Scientific reports of Bukhara State University

Volume 1
Issue 4 *BSU 2018 issue 3*

Article 7

9-26-2018

THE PRINCIPLES OF THE CORPUS ORGANIZATION

Shahlo Mirdjanovna Hamroyeva
Teacher at Uzbek linguistics department, Bukhara state university

УДК 83.

КОРПУС ТУЗИШ ТАМОЙИЛЛАРИ
ПРИНЦИПЫ СОЗДАНИЯ КОРПУСОВ
THE PRINCIPLES OF THE CORPUS ORGANIZATION

Hamroyeva Shahlo Mirdjanovna
Teacher at Uzbek linguistics department, Bukhara state university

Таянч сўзлар: разметка турлари, морфологик разметка, семантик разметка, синтактик разметка, “таггинг”, “парсинг”.
Ключевые слова: виды разметки, морфологическая разметка, семантическая разметка, синтаксическая разметка, “таггинг”, “парсинг”.
Key words: descriptions of tagging, morphologic tagging, semantic tagging, syntactic tagging, “taggers”, “parsers”.

Ушбу мақолада корпус тузиш тамойил ва босқичлари, разметка турлари, лингвистик ва экстралингвистик разметка ҳақида фикр юритилган.

В данной статье рассматриваются принципы и этапы создания корпусов, виды разметки, а именно лингвистическая и экстралингвистическая.

The article deals with the principles of corpus organization, the descriptions of tagging,as well as linguistic and extralinguistic tagging.

Кириш. Корпусни лойиҳалаш ва тузиш босқичининг технологик жараёни. Корпус лойиҳаси, уни тузиш босқичлари ва кейинчалик такомиллаштириб бориш йўлларини қамраб олиши корпусни мукаммал шакллантиришнинг асосий омили ҳисобланади [1]. Корпус тушунчаси тилшунос учун тадқиқотининг муҳим қисми саналган анъанавий картотекаларнинг давоми саналиб, картотекалар XX асрга келиб компьютерлаштирилди ва ундан оммавий фойдаланиш имконияти пайдо бўлди. Картотекаларнинг корпусга айланишида, албатта, интернет тармоғи салмоқли аҳамият касб этди. Натижада, турли лингвистик тадқиқотлар олиб бориш имконини берувчи катта ҳажмли матнларнинг умумистеъмол варианти пайдо бўлди. Бу борада луғат ва грамматикалар учун асос вазифасини ўтайдиган тил материалининг кўлами ҳамда баланси масаласи кун тартибига чиқиб, хусусан, миллий корпуслар яратиш жараёнида кўндаланг турди. Корпуснинг репрезентативлик масаласи матнлар етарлилиги ва хилма-хиллиги билан ҳал этилди. В.П.Захаров ва С.Ю.Богданова фикрича, корпуснинг жанрий-мавзувий тузилиши кўриб чиқилаётганда корпус матни сифатида қандай бирлик олиниши муаммосига алоҳида эътибор қаратиш лозим бўлади. Масалан, газеталардаги кичик реклама матни алоҳида матн сифатида қараладими ёки уларни бир матнга бирлаштириш лозимми? Газета мақоласи матн саналадими ёки газетанинг битта сонини яхлит матн сифатида баҳолаш керакми? Ҳар бир шеър битта матнми ёки шеърий тўпламни яхлит ҳолда киритиш керакми? Бир-бирига жавоб тарзида ёзилган, моҳиятан бир мавзу муҳокама қилинган, нашр этилган мактублар битта матнми ёки алоҳида корпус бирлиги сифатида ёндашиш лозимми? Бу саволларга тузувчи корпуснинг тури ва кейинчалик бажарадиган вазифа-сидан келиб чиқиб жавоб беради. Миллий корпус ёки махсус корпус эканлигига қараб корпус бирлиги белгиланади. В.П.Захаров, С.Ю.Богдановалар корпусни лойиҳалаштириш жараёнининг муҳим жиҳати сифатида хронология муаммосини ҳам санаб ўтади. Масалан, тилнинг замонавий корпуси деганда нима тушунилиши лозим? Турли жанрларда корпус-нинг хронологик чегараси турлича бўлиши табиий. Корпус кенг омма фойдаланиши ҳамда хилма-хил топшириқлар бажарилиши учун (масалан, бошқа графика асосида рус тилида ёзилган матнларни ўрганиш учун) тузилади. Лойиҳалаш босқичининг яна бир масаласи корпусда матннинг бошланғич шаклидан қандай қисм олиниб, нималар чиқариб ташла-ниши, матн таркибида мавжуд бўлган расмлар тил материалига

