнай і лагічнай сувязяў. Існуюць разнастайныя спосабы члянення тэксту на адзінкі. Як вынік, немагчыма распрацаваць адзіную фармальную мадэль канкрэтнай натуральнай мовы і пабудаваць адпаведны лінгвістычны працэсар. Патрабуюцца пастаяннае папаўненне ведаў на ўсіх яе ўзроўнях і карэкцыя існуючых мадэляў. Адной з самых вялікіх складанасцяў пры апрацоўцы тэкстаў на натуральнай мове з'яўляецца двухзначнасць яе адзінак, што выяўляецца ў з'явах шматзначнасці, сінаніміі і асабліва аманіміі. Праблема аманіміі яскрава праяўляецца ў прыкладных навуках. Яна закранае такія практычныя напрамкі, як аўтаматычны пераклад, аўтаматычнае рэферыраванне тэкстаў, стварэнне рознага роду інтэлектуальных сістэм, складанне слоўнікаў [1].

Амонімы (ад грэч. *homos* – аднолькавы, *onyma* – імя) – словы, якія аднолькава гучаць і пішуцца, але маюць розныя, не звязаныя паміж сабой значэнні (напрыклад, «тур» – першабытны дзікі бык, «тур» – этап адбору і «тур» – паездка па розных мясцінах) [2]. Амонімы складаюць значны пласт лексікі (толькі амографаў у беларускай мове налічваецца каля 4000) і падзяляюцца на наступныя групы. Лексічныя – словы, якія аднолькава пішуцца і гучаць, аднак не маюць агульных элементаў сэнсу. У марфалагічных амонімах супадаюць формы аднаго і таго ж слова (лексемы), а лексіка-марфалагічныя амонімы (найбольш часты варыянт) узнікаюць пры аднолькавых словаформах дзвюх розных лексем. Сінтаксічныя амонімы апісваюць неадназначнасць сінтаксічнай структуры, што прыводзіць да розных інтэрпрэтацый. У галіне аўтаматычнай апрацоўкі мовы адной з найскладаных праблем выступае вызначэнне амонімаў для ліквідацыі неадназначнасці ў сэнсе, выпраўлення памылак і разумення адценняў у значэнні слоў.

Амографы (ад грэч. *γμός* – аднолькавы, *γράφω* – пішу) – словы, якія супадаюць у напісанні, але адрозніваюцца ў вымаўленні (у беларускай мове часцей за ўсё з-за адрозненняў у націску (напрыклад, *па+ра* і *пара+*, *му+зыка* і *музы+ка*). Да амографаў могуць адносіцца словы, якія маюць рознае значэнне, а таксама розныя формы аднаго і таго ж слова. Іх здыманне дазваляе павысіць дакладнасць і якасць апрацоўкі тэкставых даных. Для аўтаматызаванага вырашэння неадназначнасці слоў, у прыватнасці амографаў, выкарыстоўваюцца такія падыходы, як дэтэрмінаваныя правілы, якія працуюць на аснове лексічных і граматычных даных; базы ведаў аб навакольным свеце і анталогіі, якія даюць магчымасць улічваць экстралінгвістычныя даныя; імавернасныя аналізатары, якія ўлічваюць статыстычныя даныя мовы [3–6]. Кожны з прыведзеных падыходаў мае свае абмежаванні па эфектыўнасці і не дае вынікаў на самым высокім узроўні. Статыстычныя алгарытмы здымаюць аманімію на этапе марфалагічнага аналізу тэксту з выкарыстаннем статыстыкі ўжывання граматычных прыкмет слоў з карпусоў тэкстаў, размечаных уручную [7]. Дадзеныя метады часцей за ўсё прымяняюцца для аналітычных моў (напрыклад, англійскай), у якіх граматычныя адносіны маюць тэндэнцыю да перадачы ў асноўным праз сінтаксіс, асобныя службовыя словы (прыназоўнікі, мадальныя дзеясловы і г. д.), фіксаваны парадак слоў, кантэкст і (або) інтанацыйныя варыяцыі, а не праз словазмяненне з дапамогай залежных марфем (канчаткаў, суфіксаў, прыставак і г. д.). Разнастайнасць аманіміі ў беларускай і рускай мовах тлумачыцца наяўнасцю флексіі – фармантаў, якія спалучаюць адразу некалькі значэнняў, што і выклікае шматзначнасць і неаднароднасць слоў. Менавіта таму, што беларуская і руская мовы характарызуюцца адвольным парадкам слоў у сказе, узнікае складанасць прымянення матэматычных мадэлей вызначэння аманіміі. У сувязі з гэтым для дадзеных моў часцей за ўсё выкарыстоўваюць метады, заснаваныя на правілах.

У сучасных даследаваннях апісваюцца чатыры асноўныя метады здымання шматзначнасці:

1) метады, заснаваныя на ведах (dictionary- і knowledge-based methods), пераважна выкарыстоўваюць слоўнікі, тэзаўрус, лексікаграфічныя базы даных [8, с. 110]. Яны грунтуюцца на гіпотэзе, што словы, якія знаходзяцца побач у тэксце, звязаны адно з адным і гэтую сувязь можна назіраць у азначэннях слоў і іх значэннях. Два (ці больш) словы могуць аказацца блізкімі, калі ў абодвух з іх будзе выяўлена пара значэнняў з найбольшым перасячэннем слоў у іх азначэннях у слоўніку [9–12];

2) метады навучання з настаўнікам (supervised methods) выкарыстоўваюць размечаныя карпусы тэкстаў для трэніроўкі класіфікатараў [1]. Яны грунтуюцца на наступным меркаванні: кантэкст слова, што разглядаецца, дае дастатковую інфармацыю для таго, каб вылічыць, у якім значэнні ў дадзеным выпадку яно ўжываецца (а значыць, веды, атрыманыя з слоўнікаў і тэзаўрусаў, выдаляюцца як лішнія). Усе мадэлі навучання з настаўнікам прымяняліся да праблемы WSD (бел. вырашэнне лексічнай мнагазначнасці), у тым ліку звязаныя з імі тэхнікі, такія як выбар пераменных, аптымізацыя параметраў і змешаныя мадэлі (англ. ensemble learning);

3) метады частковага навучання з настаўнікам (minimally-supervised methods) базіруюцца на другасных ведах, такіх як вызначэнне тэрмінаў у тлумачэннях слоў ці выраўнаваны двухмоўны корпус [4];

4) метады навучання без настаўніка (unsupervised methods) не прадугледжваюць выкарыстанне якіх-небудзь знешніх даных і працуюць толькі з неанатаванымі карпусамі (raw unannotated corpora). Таксама яны вядомы пад тэрмінам кластарызацыі і распазнавання сэнсу слоў [13, 14]. Асноўная ідэя гэтага метаду заключаецца ў тым, што «падобныя значэнні сустракаюцца ў падобных кантэкстах» і такім чынам яны могуць быць выняты з тэксту з дапамогай кластарызацыі з ужываннем некаторай меры падабенства кантэкстаў. Таму новыя кантэксты могуць быць аднесены да аднаго з бліжэйшых кластараў.

