КОРПУСНАЯ ЛІНГВІСТЫКА ЯК АСАБЛІВЫ РАЗДЗЕЛ КАМП'ЮТАРНАЙ ЛІНГВІСТЫКІ

Горб В. А., Радзюк К.А.

Беларускі дзяржаўны ўніверсітэт інфарматыкі і радыёэлектронікі
г. Мінск, Рэспубліка Беларусь

Пятрова Н. Я. – к. філал. н., дацэнт

У рабоце разглядаюцца асаблівасці і задачы корпуснай лінгвістыкі. Вызначаецца, што такое корпус, камп'ютарная лінгвістыка. Разглядаюцца праграмы для работы з корпусам, іх недахопы. Таксама разглядаецца роля корпуснай лінгвістыкі на Беларусі. Вызначаецца практычная каштоўнасць корпуснай лінгвістыкі.

Інфармацыйныя тэхналогіі маюць у цяперашні час ключавую ролю ў працэсах атрымання і назапашвання новых ведаў. Пры гэтым, на змену традыцыйным метадам інфармацыйнай падтрымкі навуковых даследаванняў шляхам назапашвання, класіфікацыі і распаўсюджвання навукова-тэхнічнай інфармацыі прыходзяць новыя метады, заснаваныя на выкарыстанні магчымасцяў інфармацыйнай падтрымкі фундаментальнай і прыкладной навукі, якія прадастаўляюць сучасныя інфармацыйныя тэхналогіі.

Камп'ютарная лінгвістыка – міждысцыплінарная галіна, якая тычыцца заснаванага на правілах апрацоўкі статыстычнага мадэлявання, а таксама вывучэння адпаведных камп'ютарных падыходаў да моўных пытанняў [1].

Традыцыйна, камп'ютарная лінгвістыка вызначалася камп'ютарнымі навукоўцамі, якія спецыялізуюцца ў галіне прымянення ЭВМ у апрацоўцы натуральнай мовы. Сёння такія лінгвісты часта працуюць у якасці членаў міждысцыплінарнай каманды, якая можа ўключаць у сябе звычайных лінгвістаў, спецыялістаў па тэгавай мове, а таксама інжынераў. У цэлым, камп'ютарная лінгвістыка абапіраецца на дапамогу лінгвістаў, камп'ютарных навукоўцаў, спецыялістаў у галіне штучнага інтэлекту, матэматыкаў, логікаў, філосафаў, кагнітыўных навукоўцаў, кагнітыўных псіхолагаў, псіхалінгвістаў, антраполагаў і нейрабіёлагаў.

Моваznаўства або лінгвістыка – навука аб будове, функцыянаванні і развіцці моваў свету. Гэта навука аб натуральнай чалавечай мове наогул і аб усіх мовах свету як індывідуальных яе прадстаўніках. У шырокім сэнсе слова, лінгвістыка падзяляецца на навуковую і практычную. Часцей за ўсё пад лінгвістыкай маецца на ўвазе менавіта навуковая лінгвістыка. Яна з'яўляецца часткай семіётыкі як навукі аб знаках [2].

Корпусная лінгвістыка – раздзел камп'ютарнай лінгвістыкі, які займаецца распрацоўкай агульных прынцыпаў будовы і выкарыстання лінгвістычных корпусаў (корпусаў тэкстаў) з выкарыстаннем камп'ютарных тэхналогій. Корпусная лінгвістыка дала магчымасць удакладніць вынікі праведзеных раней даследаванняў мовы і правесці новы, больш шырокі і сістэмны па аб'ёме моўнага матэрыялу лінгвістычны аналіз. У цэнтры ўвагі корпуснай лінгвістыкі – моўная асоба з маўленчай дзейнасцю, масавай камунікацыяй, праблемамі яе апісання. Галоўныя мэты – лінгвістычнае апісанне моўнай сістэмы, а таксама адлюстраванне моўнага матэрыялу ў выглядзе корпуса тэкстаў, які ў сваю чаргу можа выкарыстоўвацца іншымі лінгвістычнымі дысцыплінамі [3].