тегишли бўлмаганлиги учун корпус таркибига кирган матндан чиқариб ташлаш, жадвалларни корпусга мослаб қайта ишлаш ҳам муҳим. Улар матннинг мазмунини ифодалашда аҳамиятли, лекин корпус таркибида қолдирилса, разметкалашда қийинчилик туғдиради. Цитата, кўчирма гаплар, ўзлашма бирлик (атама)лар, ўлчов бирликлари ҳам алоҳида эътибор талаб қилади. Юқорида санаб ўтилган масалалар корпусни лойиҳалаштириш босқичида ҳал этилиши лозим. Уларнинг айрими эса корпус тузиш жараёни ҳамда корпусдан малакали фойдаланишда ҳал этилади. Шу билан бирга корпусни ишга туширишдан олдин фойдаланувчи билан қайтар алоқани назарда тутиш ҳам лозим. Мутахассислар корпус тузишнинг технологик жараёнида қуйидаги босқичларни ажратишади [4]:

1. Белгиланган манбага мувофиқ ҳолда матннинг корпусга киришини таъминлаш.
2. Матнни автоматик ўқилиш шаклида қайта ишлаш. Корпусга киритиладиган электрон шаклидаги матн турли усул билан олинган бўлиши мумкин: қўлда киритилган, сканерланган, муаллифлик нусхаси, ҳадя ва айирбошлаш, Интернет, нашриётлар томонидан корпус тузувчисига бериладиган оригинал-макетлар.
3. Таҳлил ва матнга дастлабки ишлов бериш. Ушбу босқичда турли манбалардан қабул қилинган матнлар филологик текширув ва таҳрирдан ўтади.
4. Конверсиялаш ва графематик таҳлил. Баъзи матнлар қайта кодлаштириш жараёни амалга ошадиган илк машина ишловидан қайта-қайта ўтади, номатний қисмлар (расм, жадвал) ўчирилади ёки ўзгартирилади. Матндаги бўғин кўчириш, чегаралар (MS-DOS матнларида) бекор қилинади, тире ва бошқа белгилар бир хиллигига эришилади. Графематик таҳлил корпусга кирувчи матнни қисмга (сўз ва боғловчи) ажратиш, номатний элементни ўчириш каби амалларни бажаришдан иборат.
5. Ностандарт (нолексик) элементни белгилаш ва расмийлаштириш, махсус матний элементни (қисқартма асосида ёзилган ном (исм, фамилия), бошқа алифбода ёзилган ўзлашма лексема, расмга берилган ном, изоҳ, зарварақ, адабиётлар рўйхати ва б.) бир хил мезон асосида қайта кўриб чиқиш. Албатта, бу амаллар автоматик равишда матн муҳаррири томонидан бажарилади.
6. Корпусни лойиҳалаштиришнинг кейинги босқичи манбани саралаш. Корпуснинг аҳамияти унинг бир тилдаги кенг кўламли матнларни бир жойга йиғиб, тартиб берилганлигида эмас, шу сабабли уни тузишда бир неча мезон асосида иш кўрилади. Корпус материалини саралашда корпуснинг асосий бирлиги нимадан иборатлиги, унинг ҳажми қандай бўлиши (унда қанча сўз бўлгани маъқул), ёзма матн қайси манбага асосланиши ва қанча миқдорда бўлиши, унга кирувчи матн тилнинг қайси соҳасига тегишли бўлиши каби масалаларга ечим топилади. Ушбу саволнинг илк жавоби 1965-1980 йилларда Р.Г.Пиатровский ва унинг шогирдлари томонидан берилган эди. Улар частотали луғат ва лингвостатистик тадқиқот ўтказиш учун матн танлаш тамойилларини тузиб чиқишган. Бу муаммо Л.Н.Засорина таҳрири остидаги частотали луғат сўзбошисида ҳам кўтарилган. Ўшанда илк маротаба матн танлашнинг статистик усули, ҳажми, миқдори каби омиллар санаб ўтилган. Корпуснинг асосий бирлиги сўз шакл, ўзак (негиз, лемма) ва гапни ташкил этади. Тузиладиган корпус ҳажми эса корпуснинг мақсадидан келиб чиқиб белгиланади. Агар у ҳарф, ҳарфий бирикма, товуш, дифтонгларни тадқиқ этишни мақсад қилган бўлса, у қадар катта бўлиши шарт эмас. Матннинг лексик бирликлари, морфологик ҳодиса, синтактик ва услубий хослигини тадқиқ этиш мақсадида тузилса, катта ҳажм талаб этилади. С.А.Шаровнинг фикрича [2], саралаш жараёнида қайси жанрга оид матнни (наср, драма, шеърият, илмий матн, газета, журнал материали ва ҳ.) танлаш, матннинг қайси даврни (замонавий, 10 йиллик, 50 йиллик ва мумтоз матн) қамраб олиши, матн фақат адабий тилда бўлиши ёхуд бошқа манбалар ҳам кириши каби масалалар ҳам муҳим аҳамият касб этади.

