ЖАҲОН КОРПУСЛАРИДА ГЕОГРАФИК ТЕРМИНЛАРНИНГ БЕРИЛИШИ МУАММОСИГА ДОИР

Икром Хушбоқович ИСЛОМОВ
филология фанлари бўйича (PhD) фалсафа доктори
мустақил изланувчи
Қарши давлат университети
Қарши, Ўзбекистон
ikrom75islom@gmail.com

Аннотация

Мақолада тил корпуслари, уларнинг семантик разметкаси ва уларни яратишга доир тадқиқотлар таҳлили, шунингдек, лексиканинг чегараланган бирликлари бўлган терминларнинг берилиши муаммоси географик терминлар мисолида тадқиқ қилинган. Корпусда терминологик бирликларни семантик теглаш тамойиллари ва вазифалари, уларнинг аҳамияти юзасидан тавсия ва хулосалар мақолада ўз аксини топган.

Таянч сўзлар: корпус, семантик разметка, семантик тег, қидирув имконияти, оператор ва константа теглар, географик термин, иерархик муносабат, тасниф, таксономия.

О ПРОБЛЕМЕ ИСПОЛЬЗОВАНИЯ ГЕОГРАФИЧЕСКИХ ТЕРМИНОВ В МИРОВЫХ КОРПУСАХ

Икром Хушбокович ИСЛОМОВ
доктор философии по филологическим (PhD) наукам
независмый исследователь
Каршинский государственный университет
Карши, Узбекистан
ikrom75islom@gmail.com

Аннотация

В статье анализируются языковые корпусы, их семантическая разметка и исследования по их созданию, а также на примере географических терминов рассматривается проблема подачи терминов, являющихся ограниченной единицей лексики. В статье отражены принципы семантического тегирования терминологических единиц и их задачи, рекомендации и вьводы по их значению.

Ключевые слова: корпус, семантическая разметка, семантический тег, возможности поиска, оператор и константы тега, географический термин, иерархическое отношение, характеристика, таксономия.

Жаҳон корпус лингвистикасида тил бирликлари, хусусан, лексемаларга

автоматик семантик ишлов бериш (маъносини аниқлаш, кўп маъноли сўзлар маъносини ажратиш, гипо/гиперонимик муносабати, бутун/бўлак муносабатини фарқлаш) масаласи бугун ҳам ўз ечимини кутаётган муаммолардан саналади. Ҳозиргача санокли тил корпусларида қидирувнинг семантик тури йўлга қўйилган. Зеро, компьютер дастурлари, асосан, тил бирликларини шаклига қараб таҳлил қила олади, маънони таҳлил қилишга қисман эришилиши мумкин, аммо тўла семантик таҳлилга эришиш қийин. Шундай бўлса-да, жаҳон корпус лингвистикасида бу муаммонинг ечимини топишга қаратилган ишлар бажариляпти: тезаурус, автоматик аннотациялаш, WordNet, концепт асосида сўз маъносини аниқлаш амалиётини шундай ишлар сирасига киритиш мумкин.

Е.В.Рахилина, Г.И.Кустова, О.Н.Ляшевская, Т.И.Резникова, О.Ю.Шеманаевалар нинг "Задачи и принципы семантической разметки лексики в НКРЯ" номли мақоласида [7] Рус тили миллий корпусида лексик бирликларни семантик теглаш тамойиллари ва вазифалари тавсифланган бўлса, В.В.Куканова "Принципы семантической разметки национального корпуса калмыцкого языка" номли мақоласида [4] қалмиқ тили миллий корпусининг семантик разметкаси масалаларини таҳлил этган. Е.В.Бирюльцев, А.М.Елизаров, Н.Г.Жильцов, В.В.Иванов, О.А.Невзорова, В.Д.Соловьевлар математик терминлар мавжуд ҳужжатларда семантик қидирувни йўлга қўйиш муаммоларига жавоб излашган [1; 296-300].