Існуюць іншыя метады, якія адрозніваюцца ад вышэйпералічаных прынцыпаў: вызначэнне дамінантнага значэння слова (Determining Word Sense Dominance); вырашэнне, заснаванае на тэмах корпуса (Domain-Driven Disambiguation); WSD, якое выкарыстоўвае крос-моўныя даныя (Cross-Lingual Evidence) [1].

Электронныя лінгвістычныя рэсурсы для аўтаматызаванага здымання беларускамоўнай амаграфіі. Агляд наяўных навукова-папулярных крыніц і разнастайных даследаванняў сведчыць аб вялікай зацікаўленасці навукоўцаў праблемай аўтаматычнага здымання

аманіміі. Гэта прыведзена ў працах [1–14]. Аднак у адкрытым доступе адсутнічае інфармацыя аб прыкладной рэалізацыі праграмных прадуктаў здымання аманіміі для беларускай мовы. На падставе аналізу праблемнага поля супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі (https://ssrlab.by/) была распрацавана мадэль аўтаматызаванай сістэмы па вызначэнні амографаў, двухсэнсоўных слоў у тэксце беларускай літаратурнай мовы [15]. Асноўны кірунак дзейнасці лабараторыі – высакаякасны сінтэз маўлення па тэксце, які заключаецца ў шматэтапным пераўтварэнні тэксту ў маўленне. Немалаважнай задачай з'яўляецца перапрацоўка электроннага тэксту, якая, акрамя этапаў ачысткі тэксту, дэшыфроўкі лікаў, абрэвіятур, замежных слоў і карэкціроўкі «ё», уключае вызначэнне амонімаў [16]. Для вырашэння праблемы здымання шматзначнасці слоў быў распрацаваны самастойны прататып у форме вэб-сэрвіса *«Ідэнтыфікатар амографаў»*, які даступны для вольнага выкарыстання ў Інтэрнэце па адрасе https://corpus.by/HomographIdentifier/?lang=be [17]. Для яго стварэння аўтарамі даследавання быў выкарыстаны метад, заснаваны на ведах (dictionary-based method). Дадзены падыход заснаваны на вялікім наборы слоўнікаў, згодна з якімі памылковыя значэнні шматзначных слоў выключаюцца з улікам іх значэнняў, што параўноўваюцца паміж сабой.

На ўваход сэрвісу падаецца электронны тэкст. Па выніках апрацоўкі карыстальнік атрымлівае спіс знойдзеных у тэксце амографаў з іх падрабязнымі данымі і, згодна сэнсу апрацаванага тэксту, сам выбірае правільны варыянт. Бо́льшая ўвага надаецца амографам, таму што сістэма апрацоўвае электронны тэкст, гэта значыць аналізуе графічныя знакі. Такім чынам, здаецца магчымым апрацоўваць толькі графічную форму слова з рознымі значэннямі, у адрозненне ад амафонаў, якія маюць адзіную форму вымаўлення і розную графічную сістэму.

У аснову прататыпа пакладзены наступныя электронныя слоўнікі, згодна якім адбываецца вызначэнне амографаў:

SBM1987 (паводле публікацыі «Слоўнік беларускай мовы. Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / пад рэд. М. В. Бірылы. – Мінск, 1987»);

SBM2012initial (пачатковыя формы паводле публікацыі «Слоўнік беларускай мовы / навук. рэд. А. А. Лукашанец, В. П. Русак. – Мінск : Беларус. навука, 2012»);

NOUN2013 (паводле публікацыі «Граматычны слоўнік назоўніка / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

ADJECTIVE2013 (паводле публікацыі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

NUMERAL2013 (паводле публікацыі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

PRONOUN2013 (паводле публікацыі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

VERB2013 (паводле публікацыі «Граматычны слоўнік дзеяслова / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

ADVERB2013 (паводле публікацыі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»);

ZALIZNIAK (паводле публікацыі «Грамматический словарь русского языка: Словоизменение / А. А. Зализняк. – М. : Русский язык, 1980. – 880 с.»);

CMU (паводле «Carnegie Mellon University Pronouncing Dictionary»).

Акрамя вышэй апісаных выданняў, распрацаваны дадатковыя слоўнікі *UWP_BE, UWP_RU,* якія папаўняюцца неалагізмамі і словамі, адсутнымі ў папярэдніх слоўніках, за кошт апрацоўкі электронных тэкстаў сэрвісамі платформы для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by (https://corpus.by/). Карыстальнік можа абраць з дадзенага спісу тыя слоўнікі, якія яму патрэбны для аналізу тэксту. Напрыклад, для пошуку амаграфіі ў тэксце на беларускай мове мэтазгодна абраць слоўнікі *SBM1987, SBM2012initial* і іншыя, на рускай мове – толькі слоўнік *ZALIZNIAK*, на англійскай – толькі слоўнік *CMU*.

Апрацоўка тэксту сэрвісам на дадзены момант мае наступныя асаблівасці:

1. Пошук ажыццяўляецца асобна па кожным абраным карыстальнікам слоўніку. Гэта значыць, што выпадкі міжмоўнай амаграфіі, напрыклад: бел. «выго́да» 1) усё, што задавальняе максімальныя запатрабаванні, чым зручна карыстацца; 2) прыволле) – рус. «вы́года» (прыбытак, даход, які атрымліваецца з чаго-небудзь) ці бел. «свая́» (прыналежны займеннік) – рус. «сва́я» (брус, які забіваецца ў грунт для апоры збудавання), выяўлены не будуць. Калі міжмоўная амаграфія прысутнічае, але для слова могуць быць знойдзены амаграфічныя варыянты ў межах адной мовы, напрыклад, як у выпадках «го́дны» і «годны́» для рускай мовы (адна з форм прысутнічае таксама ў беларускай мове) ці «ра́са» і «раса́» для беларускай мовы (адна з форм прысутнічае таксама ў рускай мове), сэрвісам будзе прадэманстравана толькі амаграфія ў межах адной мовы.

2. Для ажыццяўлення карэктнага пошуку ў слоўніках словы прыводзяцца да аднаго рэгістру – верхняга ці ніжняга, у залежнасці ад слоўніка. Таксама на дадзены момант слоўнікі, якія можна падключыць, не змяшчаюць уласныя імёны. Выпадкі кшталту рус. «Ко́ли» (мужчынскае імя ў родным склоне) і «коли́» (загадны лад 2-й асобы адзіночнага ліку дзеяслова «колоть») ці «Ма́ши» (жаночае імя ў родным склоне) і «маши́» (загадны лад 2-й асобы адзіночнага ліку дзеяслова «махать») не будуць разглядацца сэрвісам як амаграфічныя.

3. Літары рускага алфавіта «е» і «ё» лічацца сэрвісам рознымі. Падобныя прыклады: «же́ны» («жёны») і «жены́», не будуць пазначаны як амаграфічныя.