Тэрмін «корпус» звычайна абазначае збор тэкстаў канечнага фіксаванага памеру. З цягам часу аб'ём і склад корпуса можа змяняцца, аднак гэтыя змены не павінны ўплываць на яго рэпрэзентатыўнасць. Аб'ём першых корпусаў складаў каля 1 млн. словаўжыванняў. Зараз лічыцца, што аб'ём агульнамоўнага корпуса павінен быць не менш за 100 млн. словаўжыванняў.

Задача стваральнікаў корпуса – сабраць як мага большую колькасць тэкстаў, але галоўнае не толькі і не столькі ў колькасці матэрыялу, колькі ў яго прапарцыянальнасці. Можна сказаць, што корпус – гэта паменшаная мадэль мовы ці падмовы. Адным з важнейшых паняццяў корпуснай лінгвістыкі з'яўляецца рэпрэзентатыўнасць. Пад рэпрэзентатыўнасцю разумеюць неабходна-дастатковае і прапарцыянальнае прадстаўленне ў корпусе тэкстаў розных перыядаў, жанраў, стыляў, аўтараў і інш. [4].

Для рашэння розных лінгвістычных задач аднаго масіва тэкстаў мала. Неабходна, каб тэксты змяшчалі рознага роду дадатковую лінгвістычную і экстралінгвістычную інфармацыю. Так у корпуснай лінгвістыцы ўзнікла ідэя размечанага корпуса. Разметка (tagging, annotation) заключаецца ў прыпісванні тэкстам і іх кампанентам спецыяльных метак (tag, tags): вонкавых, экстралінгвістычных (звесткі аб аўтары і звесткі аб тэксце: аўтар, назва, год і месца выдання, жанр, тэматыка; звесткі аб аўтары могуць уключаць не толькі яго імя, але таксама ўзрост, пол, гады жыцця і інш., гэта кадзіраванне інфармацыі мае назву метаразметка), структурных (глава, абзац, сказ, словаформа) і ўласна лінгвістычных, якія апісваюць лексічныя, граматычныя і іншыя характарыстыкі элементаў тэксту. Сярод лінгвістычных тыпаў разметкі выдзяляюцца: марфалагічная разметка, сінтаксічная разметка, семантычная разметка, анафарычная разметка, прасадычная разметка і іншыя.

Сярод асноўных задач выдзяляюць такія, як збор тэкстаў з пэўнай мэтай, машынная апрацоўка тэкстаў, дапамога ў стварэнні слоўнікаў (лексікаграфічныя падтрымка), складанне канкардансаў (спіс, што сустракаюцца ў тэксце словаформ, размешчаны ў алфавітным парадку). У супрацьлегласць слоўніку, слова падаецца з яго так званым навакольным асяроддзем. Таксама задачамі з'яўляюцца наступныя: складанне частотных слоўнікаў, стварэнне нацыянальных корпусаў, даследаванне выкарыстання натуральнай мовы ў розных рэгістрах, праверка лінгвістычных тэорый.

Корпусны менеджар – гэта своеасаблівая аперацыйная абалонка лінгвістычнага корпуса, якая ўяўляе сабой цэлы рад магчымасцей для даследавання мовы. Пры наяўнасці адпаведнай разметкі ажыццяўляецца пошук па наборы марфалагічных прыкмет (напрыклад, пошук усіх словазлучэнняў выгляду 'прыназоўнік па + назоўнік у месным склоне') і інш. Інфармацыю, якая адпавядае узроўням лінгвістычнай разметкі, прадстаўленай у корпусе. Дзякуючы наяўнасці метаразметкі карыстальнік мае магчымасць ствараць свой падкорпус тэкстаў, адабраных па жанры, тэматыцы, часу напісання і і нш. Вынік выдачы ўяўляе сабой канкарданс (мноства кантэкстаў, у якім сустрэўся запрошаны моўны выраз). Кожны з прыкладаў забяспечваецца інфармацыяй аб крыніцы, адкуль узяты прыклад. У шэрагу корпусаў магчыма таксама атрымаць статыстычную інфармацыю аб запрошаным моўным выразе: яго адносную частату па ўсім корпусе, размеркаванне па жанрах або часовым зрэзах, інфармацыю аб частаце яго спалучальнасці [5].