Корпус тузувчиси бу жараёнда, албатта, тилшунос ва лингвостатистика мутахассиси ёки анкета методига мурожаат этади. Корпус тузиш жараёнида муаллиф

ўз тажрибасига таянган ҳолда корпуснинг умумий ҳажми, матннинг нашр вақти, матн сони, элементар танлов ҳамжи, танланадиган жанр хили ва турини асосий омил ҳисоблайди. Сўровнома усули “Америка мерос корпуси” (Тҳе Америсан Ҳеритаге Интермедиате Сорпус) тузувчилари томонидан қўлланган. Мутахассислар 5 млн сўз шакл ҳамда инглиз тилида 22 турдаги болалар ва ўсмирлар жанрига оид матнларни киритишди. АҚШнинг 221 та мактабига қандай матнни танлаш мақсадга мувофиқлигини аниқловчи сўровнома юборишди. Сўровнома натижаси ўрганилгач, 19 000 номдаги китоблар рўйхати тузилди. Бу асосда ҳар бири 500 сўз шаклдан иборат 1045 матн танлаб олинди. Хулоса сифатида корпусни лойиҳалаштириш жараёнида материал (матн) танлаш, саралаш, уни техник жиҳатдан корпусга мослаштириш энг асосий босқич эканлигини қайд этиш жоиз.

Разметканинг корпус тузишдаги аҳамияти. Разметка, унинг тур ва хусусияти корпус лингвистикаси мутахассислари томонидан кенг тавсифланган. Корпусни оддий электрон тўплам ёки виртуал кутубхонадан фарқловчи, матн устида турли лингвистик амалларни бажаришга имкон берувчи энг асосий восита унинг разметкасидир. Аввалги бобда корпус ёрдамида бажарилиши мумкин бўлган бир қанча ишлар ҳақида айтиб ўтгандик. Бу вазифаларни бажариш учун фақатгина корпуснинг ўзи мавжуд бўлиши етарли бўлмай, корпус материалига турли қўшимча лингвистик изоҳ ҳам бўлиши керак. В.П.Захаров, Б.Кутузов, С.А.Шаровлар бу борадаги ишларида [2] разметка ва унинг корпусдаги аҳамиятини атрофлича ёритиб беришган. Шунингдек, З.П.Захаров раҳбарлиги остида Л.В.Северюхина ишларида [3], В.В.Рыковнинг шу мавзудаги маърузаларида ҳам разметка масаласи яхши ёритилган. Уларнинг ҳар бири разметканинг алоҳида жиҳатини таърифлаган. Барчасини умумлаштирган ҳолда лингвистик аннотация ёки корпус разметкаси (инг. linguistic markup) атамаси корпусга матннинг қисми бўлмаган, лекин шу матн ҳақида қўшимча маълумот (метаахборот) киритиш, деб хулоса қилиш мумкин. Бу қўшимча маълумотнинг энг оддий кўриниши сўз туркуми ҳақидаги ахборот бўлиб, у қуйидаги шаклда берилиши мумкин: Осмон (от) остидаги (от) ҳамма (олмош) нарса (от) омонатдир (сифат). (Лао Цзи). Разметка корпусни автоматик таҳлил қилишда қўл келади. Матндаги сўзларни бир марта туркумга ажратсак, исталган тадқиқотни бажаришда (масалан, корпусдаги барча сифатларни ажратиш) фойдаланиш учун қўл келади. Разметка масаласига тўхталишдан олдин разметка тизимлари тарихига назар ташлаш фойдадан холи бўлмайди. Ўтган асрнинг 80-йилларида SGML (Standard Generalized Markup Language) номи остидаги электрон матнлар разметкаси стандарти қабул қилинганди. Бу стандарт дастлаб типография саноати учун ишлаб чиқилган бўлса-да, тез орада бошқа соҳаларда ҳам қўлланила бошланди. SGMLнинг моҳияти турли матн муҳаррирларида терилган ҳужжатни таҳрирлаш, таҳлил қилиш, ўзгартиришдан иборат.