М.Ю.Загорулько, И.С.Кононенко, Е.А.Сидоровалар "Система семантической разметки корпуса текстов в ограниченной предметной области" номли мақоласида чегараланган лексикани семантик разметкалаш масаласини таҳлил қилишади [2]. Олимларнинг фикрича, терминологик теглаш нафақат матнда чегараланган лексиканинг мавжудлиги ва статистикасини, балки маълум бир тилда кенг тарқалган умумистеъмол лексикадан фойдаланиш хусусиятларини ҳам аниқлайди. Юқорида санаб ўтилган мутахассислар томонидан терминологик бирликларни белгилашнинг қуйидаги тамойиллари таклиф этилади: 1) бирликнинг тури ва иерархик муносабатда бўладиган ички гуруҳларига мос келадиган хусусиятларни аниқлаш; 2) айнан терминологик

бирлик мазмуни билан боғлиқ матн парчасини ажратиб олиш

Таъкидланишича, терминологик хусусиятга эга бирликлар семантик жиҳатдан иерархик гуруҳига қараб ажратилади. Бундай тасниф корпус лингвистикасида таксономия деб ҳам юритилади. Шундан келиб чиқиб, ўзбек тили изоҳли луғатида географик термин сифатида изоҳланган бирликларни ҳам гуруҳларга ажратиш мақсадга мувофиқ.

Бу гуруҳларнинг ҳар бири иерархик муносабат асосида маълум бирликларни ўз ичига олади. Лингвистик таъминот базасида уларнинг барчасига шу гуруҳга мансублик белгисини билдирувчи тег бириктирилади.

Юқорида кўрсатилган тадқиқотлар натижасида Рус тили миллий корпусида семантик разметкани амалга оширишда қуйидаги ёндашувни кўриш мумкин.

Рус тили миллий корпуси (НКРЯ)нинг семантик тег(разметка)лаш принципларини кузатишларимиз шуни кўрсатдики, бу корпусда бир сўз шаклга уч турдаги белгилар бириктирилади: 1) разряд (ном ёки унга ишора сўз: предмет, шахс ва ҳ.); 2) лексик-семантик характеристика (лексема тегишли бўлган ЛСГ ёки семантик майдон, каузативлик ва ҳ.); 3) деривацион тавсиф [10].

Г.И.Кустова, Е.В.Падучевалар “Словарь как лексическая база данных” сарлавҳали мақоласида лексик-семантик тегларни қуйидаги майдон бўйича гуруҳлашади: 1) таксономия (лексема мансуб ЛСГ) – от, сифат, феъл, равиш сўз туркумлари учун тегишли тег; 2) мереология («бутун-бўлак»ка ишора қилувчи белги, «элемент-тўда/гуруҳ») – предмет ва нопредмет атов бирликларга тегишли тег; 3) топология (ифодаланаётган объектнинг топологик мавқеи) – нарса отларига тегишли тег; 4) каузация – феълларга тегишли тег; 5) баҳо – предмет ва нопредмет атов бирликлари, сифат ва равишга тегишли тег [6; 27-31].

Рус тили миллий корпуси семантик разметкаси муаллифлари “Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы” номли мақолада [5] миллий корпус учун

лексикани семантик теглаш масаласини ёритар эканлар, семантик теглардан тортиб, оператор ва константа теглар ишлаб чиқиш, семантик тегланган (аннотацияланган) лексемалар базасини шакллантириш муаммоларига ечим излайди.

