

Matnga semantik ishlov berish tizimlari

УДК: 81'1:81'37[=512.133]

СЕМАНТИК РАЗМЕТКА ТИЗИМИ ВА МОДЕЛИ

Ахмедова Дилдора Баходировна

daxmedova77@gmail.com

Филология фанлари бўйича фалсафа доктори (PhD)

Бухоро давлат университети катта ўқитувчиси

Аннотация. Мақолада корпус семантик разметкаси хусусияти ва зарурий воситалари ҳақида мулоҳаза юритилган. Рус тили миллий корпуси семантик разметкаси асосида ўзбек тилидаги атов бирликларини теглаш масаласига эътибор қаратилган.

Annotation. The article discusses the nature of the semantic layout of the case and the necessary tools. On the basis of the semantic layout of the National Corpus of the Russian language, attention is paid to the issue of equating the Uzbek language units.

Аннотация. В статье рассматривается сущность смысловой раскладки кейса и необходимые инструменты. На основе семантической схемы Национального корпуса русского языка уделяется внимание вопросу приравнивания единиц узбекского языка.

Калит сўзлар: корпус, семантик разметка, семантик теглар, лингвистик модел, WordNet тизими, “Лексикограф” базаси, семантик изоҳ, транскатегориал тег, интерфейс, лингвистик таъминот, полисемия, автоматик филтрлаш дастури.

Семантик разметка хусусияти, муаммо ва ечимлар ҳақида қилинган ишлар орасида Е.В.Рахилина, Г.И.Кустова, О.Н.Ляшевская, Т.И.Резникова, О.Ю.Шеманаеваларнинг “Задачи и принципы семантической разметки лексики в НКРЯ” [Е.В.Рахилина, Г.И.Кустова в.б., 2008] номли мақоласи алоҳида аҳамиятга эга. Мақолада семантик теглар тизимида қўйилган талаблар, семантик разметка тизими таркиби, кўп маънолилиқни филтрлаш масалалари атрофлича муҳокама қилинган, муаммолар ўз ечимини топган.

Кузатишларимиз шуни кўрсатадики, корпус семантик разметкаси учун зарурий воситалар сифатида қуйидагиларни ажратиш мумкин.

1. Ўзида маълум бир тилнинг лексикасини акс эттирувчи сўзлик.
2. Тил лексикасини тўла изоҳлай оладиган семантик луғат.
3. Семантик разметкани амалга ошириш учун лингвистик модел – қоидалар мажмуи.
4. Семантик разметка тизими.
5. Қўшимча дастурий восита: кўп маънолилик ва омонимликни фарқлай оладиган филтёр.

Ҳозирги вақтда семантик разметка учун бир қанча база ишлаб чиқилган: рус тили материаллари асосида ишланган бир неча лингвистик таъминот, дунё тиллари учун қўллаш мумкин бўлган WordNet тизими, VerbNet инглиз феълларининг online луғати, феъллар устида амаллар бажаришга мўлжалланган VerbOcean базаси, USAS тизими ҳамда “Лексикограф” базаси шундай ишлар сирасидан.

Рус тили миллий корпусида семантик разметка корпус ишга тушган вақтда мавжуд бўлмаган; корпусни мукаммаллаштириш жараёнида семантик разметка ҳам ниҳоясига етиб, бугунги кунда фойдаланувчи нафақат истаган маънони ифодаловчи сўзини ва унинг контекстини, балки маълум бир феълнинг қандай сўзлар билан (масалан, нарса билдирмайдиган отнинг ҳаракат феъли билан) боғланиш ҳолати) бирика олишини ҳам кузата олади. Бу кидирув, албатта, кенгайтирилган, чуқурлаштирилган кидирув. НКРЯ (Рус тили миллий корпусининг рус тилидаги қисқартмаси; кейинги ўринларда НКРЯ)нинг семантик разметкаси классификацияси “Лексикограф” базаси асосида ишланган; корпус разметкасини ишлаб чиқиш жараёнида “Лексикограф” базаси тўлдирилган. [Кустова Г.И., Ляшевская О.Н. в.б., 2005]

Базани ишлаб чиқишда ҳам маълум принципларга таянилади. Корпуслаштирилаётган тил хусусиятидан келиб чиққан ҳолда ҳар бир семантик луғат ва семантик разметканинг семантик разметка тизими ишлаб чиқилади.