4. Паводле Оксфардскага слоўніка, у англійскай мове слова «homograph» азначае *«a word that is spelt like another word but has a different meaning from it, and may have a different pronunciation, for example bow /baʊ/, bow /bəʊ/»* (бел. «слова, якое пішацца так, як іншае слова, але мае рознае з ім значэнне і можа мець рознае вымаўленне, напрыклад, *bow /baʊ/, bow /bəʊ/*»). Англійскія амографы вызначаюцца сэрвісам па іншых правілах. Напрыклад, амографам будзе з'яўляцца слова «awesome», паколькі яно мае розныя варыянты вымаўлення пачатковага галоснага, а таксама «record», паколькі мае адно напісанне для слоў з рознымі значэннямі (дзеяслова і назоўніка), да таго ж вымаўленне дадзеных слоў адрозніваецца націскамі.

Алгарытм працы сэрвіса «Ідэнтыфікатар амографаў». Алгарытм дае магчымасць атрымаць спіс амографаў згодна ўбудаваным слоўнікам падчас апрацоўкі электроннага тэксту на беларускай, рускай і англійскай мовах. Больш якасныя вынікі сістэма прапануе для беларускай мовы, што звязана з колькасцю ўбудаваных слоўнікаў для яе. Аднак, згодна ўніверсальнаму характару алгарытму, пры больш багатым слоўнікавым напаўненні сістэмы ёсць магчымасць апрацоўваць амографы іншых моў.

Уваходныя даныя алгарытму ўяўляюць сабой наступны набор інструментаў:

– карыстальніцкі тэкставы ўвод, *UText*;
– мноства слоўнікаў, падключаных карыстальнікам, *Dictionaries*;
– мноства кірылічных сімвалаў у верхнім і ніжнім рэгістрах, *LettersCyr*;
– мноства лацінскіх сімвалаў у верхнім і ніжнім рэгістрах, *LettersLat*;
– мноства дадатковых сімвалаў (апострафы, сімвалы націскаў, злучок), *AdditionalCharacters*.

Алгарытм складаецца з чатырох асноўных каманд, якія падзяляюцца на больш падрабязныя крокі:

Крок 1. Апрацоўка ўваходнай інфармацыі.

1.1. Стварэнне мноства асноўных сімвалаў *MainCharacters* шляхам зліцця мностваў *LettersCyr* і *LettersLat*.

1.2. Раздзяленне *UText* на мноства *ParagrapsArr* паводле сімвала пераводу радка. Стварэнне пустога асацыятыўнага масіву U*niqueWordsArr* для наступнага запісу ўнікальных слоў, знойдзеных у *ParagrapsArr*.

1.3. Для кожнага элемента *Paragraph* мноства *ParagrapsArr* выканаць 1.3.1–1.3.4.

1.3.1. Выдаліць з *Paragraph* усе пачатковыя і канцавыя сімвалы водступаў.

1.3.2. Калі *Paragraph* = ∅, перайсці да наступнага элемента *Paragraph* і выканаць крок 1.3.1.

1.3.3. Стварэнне масіву *WordsArr* і запаўненне яго элементамі, складзенымі са зместу *Paragraph*. Элементам лічыцца спецыяльны набор сімвалаў ад пачатку радка да прабелу, ад

прабелу да прабелу і ад прабелу да канца (або сімвала пераводу) радка, які адпавядае шаблону <[1 сімвал MainCharacters]+[0 або больш сімвалаў MainCharacters і AdditionalCharacters]> альбо шаблону <[0 або больш сімвалаў MainCharacters]>. Стварэнне пераменнай *WordsCnt* і запіс у яе колькасці элементаў *WordsArr*.

1.3.4. Для I = 0; I < WordsCnt, I++ выканаць 1.3.4.1 і 1.3.4.2.

1.3.4.1. Стварэнне пераменнай *Word = WordsArr[I]*.

1.3.4.2. Калі *Word* ≠ *Ø*, прывесці ўсе сімвалы Word да ніжняга рэгістра і замяніць усе пачатковыя сімвалы «ў» на «у» (неабходна для наступнага карэктнага пошуку ў слоўніках). Паспрабаваць запісаць Word у асацыятыўны масіў *UniqueWordsArr*, выкарыстоўваючы значэнне Word у якасці ключа і прысвоіўшы яму значэнне 1. Калі слова паводле ключа Word ужо прысутнічае ў *UniqueWordsArr*, інкрэментаваць адпаведнае яму значэнне.

Крок 2. Пошук амографаў.

2.1. Стварэнне асацыятыўнага масіву *UniqueHomographsArr* для наступнага запісу ўнікальных амографаў і спадарожнай інфармацыі.

2.2. Для кожнага *UniqueWord* у *UniqueWordsArr* і кожнага *Dictionary* у *Dictionaries* выканаць 2.2.1–2.2.3.

2.2.1. Стварыць асацыятыўныя масівы *AccentArr* і *ResultArr*. Калі *Dictionary=Homographs_*Be (гэта значыць, што на дадзеным кроку цыкла праглядаецца слоўнік з такою назвай), праверыць, ці прысутнічае ў *Dictionary* элемент, ідэнтычны бягучаму элементу *UniqueWord*. Калі прысутнічае, запісаць у масіў *AccentArr* значэнне элемента слоўнікавага артыкула *Category*, які адпавядае бягучаму *UniqueWord*, а ў масіў *ResultArr* – значэнне адпаведных элементаў слоўнікавага артыкула *AccentedWord*, *Category* і *LexemeId*, пасля чаго перайсці да 2.2.3. Інакш – перайсці да наступнага кроку.

2.2.2. Калі *Dictionary≠Homographs_Be*, выканаць 2.2.2.1–2.2.2.3.

2.2.2.1. Стварыць пераменную *UniqueWordWithSlashes* і запісаць у яе значэнне бягучага *UniqueWord*, акружанае сімваламі слэша. Калі пры гэтым *UniqueWordWithSlashes* складаецца толькі з сімвалаў, якія належаць да мноства *LettersLat*, акружаных сімваламі слэша, прывесці ўсе сімвалы *UniqueWordWithSlashes* да верхняга рэгістра. Стварыць пераменную *Query* для захоўвання запыту да слоўнікавай базы і сфарміраваць яе значэнне згодна з шаблонам «SELECT * FROM %s WHERE word='%s», дзе %s – значэнне *UniqueWordWithSlashes*. Дадзены крок неабходны для ажыццяўлення карэктнага пошуку па слоўнікавых базах.

2.2.2.2. Калі паводле запыту, запісанага ў пераменнай *Query*, у бягучым *Dictionary* удалося знайсці якія-небудзь даныя, то да таго моманту, пакуль стандартная функцыя звароту да базы даных не дойдзе да апошняга радка Row у табліцы вынікаў, выконваць 2.2.2.2.1–2.2.2.2.3.

2.2.2.2.1. Стварыць пераменную *AccentedWord*. Калі ў бягучым *Row* зададзены элемент слоўнікавага артыкула *Accent*, запісаць у *AccentedWord* яго значэнне, прыведзенае да ніжняга рэгістра; калі пры гэтым у *AccentedWord* адсутнічае сімвал «+», завяршыць дадзены крок цыкла, перайсці да наступнага *Row* і выканаць 2.2.2.2.1 яшчэ раз. Калі элемент *Accent* не зададзены, запісаць у *AccentedWord* значэнне элемента *Transcription*, прыведзенае да ніжняга рэгістра.