Лексикограф базасида бўлгани сингари корпусда ҳам ҳар бир сўз туркуми учун теглар мажмуи ўзига хос бўлади. Танлаб олинган семантик изоҳлардан қуйидагиларни келтириш мумкин: 1) феъл туркуми учун: ҳаракат, физик таъсир, яратиш, йўқ қилиш, эгалик қилиш, ҳис-ҳис-туйғу, нутқ, инсон хулқ-атвори; 2) сифат туркуми учун: ўлчов, шакл, ранг, ҳарорат, маза-таъм, ҳид, макон, замон, инсон хусусиятлари; 3) нопредмет отлар учун: уларнинг кўпчилиги феъл ва сифат кесишувида ҳосил бўлганлиги сабабли шу туркумга хос бўлган белгилар (ҳаракат, физик таъсир, тузиш, йўқ қилиш, эгалик қилиш, ҳис-туйғу, нутқ, макон, замон, белги-хусусият, ранг, ҳарорат, маза-таъм ва ҳ.), шу билан бирга уларнинг махсус гуруҳлари: тадбир, касаллик, спорт, ўйин, ўлчов бирликлари; 4) предметни атовчи отлар учун: шахс, ҳайвон, ўсимлик, модда ва материал, бино, иншоот, асбоб-ускуна, транспорт воситаси ва ҳ.к. [8]

Э.Г.Шимчук рус лексикографиясига бағишланган қўлланмасида [9] соҳа луғатларини терминологик разметкаланган матн фрагментлари воситасида тўлдириш бир неча босқичлардан иборат бўлиши таъкидланади:

1.Семантик белгилар иерархиясини луғатда мавжуд семаларга мувофиқ ҳолда киритиш

2.Фрагментларга луғат тузишнинг морфологик ва синтактик компонент технологияси асосида ишлов бериш

3.Атамаларни унификация қилиш, нормаллаштириш

4.Теглаш хусусиятларига мувофиқ семантик хусусиятларга эга бўлган атамалар билан таъминлаш

Терминологик луғатни ишлаб чиқиш жараёнида техник муаммолар ҳам туғилади, уларнинг ечими алоҳида эътиборни талаб қилади.

1.Морфологик ва лексик омонимия. Мутахассис морфологик ва синтактик теглашни амалга оширмаганлиги сабабли атамага шаклан мос келадиган барча

омонимлар автоматик равишда луғатга қўшилиб қолади ва семантик маънолари фарқланмай қолади.

2.Сўз бирикма таркибига кирган умумистеъмол лексика (мураккаб таркибли атамалар). Бундай ҳолда, луғат умумистеъмол сўзларни ҳам қамраб олади ва уларнинг семаларига нотерминлик пометаси қўйилади.

3.Рус тилининг луғатида мавжуд бўлмаган, умумлексикага мансуб бўлмаган сўз. Бундай номаълум сўзлар бир таркибли атама сифатида пайдо бўлади ёки таркибли атаманинг таркибида учрайди. Семантик разметкалаш жараёнида бу ҳолат ҳам ноаниқликни келтириб чиқариши мумкин.

4.Сўзлар орасидаги пробел синтактик бирликдек қабул қилиниши мумкин, чунки дастурдаги синтактик шаблонга мос келиб қолиш ҳоллари учрайди. Бундай бирликлар ҳам махсус ишлов беришга эҳтиёж сезади.

5.Луғат бирлиги ҳисобланмайдиган (нолексик) бирликлар ҳам атама таркибига кириб қолиши мумкин. Бундай ҳолатда бошқа бирликлар билан семантик, синтактик алоқа ҳосил қилади.

Шундай қилиб, Э.Шимнукнинг тажрибалари шуни кўрсатадики, танланган лексикографик манбалардан маълумотни автоматик экспорт қилиш усули билан терминологик базани тўлдириш усули ўзини тўла оқламайди.

Биринчидан, моделлар билан бошқариш усулида ҳамма атамаларни ҳам модел асосида ажратиб олиш тўғри натижа беравермайди. Чунки феълнинг ҳар бир маъноси учун бошқарув моделлари рўйхати қанчалик тўлиқ бўлмасин, ҳақиқий корпусда ушбу моделлар қамраб олмаган мисолларнинг миқдори юқорилигича қолаверади. Нотўлиқ бошқарув моделларини такомиллаштириш қўшимча синтактик модулни киритиш ҳамда корпусдан ушбу бирликларни тадқиқ этиши талаб қилинади.