Е.В.Рахилинанинг фикрига кўра, “Лексикограф” базасида бўлгани сингари корпусда ҳам ҳар бир сўз туркуми учун теглар мажмуи ўзига хос. Танлаб олинган семантик изоҳлардан қуйидагиларни келтириш мумкин:

- 1) феъл туркуми учун: ҳаракат, физик таъсир, яратиш, йўқ қилиш, эгалик қилиш, 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. [Рахилина Е.В., Кобрицов Б.П. в.б., 2005]

Албатта, бирликка берилган семантик изоҳ қанча кўп бўлса, унинг бошқа бирликлар билан учровчанлик хусусияти корпусда шунча кўпроқ намоён бўлади. Бундай вазифани транскатегориал белгилар акс этган лексик маълумотлар базаси жуда яхши уддалайди. Масалан, “ҳаракат” теги *бормоқ* феъли, *пиёда* сифати, *оёқ* отини ифодалаб келиши мумкин. Мана шундай бир неча туркумни изоҳлай оладиган тег транскатегориал тег саналади. Изоҳлашнинг бундай кўриниши нега кераклигини шарҳлашга ҳаракат қиламиз. Маълум бир фикрни ифодаламоқчи бўлган муаллиф (ўз она тилида ёзаётган/гапираётган бўлиши, таржима қилаётган бўлиши мумкин) фақат бир сўз туркумидаги сўзни қидириш учун сўров бериши мумкин. Аммо айтиш билан бир маънони ифодаловчи сўзлар қайси туркумда мавжудлигини фойдаланувчи билмайди. Шундай пайтда транскатегориал тег бириктирилган барча сўз сўров натижаси сифатида интерфейсда кўринади. Бу эса фойдаланувчига ифодаламоқчи бўлаётган фикрини кенгроқ планда бериш имконини яратади. Ҳатто бундай сўровни грамматик шаклланган ҳолда амалга ошириш, сўраш имкони мавжуд бўлади. Масалан: *сущ. & им.п. + глагол: движение & личная форма*. Бундай сўров берилиши натижаларни нисбатан аниқроқ кўрсатади. Электрон/қоғоз шаклидаги идеографик луғат ҳам бундай қулайлик ярата олмайди, корпус шундай воситалардан юқоридаги хусусияти билан устун туради. Семантик разметкага эга корпуснинг яна бир афзал томони шундаки, юқорида кўрсатилган вазифани бажаришга мўлжалланган луғатларда бирикувнинг чегараланган миқдори келтирилади. Яна бир жиҳат: *ҳаракат*ни ифодаладиган феъл билан бирикувчи *йўл, соат, газ, тутун* каби сўзлар ишлатилган саноксиз контекстларни кўрсатади; тадқиқотчи турли позицияга эга конструкцияни танлайверади, яъни корпус тил имкониятларини луғатга нисбатан минг, юз минг марта кўпроқ намоён эта олади.

Е.В.Рахилина фикрича, НКРЯда лексик классификацияга ёндашув соф семантик аспектда.[Рахилина Е.В., Кустова Г.И. в.б.,2008] Шундай феъллар

борки, улар фақат контекстда ҳаракат маъносини билдиради, контекстдан ташқарида бундай хусусиятга эга эмас. Шундай бирликларни теглаш ҳам лингвистик таъминот асосида амалга оширилади.