2.2.2.2.2. Калі ў бягучым Row зададзены і элемент *LexemeId*, і элемент *Tag*, рэініцыялізаваць *Query* значэннем «SELECT * FROM tags WHERE tag='%s' LIMIT 1», дзе %s – значэнне элемента *Tag*, і паспрабаваць знайсці якія-небудзь даныя паводле гэтага запыту. Калі даныя ўдалося знайсці, то для кожнага радка табліцы вынікаў запісаць у створаную пераменную *CategoryVar* значэнне элемента слоўнікавага артыкула *Category* (калі ён зададзены), а ў пераменную *LexemeIdVar* – значэнне элемента слоўнікавага артыкула *LexemeId*, пасля чаго вызваліць сістэмныя рэсурсы ад вынікаў запыту.

2.2.2.2.3. Запісаць у масіў *AccentArr* значэнне *CategoryVar*, у масіў *ResultArr* – значэнні *AccentedWord*, *CategoryVar* і *LexemeIdVar*.

2.2.2.3. Ачысціць сістэмныя рэсурсы ад вынікаў запытаў да слоўнікавых баз даных.

2.2.3. Калі ў *AccentArr* прысутнічае больш за адзін элемент, выканаць 2.2.3.1–2.2.3.8.

2.2.3.1. Калі ў *UniqueHomographsArr* не зададзены элемент, роўны бягучаму *UniqueWord*, задаць такі элемент у выглядзе пустога масіву. Стварыць пераменную *AccentList* і ініцыялізаваць яе пустым значэннем.

2.2.3.2. Для кожнага элемента *AccentedWord* у *AccentArr* (дадзенаму элементу адпавядае свой масіў *CategoryArr*) здзейсніць наступныя дзеянні ў прамым парадку:

– дадаць да *AccentList* значэнне бягучага *AccentedWord*;

– калі *CategoryArr≠Ø,* дадаць да *AccentList* радок « <small>(», значэнне ўсіх унікальных элементаў *CategoryArr,* аб'яднаных у радок з дапамогай радка «, », і радок «)</small>»;

– дадаць да *AccentList* радок «
\n».

2.2.3.3. Замяніць у *AccentList* усе камбінацыі [любы сімвал + сімвал «=»] на той жа сімвал, але забяспечаны сімвалам пабочнага націску і акружаны тэгамі HTML для надання яму сіняга колеру. Замяніць у *AccentList* усе камбінацыі [любы сімвал + сімвал «+»] на той жа сімвал, але забяспечаны сімвалам асноўнага націску і акружаны тэгамі HTML для надання яму чырвонага колеру.

(2.2.3.2 і 2.2.3.3 – неабходны для фарміравання карэктнай выдачы.)

2.2.3.4. Выдаліць з *ResultArr* значэнні, якія паўтараюцца. Стварыць асацыятыўныя масівы *LexemeIdArr* і *CategoryArr*.

2.2.3.5. Калі элемент *HomographArray*, які адпавядае бягучаму *UniqueWord*, які, у сваю чаргу, адпавядае бягучаму *Dictionary*, не зададзены (тут і далей гэты элемент будзе названы як X), то для кожнага элемента *WordInfo* у *ResultArr* праверыць, ці зададзены ў ім элемент слоўнікавага артыкула *Category* і элемент *LexemeId*. Калі хаця б адзін з дадзеных элементаў не зададзены, запісаць у X значэнне *Accents*, роўнае *AccentList*, і значэнне *Type*, роўнае «-» (гэта значыць тып амаграфіі не вызначаны).

2.2.3.6. Калі пасля папярэдняга кроку X не зададзены, то для кожнага элемента *WordInfo* у *ResultArr* праверыць, ці зададзены ў *LexemeIdArr* элемент *WordInfo*, які ў сваю чаргу адпавядае элементу слоўнікавага артыкула *LexemeId*. Калі не зададзены, зрабіць яго роўным адзінцы, інакш запісаць у X значэнне *Accents*, роўнае *AccentList*, і значэнне *Type*, роўнае «one paradigm» (гэта значыць амографы належаць да адной парадыгмы скланення або спражэння).

2.2.3.7. Калі пасля папярэдняга кроку X не зададзены, то для кожнага элемента *WordInfo* у *ResultArr* праверыць, ці зададзены ў *CategoryArr* элемент *WordInfo*, які ў сваю чаргу адпавядае элементу слоўнікавага артыкула *Category*. Калі не зададзены, зрабіць яго роўным адзінцы, інакш запісаць у X значэнне *Accents*, роўнае *AccentList*, і значэнне *Type*, роўнае «one part of speech» (гэта значыць амографы належаць да адной часціны мовы).

2.2.3.8. Калі пасля папярэдняга кроку X не зададзены, то запісаць у X значэнне *Accents*, роўнае *AccentList*, і значэнне *Type*, роўнае «different parts of speech» (гэта значыць амографы належаць да розных часцін мовы).

Крок 3. Фарміраванне кантэкстаў. Для кожнага элемента *Paragraph* мноства *ParagrapsArr* выканаць 3.1–3.4.

3.1. Выдаліць з *Paragraph* усе пачатковыя і канцавыя сімвалы водступаў.

3.2. Калі *Paragraph=Ø*, перайсці да наступнага элемента *Paragraph* і выканаць 3.1.

3.3. Стварэнне масіву *WordsArr* і запаўненне яго элементамі, складзенымі са зместу *Paragraph*. «Элементам» лічыцца спецыяльны набор сімвалаў ад пачатку радка да прабелу, ад прабелу да прабелу і ад прабелу да канца (або сімвала пераводу) радка, які адпавядае шаблону <[1 сімвал MainCharacters]+[0 або больш сімвалаў MainCharacters і AdditionalCharacters]> альбо шаблону <[0 або больш сімвалаў MainCharacters]>. Стварэнне пераменнай *WordsCnt* і запіс у яе колькасці элементаў *WordsArr*.

3.4. Для *I = 0; I < WordsCnt, I++* выканаць 3.4.1–3.4.3.

3.4.1. Стварэнне пераменнай *Word = WordsArr[I]*.

3.4.2. Калі *Word≠Ø,* прывесці ўсе сімвалы Word да ніжняга рэгістра і замяніць усе пачатковыя сімвалы «ў» на «у».