Юқоридагилардан келиб чиққан ҳолда айтиш мумкинки, терминологик маълумотлар базасини яратиш учун маълумотлар омбори ишлаб чиқилиши лозим.

Рус тили миллий корпуси семантик разметка тизими учун нисбатан содда таксономик гуруҳлар ишлаб чиқиш принципига амал қилинган. Биринчидан,

иерархик бўлмаган, яъни тўғридан-тўғри қидирув фойдаланувчига қулайлик яратади. Иккинчидан, семантик майдон ва гуруҳлар битта ойнада кўриниб турса, фойдаланувчи уларда қидирув сўровини бериш бўйича тезроқ қарор қабул қилади ва семантик майдонни танлашда қийинчилик сезмайди [3].

Қуйида бир неча тил корпусида семантик белгилар асосидаги қидирувда географик терминларни қидириш имкониятини кузатамиз ва таҳлил қиламиз. Бунинг учун Рус тили миллий корпусининг [10] семантик қидирув имкониятидан фойдаланиш мумкин. Бу корпус семантик қидирув ойнасида географик терминларни билдирувчи ёки шундай сўзларнинг семантик майдони/лексик-семантик гуруҳини ифодаловчи махсус параметр мавжуд эмас. Таксономия қисмида "пространство и место" деган белги мавжуд: бу белги асосида сўров берилганда, қуйидаги натижани кўриш мумкин: **S, r:concr & t:space** теги, яъни макон/ўрин-жой лексик-семантик гуруҳга мансуб сўзлар қидируви сўрови берилганда, 119 608 та метариал, 8 399 547 та сўз шакл топилган.

Эътибор қаратсак, сўз ҳақида берилган маълумотлар сирасида *(см в словаре)* гиперҳаволаси кўринади: бу ҳаволага мурожаат қилинганда маълум бир сўзнинг луғатлардаги [12] изоҳи очилади.

Юқоридаги тавсифлардан хулоса қилиш мумкинки, семантик разметкаланган Рус тили миллий корпусида корпус ичида туриб географик термин ҳақида тўла маълумот олиш имкони мавжуд эмас, ўрин-жой лексик-семантик гуруҳига мансуб сўзлар ажратиб кўрсатилган, гипер ҳавола орқали луғатлардан унинг изоҳини ўқиш мумкин бўлади. Мисолларда ажратиб кўрсатилган сўзларни таҳлил қилиш шуни кўрсатадики, бу сўзлар умумий ҳолатда ўрин-жой мазмунини билдиради. Масалан: *страна, граница, штат, область, линия, город, край* [13]. Бу бирликлар умумистеъмол сўз мавқеида бўлади, терминологик маъноси махсус белги ёки помета билан ажратилмайди. Корпусда берилган *море, озеро, лес* каби географик жойни ифодаловчи лексик бирликлар географик ҳодиса маъносини билдиради, табиийки, бундай сўзларга махсус белгилар қўйиш ва шу тег/белги ёрдамида унинг қидирувини ташкил

қилиш шундай сўзлар устида турли амалларни бажаришга ёрдам беради. Шундай белгиси бўлмаганлиги сабабли Рус тили миллий корпуси семантик разметкаси ҳақида сўзнинг терминологик маъноси махсус тегланмаган (аннотацияланмаган) корпус деб хулосалаш мумкин. Кузатишларимиз шуни кўрсатадики, сўз изоҳи ҳақида маълумот олишнинг (ҳозирча) корпус бирлигини электрон луғатга гиперҳавола орқали бириктириш имкониятигина мавжуд.

Туркий тилларда мавжуд корпуслар: олтой, бошқирд, қозоқ, татар, қрим-татар, тува, турк, ўзбек, хакас, шор, ёқут тили корпуси фақат морфологик белги асосидаги қидирув имкониятига эга: улар семантик жиҳатдан аннотацияланмаган.