Тег лингвистик восита сифатида. SGML тег концепциясини олиб кирди. “Тег”лар (инг. tags) матндаги ишчи изоҳ бўлиб, матн ҳақидаги маълумотни қамраб олади. Корпус ёрдамида статистик ҳисоб жараёнида тилимизда мавжуд сўзларнинг фақат частотасини аниқлаш эмас, балки яна бир қанча маълумотларни олишимиз мумкин. Масалан, ҳар бир сўз билан ёнма-ён унинг туркуми белгиланган бўлса, тилда турли нутқий вазиятда сўз туркумларининг қўлланилиш даражасини ҳам аниқлаш мумкин. Лингвистик разметка ҳар бир сўзнинг маълум кодга эга бўлиши билан характерланади. Ушбу код тег, сўзни кодлаш эса теггинг (инг. таггинг) дейилади. Бугунги кунда матнга лингвистик ва бошқа маълумотларни қўшишнинг умумэътироф қилинган стандарти мавжуд эмас. Лекин Тext Encoding Initiative (TEI) махсус халқаро лойиҳаси разметканинг стандарт воситасини ишлаб чиқишга мўлжалланган. Бунинг учун ҳужжат разметкасининг бутун халқаро қабул қилинган тили − SGML ва XML мавжуд. XML кенгайтмаси XML (инг. extensible Markup Language) базаси асосида қурилган ва луғат тег ва атрибути, қоидалар базасини қамраб олган аниқ грамматика тили. SGML эса ҳужжат учун разметка тилини аниқлайдиган метатил (метаязык).

Анъанавий теглар жуфт ҳолда қўлланувчи (очилувчи ва ёпилувчи) учбурчак қавсдан иборат бўлади. Масалан, <ф> очилувчи, </ф> ёпилувчи тег бўлиб, очилувчи тег бераётган маълумотнинг ёпилганлигини билдиради. <ds> Бу, биринчи навбатда сизга боғлиқ </ds>, -деди у.

Бу, биринчи навбатда сизга боғлиқ, -деди у матни “кўчирма гап” (инг., direct speech) маъносини берувчи <ds>, унинг ичида келган биринчи навбатда киритмаси ажратувчи белгиси (тег) орқали ажратилган. Теглар якка ҳолда (жуфтлашмай) ишлатилиши ҳам мумкин. Масалан, оғзаки корпусда маълум ўринда тўхтамни билдириш учун <pause> теги ишлатилади. Бу якка тег ҳисобланади. Одатда, тегнинг ўзи фойдаланувчига кўринмайди, разметкаланган матнни кўрсатаётган дастур тегни қабул қилинган ва айни шу матн учун қўллашга келишилган шаклда кўрсатиши мумкин.

Матн турли томондан: alternative views. В.П.Захаров SGML типи разметкасининг ўзи-га хос имкониятидан бири матннинг турли кўринишини намойиш эта олиш эканлигини таъ-кидлайди. Бир марта разметкаланган матн қўйиладиган топшириққа биноан турли шаклда берилиши мумкин. Корпусдан фақат кўчирма гапли қўшма гапнинг муаллиф гапини ажра-тиб олмоқчи бўлсак, матнни кўриб чиқаётган дастуримиз <ds> белгиси ичига кирган қисмни яширади ва натижа қуйидагича кўринади: - деди у. Ёки биз кўраётган матнда муҳим қисм (тадқиқ манбаси) кўк рангда, муаллиф гапи қора рангда берилишини буюрамиз. Натижа қуйидагича кўринади: Бу, биринчи навбатда сизга боғлиқ, -деди у. Бундан ҳам мураккаброқ “alternative views”ни кўришимиз мумкин. Масалан, драматик асарда турли қаҳрамон нутқи алоҳида теглар билан белгиланган бўлиб, керакли ҳолда бир қаҳрамон нутқи ёки маълум икки қаҳрамон диалоги руйхатини битта буйруқ билан ажратиб олишимиз мумкин.