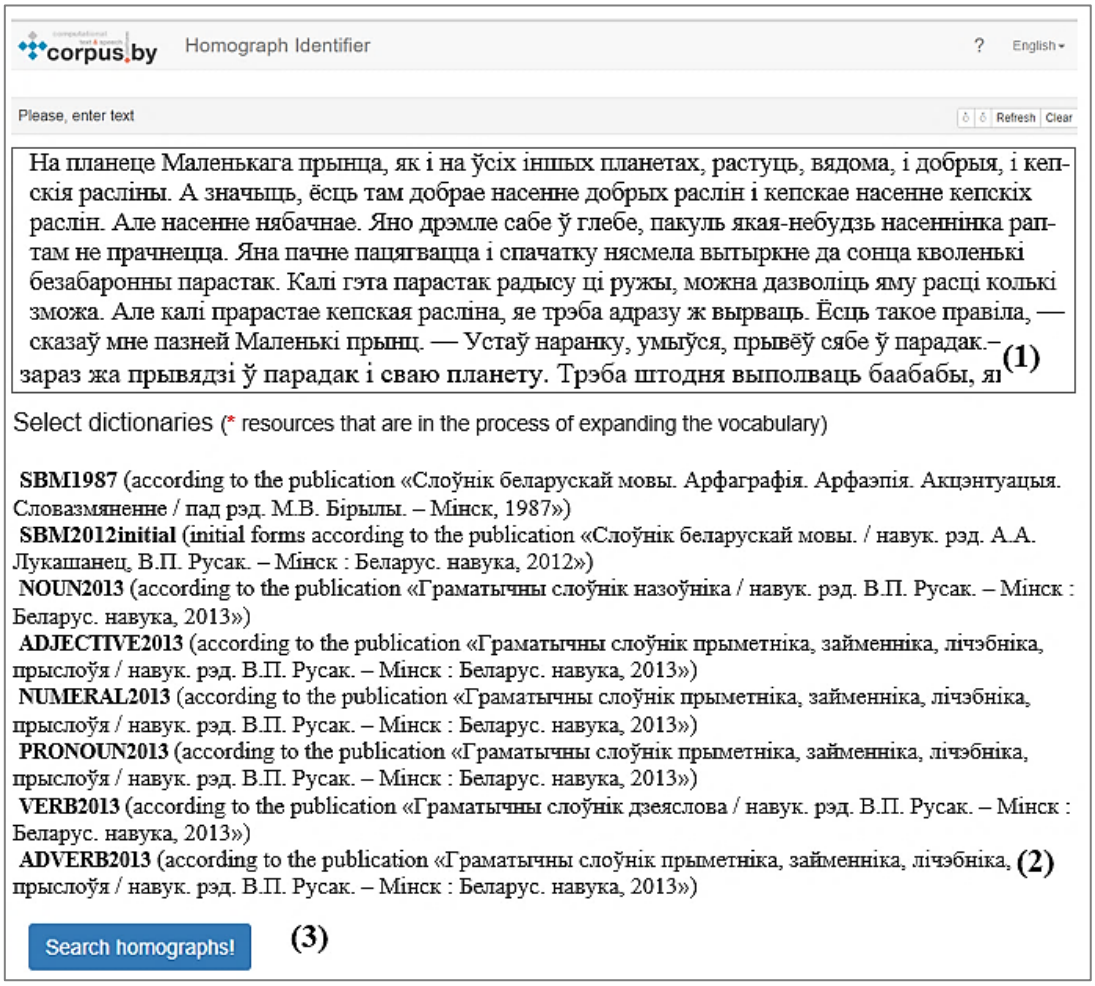
3.4.3. Для элемента масіву *UniqueHomographsArr,* роўнага значэнню *Word,* запісаць у *UniqueHomographsArr Y* элементаў, якія ідуць у *WordsArr* да элемента Word, сам элемент

Word і Y элементаў, якія ідуць у *WordsArr* пасля элемента *Word*, $0 \leqslant Y \leqslant 3$, $Y \in N$, дзе N – мноства натуральных лікаў.

Крок 4. Фарміраванне выдачы. Для кожнага знойдзенага амографа прадставіць варыянты націску, тып амографа, колькасць разоў, якія дадзены амограф сустракае ў зыходным тэксце, кантэксты ўжывання і назву слоўніка, у якім дадзены амограф быў знойдзены.

Канец алгарытму.

Інтэрфейс прататыпа і яго выкарыстанне. Сістэма ў форме вэб-сэрвіса знаходзіцца ў вольным доступе, што зручна для апрабацыі, тэсціравання і нагляднасці працаздольнасці апісанага вышэй алгарытму па спасылцы https://corpus.by/HomographIdentifier/?lang=be. Карыстальніцкі інтэрфейс сэрвіса прадстаўлены на мал. 1, які змяшчае наступныя вобласці: поле ўводу электроннага тэксту (1); поле выбару слоўнікаў (2); кнопка «Шукаць амографы!», якая запускае апрацоўку (3); поле выводу вынікаў, якое з’яўляецца пасля апрацоўкі і мае выгляд табліцы.



Мал. 1. Графічны інтэрфейс сэрвіса «Ідэнтыфікатар амографаў»

Fig. 1. Graphical interface of the "Omograph ID" service

Поле выводу вынікаў адлюстроўвае наступныя даныя: слова-амограф; варыянты націску ў словах-амографах; тып амографа (вызначэнне часціны мовы); колькасць разоў ужывання амографа ў тэксце; кантэксты ўжывання; слоўнік(і), па якім (якіх) быў праведзены пошук амографаў (мал. 2). Таксама сэрвіс прапануе карыстальніку атрымаць амографы ў форме звычайнага спіса. Каб атрымаць такі спіс, патрэбна націснуць на адпаведную спасылку ў полі выводу. Магчымы вынік працы сэрвіса прадстаўлены на мал. 2.

Homograph	Accent variant	Homograph type	Amount	Contexts	Dictionary
значыць	зна́чыць значы́ць	–	1	А значыць, ёсць там	SBM2012INIT
зараз	за́раз зара́з	–	1	– зараз жа прывядзі ў …	SBM2012INIT
яму	я́му (назоўнік) ямў (займеннік)	one part of speech	1	можна дазволіць яму расці колькі зможа. …	SBM1987
распазнаеш	распазнае́ш (дзеясл) распазна́еш (дзеясл)	one part of speech	1	як толькі распазнаеш. Дык вось, на …	SBM1987
сказаў	ска́заў (назоўнік) сказа́ў (дзеяслоў)	different parts of speech	1	… Ёсць такое правіла, - сказаў мне пазней Ма- ленькі …	SBM1987
зараз	за́раз (прыслоўе) зара́з (прыслоўе, назоўнік)	one part of speech	1	— зараз жа прывядзі ў …	SBM1987
часам	ча́сам (прыслоўе, назоўнік) часа́м (назоўнік)	one paradigm	1	… дык гэта спатрэбіц- ца. Часам работу можна адкласці …	SBM1987
павучаць	павуча́ць (дзеясл) павў́чаць (дзеясл)	one part of speech	1	… страшэнна не люб- лю павучаць людзей. Але небяспека …	SBM1987
адкідаю	адкі́даю (дзеясл) адкіда́ю (дзеясл)	one part of speech	1	… што сёння я адкідаю сваю стрыма- насць і …	SBM1987

Мал. 2. Вынік працы сэрвіса «Ідэнтыфікатар амографаў»

Fig. 2. The result of the "Omograph ID" service

Сэрвіс прапаноўвае два сцэнарыя працы: пошук па слоўніках, абраных па змоўчанні, ці пошук па слоўніках, абраных карыстальнікам. У першым сцэнары выкарыстоўваюцца ўсе слоўнікі, убудаваныя ў сістэму, у другім – вызначэнне амографаў адбываецца па асобных слоўніках, абраных карыстальнікам. Далей карыстальнік выбірае той амограф, які падыходзіць па кантэксце згодна сэнсу выказвання.