Фойдаланувчи семантик разметкадан унумли фойдалана олиши учун қидирув/сўров шакллари мукамал ишлаш муҳим: бу интерфейсининг интуитив тушунарли бўлишини таъминлашни талаб қилади. Шунингдек, мақола муаллифлари семантик разметканинг хусусиятлари таҳлили асосида семантик синф ва унинг муҳим таксономлари мустақил (1), базавий (2), катта синфларга бўлинган (3), натижани аниқ акс эттирадиган (4), битта мукамал сўров орқали конструкцияни аниқ акс эттирадиган (5) бўлиши кераклиги ҳақида ҳулоса қилишади.

Корпус семантик разметкасида шундай бир синф борки, у ҳам нарсаларга, ҳам нопредметларга тегишли. Масалан, атоқли отларга *асбоб*, *модда*, *пайт оти*, *товушни билдирувчи бирлик*, *мавҳум отлар* кирмаганидек, атоқли отлар ҳам, нарсаларга ҳам нопредмет отларни аташи мумкин. Шунинг учун атоқли отларнинг разметкаси мустақил, алоҳида амалга оширилади. Улар лингвистик таъминот асосида автоматик тегланиши қийин. Ҳозирча НКРЯнинг атоқли отлар синфи фақат исм, фамилия, отчества, топоним каби гуруҳларга бўлинган.

Атоқли от полисемияси разметкани мураккаблаштиради: *Волга* топоним (дарё) ва маиший предмет номи (машина) бўлиб келади; *Форд* шахс номи ва машина маркасини номлайди. Лисоний полисемия корпусда омонимияни ҳосил қилади. Компьютер учун икки ҳолатда ҳам *Форд*, яъни омоним: у кўп маъноли бўлганда ҳам, омоним бўлганда ҳам дастур буни фарқлай олмайди, дастур бу бирликни $f+o+r+d$ каби ҳарфлар бирикмасидан ташкил топган бирлик сифатида ўқийди, холос. Бу муаммони ечиш учун омонимликни автоматик фарқлаш дастури яратилиши ҳамда у разметка учун база сифатида ишлаши талаб қилинади. Омонимни фарқлаш дастури модуллар мажмуи асосида яратилади. НКРЯнинг лингвистик разметкаси (морфологик, семантик, синтактик) доимий равишда мукамаллаштириб борилишини инobatга олсак, корпус муаллифлари бу муаммони ҳам ечишлари аниқ.

НКРЯнинг семантик белгилар интерфейсида *инсон хусусияти* ($t:humq$) белгисига қарама-қарши ўлароқ *физик хусусият* ($t:physq$) параметри алоҳида белгиланган. Бу белгидан тег сифатида *инсонга хос белгининг нарсага кўчирилиши* ҳоллари (метафора)да фойдаланилади: *юмшоқ нон* - *юмшоқ киши*. НКРЯ семантик разметкасининг яна бир афзаллик томони шундаки, унга кўп

маънолиликни автоматик филтрлаш дастурини киритиш кўзда тутилган: бирлик кўп маънолилик апробациясидан ўтади [Шеманаева О.Ю., Кустова Г.И. в.б., 2007. – С. 582-587], омонимликни автоматик аниқлаш ва бартараф этиш функцияси эса назарий жиҳатдан ўрганилган, амалиётга киритиш босқичида.

Фойдаланилган адабиётлар:

1. Красильщик И.С., Рахилина Е.В. Предметные имена в системе “Лексикограф” // НТИ, сер. 2. – 1992. – № 9. – С. 24-31.; Кустова Г.И., Падучева Е.В. Словарь как лексическая база данных // Вопросы языкознания, 1994. – № 4.
2. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. – Москва: Индрик, 2005.
3. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Лексико-семантическая разметка в национальном корпусе русского языка. (Электрон ресурс): <http://ruscorpora.ru/sbornik2005/10kustova.pdf>.
4. 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>.
5. Рахилина Е.В., Ляшевская О.Н., Кобрицов Б.П., Кустова Г.И., Шеманаева О.Ю. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции “Диалог 2006”, – 2006. – С. 445-450.
6. Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Иомдин Л.Л., Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». – 2007. – С. 582-587.