Канкардансер – гэта спецыяльная праграма, якая дазваляе аналізаваць вялікія масівы тэкстаў на прадмет пошуку заканамернасцей выкарыстання слоў і выразаў у мове. Канкандансер ажыццяўляе пошук зададзенага слова ў корпусе і выдае ў новым акне некалькі фрагментаў сказаў з розных тэкстаў, у якіх выкарыстоўваецца дадзенае слова ці выраз. Аднак такія праграмы таксама робяць памылкі. Напрыклад, могуць сустракацца недакладнасці пры лексіка-граматычным пошуку [6].

Сярод спецыяльных праграм для апрацоўкі мовы асабліваемесца займаюць праграмы аўтаматычнай разметкі. Разметка корпусаў уяўляе сабой працаёмістую аперацыю, асабліва ўлічваючы вялікія памеры сучасных корпусаў тэкстаў. Для некаторых відаў разметкі існуюць розныя праграмы, але ёсць і такія віды разметкі, дзе асноўная частка працы праводзіцца ўручную. Разметка (tagging, annotation) заключаецца ў прыпісванні тэкстам і іх кампанентам спецыяльных метак (тэгаў). Тэгі дзеляцца на ўласна лінгвістычныя, якія апісваюць лексічныя, граматычныя і іншыя характарыстыкі элементаў тэксту, а таксама знешнія, экстралінгвістычныя (звесткі аб аўтары і звесткі аб тэксце) [7].

Такім чынам, корпусная лінгвістыка адыгрывае важную ролю ў апрацоўцы натуральнай мовы і з'яўляецца значным рэсурсам для розных тыпаў адукацыйных праграм, праграм машыннага перакладу, спатрэбіцца для правядзення лінгвістычных даследаванняў у галіне лексікаграфіі, а таксама для распрацоўкі тэрміналагічнай базы беларускай мовы.

Спіс выкарыстаных крыніц:

1. Баранов, А. Н. Компьютерная лингвистика [Электронны рэсурс] / Баранов А. Н. – Рэжым доступу: https://bigenc.ru/linguistics/text/2087783. – Дата доступу: 01.04.2020.

2. Крылов, С. А. ,Лингвистика [Электронны рэсурс] / Крылов С. А. – Рэжым доступу: https://www.krugosvet.ru/enc/kultura_i_obrazovanie/literatura/LINGVISTIKA_YAZIKOZNANIE_YAZIKOVEDENIE.html. – Дата доступу: 01.04.2020.

3. Городецкий, Б.Ю. Компьютерная лингвистика и моделирование языкового общения [Электронны рэсурс] / Б.Ю. Городецкий – Рэжым доступу: https://classes.ru/grammar/165.new-in-linguistics-24/source/worddocuments/_.htm. – Дата доступу: 01.04.2020.

4. Марчук, Ю.Н. Основы компьютерной лингвистики [Электронны рэсурс] / Ю.Н. Марчук. – Рэжым доступу: https://obuchalka.org/20191023114865/komputernaya-lingvistika-marchuk-u-n-2007.html. – Дата доступу: 01.04.2020.

5. Барковіч, А. А. Корпусная лінгвістыка і інтэрнэт [Электронны рэсурс] / Барковіч А. А. – Рэжым доступу: http://elib.bsu.by/bitstream/123456789/166481/1/33-36.pdf. – Дата доступу: 01.04.2020.

6. Коваль, В.И. Компьютерная лингвистика: научное направление и учебная дисциплина [Электронны рэсурс] / В. И. Коваль. – Рэжым доступу: http://lab314.brsu.by/kmp-lite/kmp-video/CL/CL-Gomel%202010.pdf. – Дата доступу: 01.04.2020.

7. Захаров, В.П. Корпусная лингвистика [Электронны рэсурс] / В.П. Захаров. – Рэжым доступу: http://lab314.brsu.by/kmp-lite/kmp-video/CL/%D0%97%D0%B0%D1%85%D0%B0%D1%80%D0%BE%D0%B2-%D0%9A%D0%9B.pdf. – Дата доступу: 01.04.2020.