«Ідэнтыфікатар амографаў» з'яўляецца адным з асноўных сэрвісаў, які выкарыстоўваецца для вычыткі электроннага тэксту праз праграмнае забеспячэнне, распрацаванае супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі. Прапанаваным праграмным забеспячэннем выступаюць сэрвісы апрацоўкі электроннай тэкставай інфармацыі, якія размешчаны на інтэрнэт-платформе www.corpus.by. Выкарыстанне спецыялізаванай методыкі вычыткі электроннага тэксту, якая прадстаўлена на афіцыйным сайце лабараторыі праз спасылку https://ssrlab.by/5406, дазваляе атрымаць арфаграфічна правільны тэкст на беларускай мове. У межах задання «Мадэлі, метады, алгарытмы і праграмныя сродкі інтэлектуальнай апрацоўкі, аналізу і распазнавання медыка-біялагічных даных, малюнкаў, маўленчай і тэкставай інфармацыі і распрацоўка на іх аснове інфармацыйных тэхналогій і сістэм медыцынскага і сацыяльнага прызначэння» сэрвіс «Ідэнтыфікатар амографаў» выкарыстоўваецца для стварэння і кампіляцыі

мадэляў паралельных электронных карпусоў маўлення і тэкстаў медыцынскай, сацыяльнай і прававой тэматыкі на рускай, беларускай і англійскай мовах для распрацоўкі. Акрамя таго, сэрвіс прайшоў апрабацыю падчас складання тэкстаў рэлігійнай тэматыкі [18].

Доступ да сэрвіса праз API. Для доступу да сэрвіса «Ідэнтыфікатар амографаў» праз API неабходна адправіць *AJAX-запыт* тыпу *POST* на адрас https://corpus.by/HomographIdentifier/api.php. Праз масіў *data* перадаюцца наступныя параметры:

text – адвольны ўваходны тэкст;

маркеры выкарыстання слоўнікаў:

sbm1987 – «Слоўнік беларускай мовы. Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне / пад рэд. М. В. Бірылы. – Мінск, 1987»;

sbm2012initial – «Слоўнік беларускай мовы / навук. рэд. А. А. Лукашанец, В. П. Русак. – Мінск : Беларус. навука, 2012»;

noun2013 – назоўнікі паводле кнігі «Граматычны слоўнік назоўніка / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

adjective2013 – прыметнікі паводле кнігі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

numeral2013 – лічэбнікі паводле кнігі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

pronoun2013 – займеннікі паводле кнігі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

verb2013 – дзеясловы паводле кнігі «Граматычны слоўнік дзеяслова / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

adverb2013 – прыслоўі паводле кнігі «Граматычны слоўнік прыметніка, займенніка, лічэбніка, прыслоўя / навук. рэд. В. П. Русак. – Мінск : Беларус. навука, 2013»;

zalizniak – «Грамматический словарь русского языка: Словоизменение / А. А. Зализняк. – М. : Русский язык, 1980. – 880 с.»;

cmu – «Carnegie Mellon University Pronouncing Dictionary»;

uwp_be – беларускія словы, сабраныя сістэмай «Апрацоўка невядомых слоў»;

uwp_ru – рускія словы, сабраныя сістэмай «Апрацоўка невядомых слоў».

Ніжэй прадстаўлены прыклад AJAX-запыту:

```
$.ajax({
type: “POST”,
url: “https://corpus.by/HomographIdentifier/api.php”,
data:{
“text”: “– Маё жыццё такое аднастайнае. Я палюю на курэй, людзі палююць на мяне. Усе куры падобны адна на адну, і ўсе людзі падобны адзін на аднаго. З гэтай прычыны мне і сумнавата. Але, калі ты прыручыш мяне, жыццё маё нібы сонцам азарыцца. Я навучуся распазнаваць твае крокі сярод тысячы іншых. Калі я чую людскія крокі, я ўцякаю і хаваюся. Твае ж паклічуць мяне з нары як музыка. І потым – паглядзі! Бачыш, там, удалечыні, жытняе поле? Я не ем хлеба. Жыта мне ні да чаго. Збажына нічога не напамінае мне. І гэта так сумна! А ў цябе залатыя валасы. І як цудоўна было б, калі б ты прыручыў мяне! Залатое жыта заўсёды было б мне напамінкам пра цябе… Я палюбіў бы песню ветру ў калоссі…”,
“sbm1987”: 1,
“sbm2012initial”: 1
},
success: function(msg){ },
error: function() { }
});
```

Сервер верне JSON-масіў з уваходным тэкстам (параметр text), спісам знойдзеных у тэксце амографаў (параметр result), масівам з дэталямі выніку (параметр resultArr), колькасцю выяўле-

ных амографаў (параметр resultCnt) і спасылкай на спіс выяўленых амографаў (параметр resultUrl). Напрыклад, па вышэй прыведзеным AJAX-запыце будзе сфарміраваны наступны адказ:


```
[
{
“text”: “Груша цвіла апошні год.”,
“result”: “куры
людскія
нары
музыка
музыка”,
“resultArr”: {
“SBM1987”: {
“куры”: {
“accents”: “ку́ры куры́”,
“type”: “адна часціна мовы”,
“count”: 1,
“contexts”: “… на мяне. Усе куры падобны адна на …”
},
“людскія”: {
“accents”: “лю́дскія людскі́я”,
“type”: “адна часціна мовы”,
“count”: 1,
“contexts”: “… Калі я чую людскія крокі, я ўцякаю …”
},
“нары”: {
“accents”: “нары́ на́ры”,“type”: “адна часціна мовы”,
“count”: 1,
“contexts”: “… паклічуць мяне з нары як музыка. І …”
},
“музыка”: {
“accents”: “музы́ка му́зыка”,
“type”: “адна часціна мовы”,
“count”: 1,
“contexts”: “… з нары як музыка. І потым – паглядзі! …”
}
},
“SBM2012initial”: {
“музыка”: {
“accents”: “музы́ка му́зыка”,
“type”: “–”,
“count”: 1,
“contexts”: “… з нары як музыка. І потым – паглядзі! …”
}
}
},
“resultCnt”: “5”,
“resultUrl”: “https://corpus.by/<…>”,
}
]
```


Заключэнне. Прыклад працы прататыпа сістэмы «Ідэнтыфікатар амографаў» дэманструе працаздольнасць распрацаваных алгарытмаў. Карэктнасць працы алгарытму можа быць адлюстравана ў працэсе выкарыстання сістэмы для пошуку амографаў і аналізу вынікаў апрацаванага электроннага тэксту. Сэрвіс будзе карысным для вырашэння ўсіх прыкладных задач, рэалізацыя якіх ускладняецца наяўнасцю амографаў. Такія патрэбы могуць узнікнуць у наступных выпадках: пры правядзенні даследавання пэўнага тэксту на наяўнасць і функцыянаванне ў ім амографаў, што запатрабавана ў корпуснай лінгвістыцы; пры падрыхтоўцы тэксту да апрацоўкі сінтэзатарам маўлення з мэтай атрымання агучанага тэксту. Своечасовае выяўленне амографаў і правільная расстаноўка націскаў дазволяць значна павысіць якасць працы беларускамоўных сістэм сінтэзу маўлення. Немалаважна выкарыстанне праграмы падчас вычыткі тэкстаў вялікага памеру. Сэрвіс дапамагае выявіць амографы для далейшай расстаноўкі націскаў у тых выпадках амаграфіі, ад якіх залежыць сэнс тэксту.