Юқорида келтирилган назарий материаллар ҳамда корпус қидирув имкониятини таҳлил қилиш асосида қуйидаги хулосаларга келиш мумкин:

Биринчидан, тил корпуслари, асосан, матнлар массивидан иборат бўлиб, маълум тилдаги тегишли тил бирлиги қидирувини амалга ошириш мақсадида тузилади. Таркибидаги бирликларига қўшимча изоҳ бириктирилишига қараб корпуслар аннотацияланган ва аннотацияланмаган корпусларга ажратилади. Ўз навбатида, аннотацияланган корпуслар қидирув имконияти билан бир-биридан фарқ қилади. Айрим корпусларда қидирувнинг содда шакли – сўзшакл ва лексема қидирувигина мавжуд бўлса, айрим корпуслар лемма, сўзшакл, морфологик, синтактик ва семантик белгилар асосида қидирувни амалга ошириш имкониятига эгалиги билан фарқланади.

Иккинчидан, семантик аннотацияланган корпуслар тилнинг лексик бойлиги жамланган маълумотлар базасига эҳтиёж сезади, яъни тил корпусида семантик параметр асосидаги қидирувни йўлга қўйиш учун лемма/ лексемалар семантик тегланган бўлиши, ҳар бир сўзга унинг семантик хусусиятлари ҳақида маълумот бириктирилиши лозим.

Учинчидан, семантик аннотациялаш учун маълумотларни тилда мавжуд (агар мавжуд бўлса) электрон луғатлардан экспорт қилиш йўли билан яримавтоматик усулда олиш тўғри танлов бўлади. Шунда ҳам электрон луғатдаги маълумотларни олишда маълум маълумотлар қўлда таҳрирланиши

мақсадга мувофиқ бўлади.

Тўртинчидан, электрон (манба) маҳсулотлар мавжуд бўлмаган тилларнинг корпусини яратишда, албатта, маълумотлар омбори яратилишига эҳтиёж сезилади. Хусусан, ўзбек тили бирликларига автоматик ишлов бериш учун батафсил семантик ахборотга эга бирликлар базасини яратиш мақсадга мувофиқ.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР:

1. Бирюльцев Е.В., Елизаров А.М., Жильцов Н.Г., Иванов В.В., Невзорова О.А., Соловьев В.Д. Модель семантического поиска в коллекциях математических документов на основе онтологий // Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010. – Казань, 2010. – 314 с.

2. Загорулько М.Ю., Кононенко И.С., Сидорова Е.А. Система семантической разметки корпуса текстов в ограниченной предметной области // http://www.dialog-21.ru/media/1372/94.pdf

3. Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю. Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей // https://cache-samaramgf03.cdn.yandex.net/download.yandex.ru/IMAT2007/kobricov.pdf

4. Куканова В.В. Принципы семантической разметки национального корпуса калмыцкого языка // http://kalmcorpora.ru/sites/default/files/kukanova_25.pdf

5. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в национальном корпусе русского языка: принципы, проблемы, перспективы // http://ruscorpora.ru/sbornik2005/10kustova.pdf

6. Кустова Г.И., Падучева Е.В. Словарь как лексическая база данных // ВЯ. – 1994. – № 4. – С. 138.

7. Рахилина Е.В., Кустова Г.И., Ляшевская О.Н., Резникова Т.И., Шеманаева О.Ю. Задачи и принципы семантической разметки лексики в НКРЯ

// http://ruscorpora.ru/sbornik2008/10.pdf

8. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Лексико- семантическая разметка в национальном корпусе русского языка. // http://ruscorpora.ru/sbornik2005/10kustova.pdf ;

9. Шимчук Э.Г. Русская лексикография: учеб. пособие для студ. филол. фак. высш учеб. заведений / Э.Г.Шимчук. – Москва: Издательский центр «Академия», 2009. – 336 с.

10. https://ruscorpora.ru/new/reqsem.html

11. https://processing.ruscorpora.ru/search.xml

12. https://academic.ru/searchall.php

13. https://processing.ruscorpora.ru/search.xml