Рэалізацыя метаду, заснаванага на ведах у дадзенай сістэме, дазваляе з упэўненасцю сцвярджаць, што гэта эфектыўны рэсурс для аўтаматычнага пошуку амографаў беларускай, рускай і англійскай моў і іх вызначэння для далейшай апрацоўкі натуральнай мовы камп’ютарнымі сродкамі. Прататып з’яўляецца першым рэсурсам па здыманні шматзначнасці для беларускай мовы ў адкрытым доступе, што дазваляе кожнаму зацікаўленаму выкарыстаць яго для асабістых мэт. У сваю чаргу, для распрацоўшчыкаў пастаяннае выкарыстанне сэрвіса дае магчымасць працягваць бесперапыннае тэсціраванне працаздольнасці алгарытмаў. Паляпшэнне якасці прыкладання можа быць дасягнута дабаўленнем іншых слоўнікаў. Акрамя таго, адкрыты

код робіць магчымым выкарыстоўваць сэрвіс для іншых славянскіх моў шляхам яго адаптацыі да канкрэтнай мовы і дадання новых слоўнікавых рэсурсаў. На далейшым этапе распрацоўкі плануецца дапрацаваць сістэму ідэнтыфікацыі амографаў дабаўленнем дадатковай опцыі пошуку амографаў у тэкстах з пазнакай адрозных часцін мовы пры аднолькавых націсках слоў.

Уклад аўтараў. *Ю. С. Гецэвіч* прапанаваў канцэпцыю аўтаматызаванай экстракцыі амонімаў і выканаў адбор метадаў і алгарытмаў для распрацоўкі прататыпа сэрвіса «Ідэнтыфікатар амографаў». *А. А. Бакуновіч* распрацаваў пакрокавы алгарытм працы сэрвіса і ажыццявіў яго рэалізацыю на API, унёс карэкціроўкі пасля тэсціравання прататыпа. *А. Я. Драгун* падрыхтавала тэкставы корпус літаратурнага дамену з дастатковым аб'ёмам даных, якія змяшчаюць усе тыпы амографаў, для тэсціравання сэрвіса. *Я. С. Зяноўка* і *Д. І. Латышэвіч* правялі дэтальнае тэсціраванне прататыпа з вызначэннем тэхнічных і лінгвістычных памылак апрацоўкі тэкстаў прататыпам. *М. А. Казлова* прадставіла падрабязную інструкцыю выкарыстання «Ідэнтыфікатара амографаў».

Спіс выкарыстаных крыніц

1. Word Sense Disambiguation: Algorithms and Applications / eds.: E. Agirre, P. Edmonds. – Springer, 2007. – Series: Text, Speech and Language Technology. – Vol. 33. – 377 p.

2. Ширшикова, А. А. О проблемах омонимии / А. А. Ширшикова // Альманах современной науки и образования. – Тамбов : Грамота, 2012. – № 2(57). – С. 190–192.

3. Tian, T. Improving web search results for homonyms by suggesting completions from an ontology / T. Tian, J. Geller, S. A. Chun // Current Trends in Web Engineering – 10th Intern. Conf. on Web Engineering, ICWE 2010 Workshops, Vienna, Austria, July 2010. – Vienna, Austria, 2010. – P. 41–44.

4. Van den Beukel, S. Homonym detection for humor recognition in short text / S. van den Beukel, L. Aroyo // Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Belgium, 31 October 2018. – Brussels, Belgium, 2018. – P. 286–291.

5. Pozdniakov, K Regular homophones: a tool for semantic typology and for linguistic reconstruction / K. Pozdniakov, G. Segerer // Africana Linguistica. – 2019. – Vol. 25. – P. 231–279.

6. Roll, U. Using machine learning to disentangle homonyms in large text corpora / U. Roll, R. A. Correia, O. Berger-Tal // Conservation Biology. – June 2018. – Vol. 32, iss. 3. – P. 716–724.

7. Рысаков, С. В. Статистические методы снятия омонимии / С. В. Рысаков, Э. С. Клышинский // Новые информационные технологии в автоматизированных системах. – 2015. – № 18. – С. 555–563.

8. Navigli, R. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation / R. Navigli, P. Velardi // IEEE Transactions on Pattern Analysis and Machine Intelligence. – July 2005. – Vol. 27, iss. 7. – P. 1075–1086.

9. Гатауллин, Р. Р. Аналитический обзор методов разрешения морфологической многозначности / Р. Р. Гатауллин // Электронные библиотеки. – 2016. – Т. 19, № 2. – С. 98–114.

10. Зеленков, Ю. Г. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиции соседних слов / Ю. Г. Зеленков, И. В. Сегайлович, В. А. Титов // Компьютерная лингвистика и интеллектуальные технологии : тр. Междунар. конф. «Диалог-2005», Звенигород, 1–6 июня 2005 г. – М. : Наука, 2005. – С. 616–638.

11. Мухамедшин, Д. Р. Модуль разрешения морфологической неоднозначности: архитектура и организация базы данных / Д. Р. Мухамедшин, Д. Ш. Сулейманов // Программные продукты и системы. – 2020. – Т. 33, № 1. – С. 38–46.

12. Порохнин, А. А. Анализ статистических методов снятия омонимии в текстах на русском языке / А. А. Порохнин // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2013. – № 2. – С. 168–174.

13. Лесько, О. Н. Использование онтологии предметной области для снятия омонимии в естественно-языковых текстах / О. Н. Лесько, Ю. В. Рогушина // Проблеми програмування : науковий журнал. – 2017. – № 2. – С. 61–71.

14. Зинькина, Ю. В. Разрешение функциональной омонимии в русском языке на основе контекстных правил / Ю. В. Зинькина, Н. В. Пяткин, О. А. Невзорова // Компьютерная лингвистика и интеллектуальные технологии : тр. Междунар. конф. «Диалог-2005», Звенигород, 1–6 июня 2005 г. – М. : Наука, 2005. – С. 198–202.

15. Okrut, T. Context-sensitive homograph disambiguation with NooJ in Belarusian and Russian electronic texts / T. Okrut, B. Lobanov, Y. Yakubovich // Intern. Scientific Conf. on the Automatic Processing of Natural-Language Electronic Texts “NooJ’2015”, Minsk, Belarus, 11–13 June 2015. – Minsk : UIIP NASB, 2015. – P. 48.

16. Камп’ютарна-лінгвістычныя сэрвісы www.corpus.by для аўтаматычнай апрацоўкі тэкстаў / Я. С. Качан [і інш.] // Нацыянальна-культурны кампанент у літаратурнай і дыялектнай мове : зб. навук. арт. – Брэст : БрДУ імя А. С. Пушкіна, 2016. – С. 93–104.

17. The problem of automatic search and determonation of homonyms for the Belarusian and Russian languages / Ya. Zianouka [et al.] // Информационные технологии в промышленности, логистике и социальной сфере. – Минск : Объединенный институт проблем информатики НАН Беларуси, 2021. – С. 182–184.

18. Новы Запавет – Кніга Прыповесьцяў : пер. А. Бокуна. – Мінск : Пазітыў-цэнтр, 2016. – 511 с.

References

1. Agirre E., Edmonds P. (eds.). *Word Sense Disambiguation: Algorithms and Applications.* Springer, 2007, Series: Text, Speech and Language Technology, vol. 33, 377 p.

2. Shirshikova A. *On the problems of homonymy*. Al'manakh sovremennoy nauki i obrazovaniya [*Almanac of Modern Science and Education*], Tambov, Gramota, 2012, no. 2(57), pp. 190–192 (In Russ.).

3. Tian T., Geller J., Chun S. A. Improving web search results for homonyms by suggesting completions from an ontology. *Current Trends in Web Engineering: 10th International Conference on Web Engineering, ICWE 2010 Workshops, July 2010, Vienna, Austria, July 2010*. Vienna, Austria, 2010, pp. 41–44.

4. Van den Beukel S., Aroyo L. Homonym detection for humor recognition in short text. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Brussels, Belgium, 31 October 2018.* Brussels, Belgium, 2018, pp. 286–291.

5. Pozdniakov K., Segerer G. Regular homophones: a tool for semantic typology and for linguistic reconstruction. Africana Linguistica, 2019, vol. 25, pp. 231–279.

6. Roll U., Correia R. A., Berger-Tal O. Using machine learning to disentangle homonyms in large text corpora. *Conservation Biology*, June 2018, vol. 32, iss. 3, pp. 716–724.

7. Rysakov S. V., Klyshinsky E. S. *Statistical methods of homonymy removal.* Novye informacionnye tehnologii v avtomatizirovannyh sistemah [*New Information Technologies in Automated Systems*], 2015, no. 18, pp. 555–563 (In Russ.).

8. Navigli R., Velardi P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2005, vol. 27, no. 7, pp. 1075–1086.

9. Gataullin R. R. *Analytical review of methods for resolving morphological polysemy*. Elektronnyye biblioteki [*Electronic Libraries*], 2016, vol. 19, no. 2, pp. 98–114 (In Russ).

10. Zelenkov Yu. G., Segalovich I. V., Titov V. A. *Probabilistic model for removing morphological homonymy based on normalizing substitutions and positions of neighboring words*. Komp'juternaja lingvistika i intellektual'nye tehnologii : trudy Mezhdunarodnoj konferencii «Dialog-2005», Zvenigorod, 1–6 ijunja 2005 g. [*Computer linguistics and intellectual technologies: proceedings of the international conference "Dialogue-2005", Zvenigorod, 1–6 June 2005*], Moscow, Nauka, 2005, pp. 616–638 (In Russ.).

11. Mukhamedshin D. R., Suleymanov D. Sh. *Module of morphological ambiguity resolution: database architecture and organization*. Programmnyye produkty i sistemy [*Software Products and Systems*], 2020, vol. 33, no. 1, pp. 38–46 (In Russ.).

12. Porokhnin A. A. *Analysis of statistical methods for removing homonymy in Russian texts*. Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Serija: Upravlenie, vychislitel'naja tehnika i informatika [*Bulletin of the Astrakhan State Technical University. Series: Management, Computing and Information Science*], 2013, no. 2, pp. 168–174.

13. Les'ko O. N., Rogushina Yu. V. *Using the domain ontology for removing homonymy in natural language texts*. Problemi programuvannya [*Programming Problems*], 2017, no. 2, pp. 61–71 (In Russ.).

14. Zin'kina Yu. V., Pyatkin N. V., Nevzorova O. A. *Resolution of functional homonymy in Russian based on contextual rules*. Komp'juternaja lingvistika i intellektual'nye tehnologii : trudy Mezhdunarodnoj konferencii «Dialog-2005», Zvenigorod, 1–6 ijunja 2005 g. [*Computer linguistics and intellectual technologies: proceedings of the international conference "Dialogue-2005", Zvenigorod, 1–6 June 2005*], Moscow, Nauka, 2005, pp. 198–202 (In Russ.).

15. Okrut T., Lobanov B., Yakubovich Y. Context-sensitive homograph disambiguation with NooJ in Belarusian and Russian electronic texts. *International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015", Minsk, Belarus, 11–13 June 2015*. UIIP NASB, 2015, p. 48.

16. Hiecevic Ju., Kacan Ya, Lysy S., Stanislavienka H., Hiuntar A. *Computer-linguistic services www.corpus.by for automatic text processing*. Nacyjanalna-kulturny kampanient u litaraturnaj i dyjaliektnaj movie : zbornik navukovyh artykulaŭ [*National-cultural Component in Literary and Dialect Language : Collection of Scientific Articles*], Brest, Brjescki dzjarzhaŭny ŭniversitjet imja A. S. Pushkina, 2016, pp. 93–104 (In Bel).

17. Zianouka Ya., Hetsevish Yu., Majeŭski S., Dzienisiuk Dz. *The problem of automatic search and determonation of homonyms for the Belarusian and Russian languages*. Informacionnye tehnologii v promyshlennosti, logistike i socialnoj sfere [*Information Technologies in Industry, Logistics and Social Sphere*], Minsk, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, 2021, pp. 182–184.

18. Novy Zapavet – Kniga Prypoves'cjaŭ. *The New Testament – The Book of Proverbs*. Transl. A. Bokuna. Minsk, Pazityŭ-cjentr, 2016, 511 p. (In Bel).

Інфармацыя пра аўтараў

Гецэвіч Юрась Станіслававіч, кандыдат тэхнічных навук, дацэнт, загадчык лабараторыі распазнавання і сінтэзу маўлення, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: yuras.hetsevich@gmail.com

Зяноўка Яўгенія Сяргееўна, малодшы навуковы супрацоўнік, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: evgeniakacan@gmail.com

Латышэвіч Давід Іосіфавіч, стажор малодшага навуковага супрацоўніка, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: david.latyshevich@gmail.com

Бакуновіч Андрэй Аляксеевіч, малодшы навуковы супрацоўнік, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: bakunovich.andrei.work@gmail.com

Драгун Анастасія Яўгеньеўна, малодшы навуковы супрацоўнік, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: ndrahun@gmail.com

Казлова Маргарыта Аляксандраўна, стажор малодшага навуковага супрацоўніка, Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі.
E-mail: margaryta.kazlova@gmail.com

Information about the authors

Yuras S. Hetsevich, Ph. D. (Eng.), Assoc. Prof., Head of the Speech Synthesis and Recognition Laboratory, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: yuras.hetsevich@gmail.com

Yauheniya S. Zianouka, Junior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: evgeniakacan@gmail.com

David I. Latyshevich, Trainee of Junior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: david.latyshevich@gmail.com

Andrey A. Bakunovich, Junior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: bakunovich.andrei.work@gmail.com

Anastasia Ya. Drahun, Junior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: ndrahun@gmail.com

Margarita A. Kazlova, Trainee of Junior Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: margaryta.kazlova@gmail.com