Матннинг автоматик разметкаси. Кўриняптики, катта ҳажмли корпусни қўлда разметкалаш узоқ муддатли, қимматли меҳнатни талаб этади. Шунинг учун ўтган асрнинг 70-йилларидаёқ бу вазифани компьютерга юклатиш бўйича бир неча лойиҳалар ишлаб чиқилди [1]. Ўшанда TAGGIT дастури Браун корпусидаги сўзларнинг 77 фоизини сўз туркумига тўғри ажратган эди, қолганини эса 10 йил давомида қўлда бажаришга тўғри келганди. Лекин 80-йилларга келиб CLAWS (Constituent Likelihood Automatic Word-tagging System) дастури Браун корпуси сўзларини 95 фоиз тўғри таҳлил қилди. Бугунги кунга келиб Европа тиллари учун сўз туркуми автоматик разметкаси (word-class tagging), гап бўлаклари автоматик разметкаси (parsing) ишлаб чиқилган. Бу ишнинг натижаси эса автоматик таржима ва интернет-қидирув тизимининг ишлашида намоён бўлади. Шу ўринда “Матнга автоматик ишлов бериш” устида иш олиб бораётган олимлар гуруҳининг (сайт http://www.aot.ru) бу борада қилаётган ишларини таъкидлаб ўтиш жоиз. Асосан, назарий лингвистикани замонавий ахборот технологияларига қўллаш устида иш олиб боришмоқда. Улар рус, инглиз ва немис тилларидаги матнларни таҳлил қилишнинг графем (сўз чегарасини аниқлаш), морфологик (сўз туркумини аниқлаш), синтактик (гап бўлакларини аниқлаш) ҳамда семантик (сўзлараро семантик муносабат) модулларини яратишди. Ўзбек тили сўз туркумларини ўқий оладиган дастур – таггинг ишлаб чиқиш компьютер лингвистикасининг олдида турган долзарб вазифалардан бири.

Ҳозирда қўлланилаётган разметкаларни лингвистик ва экстралингвистик каби турга ажратиш мумкин. Экстралингвистик разметканинг қуйидаги турлари фарқланади:

1. Матн форматининг ўзига хослигини акс эттирувчи (боб, хатбоши, қисм ва ҳ.) разметка;
2. Матн ва унинг муаллифига тегишли маълумотдан иборат разметка. Чунки муаллиф ҳақидаги маълумот нафақат ном, балки ёши, жинси, яшаган йили ва б. ҳам билдириши мумкин. Матн ҳақидаги маълумот эса матн (асар) номидан ташқари тили, ёзилган ҳамда нашр этилган йилларини қамраб олиши мумкин. Бундай маълумотларнинг мавжудлиги базада қидирувни анча деталлаштирилган кўринишда амалга ошириш имконини беради.

Экстралингвистик разметка ёки метаахборот (рус., метаданные), ташқи-интеллектуал маълумотларни қамраб олувчи, библиографик, типологик, тематик ва социологик тавсифни; формал − структур разметкани (матн, бўлим, боб, қисм, абзац, гап), шунинг-дек, техник-технологик разметка (кодировка, манбанинг электрон версиясини қайта ишлаш санаси) бирлаштирувчи разметка туридир. Метаахборот мажмуи корпус фойдала-нувчисига унинг имкониятларини аниқлаб беради. Бу маълумотларни танлашда корпус тузувчиси тадқиқот мақсади, тилшунослар талаби ҳамда матнга у ёки бу қўшимча белгини қўшиш имконияти билан таниш бўлмоғи лозим. Ташқи интеллектуал разметка, биринчи-дан, тилнинг ўзаро алоқаси ва мавжудлигини аниқлаш; иккинчидан, тилнинг ўзига хос хусусиятини ўрганиш учун керак. Тилга таъсир қиладиган иккита: ташқи, нолисоний ҳамда ички омил мавжуд. Дж. Синклер ташқи омилнинг уч гуруҳини ажратади:

1. Муаллиф томонидан матн яратишга алоқадорлик омили;
2. Матннинг ташқи белгиларига тегишли омил;
3. Матннинг яратилишига сабаб бўлувчи ва аудиторияга таъсир қилувчи омил;
4. Ички омилнинг икки гуруҳи – матннинг мавзу доираси ва услубий хусусияти (услуб ва жанр) ажратилади. Масалан, рус тили миллий корпусида қуйидаги метаахборотлар мажмуи мавжуд:
5. Биринчи блок:
6. Матн муаллифи: исми, жинси, туғилган вақти (тахминий ёши).
7. Матн номи.
8. Ёзилиш вақти ва жойи (аниқ шаҳар ёки давлат бўлиши мумкин).
9. Матн ҳажми: бадиий асарлар учун меъёр сифатида ҳикоя ҳажми камида 5000 сўз; қисса ҳажми 5000дан 15000 сўзгача; роман 15000 сўздан ортиқ бўлиши одат тусига кирган.

Иккинчи блок: метаизоҳ корпусдаги матннинг 3 асосий кўриниши – бадиий матн, нобадиий матн, драматик асарларни фарқлашга мўлжалланган. Масалан, рус тили миллий корпусида бадиий асар учун қуйидаги маълумотлар кўрсатилади:

1) матн жанри: автобиографик проза, детектив, болалар адабиёти, тарихий, криминал адабиёт, саргузашт, фантастика, юмор ва сатира;
2) матн типи: автобиографик проза, латифа, детектив, очерк, адабий мактуб, қисса, масал, пьеса, ҳикоя, роман, эртак, триллер, эпопея, эссе ва ҳ.;
3) матн хронотопи: тасвирланаётган воқеанинг тахминий вақти ва жойи.

Лингвистик разметканинг ҳам ўз навбатида бир қанча кўриниши мавжуд:

1. Морфологик разметка кейинги – синтактик ва семантик разметкага асос бўлувчи аҳамиятли разметка ҳисобланиб, инглиз тилида парт-оф-спеэч таггинг деб аталувчи, сўзларни туркумга ажратувчи разметкадир. Разметканинг ушбу тури теглар ёрдамида амалга оширилиб, тегнинг матнда мавжудлик даражаси ва кўлами корпуснинг хусусияти-дан келиб чиқиб ҳар хил бўлади. Тег қанча кўп бўлса, корпуснинг лингвистик амалларни бажариш имконияти шунча кенг бўлади. Лекин кейинги авлод корпуслари ҳамжининг катталиги сабабли тегни соддалаштириш йўли қулай деб ҳисобланди. Кодировканинг соддалаштирилган тизими ортиқча хатоликларнинг олдини олади, морфологик кўп маънолиликни келтириб чиқармайди, бир неча миллион сўзни қамраб олувчи катта массивли матнларнинг разметкаланишини тезлаштиради.
2. Синтактик разметка синтактик таҳлил ва парсинг (инг. parsing) натижаси саналади. У компонентларининг грамматик структурасига асосланади. Гапдаги бўлаклар орасидаги синтагмалар график ва шажара тарзида, матнда эса улар отли, феълли ҳамда мураккаб бирикмаларни, содда ва қўшма гапни кўрсатувчи очилувчи ва ёпилувчи қавслар ёрдамида кўрсатилади. Синтактик разметкага эга корпуслар треэбанкс номи билан оммалашган. Худди морфологик разметкада бўлгани каби кейинги пайтда таҳлилни тезлаштириш мақсадида синтактик разметка ҳам соддалаштирилди, натижада бу усул скелетион парсинг номини олди.

3. Семантик разметкада ҳам бошқа разметкаларда бўлганидек, ягона стандарт шакл бўлмаса ҳам, ҳарф ва рақам ёки фақат рақамдан иборат кодлардан фойдаланилади. Биринчи ҳарф ёки рақам умумий семантик маънони, кейинги белги эса сўз маъносини янада махсуслаштирувчи кичик семантик гуруҳни ифодалайди. Семантик разметка нафа-қат сўз, балки кўплаб бирикмаларни ҳам семантик гуруҳларга бирлаштиради, бундай пайтда турли бирикувдаги бир маънони билдирувчи бирикмалар битта белги билан кодланади. Идиоматик ибора таркибидаги сўзлар миқдорини билдирувчи ахборот ҳам разметкадан жой олади. Семантик разметка корпусдаги сўз маъносининг ихтисослашуви, омонимлик ва синонимлик, маъновий гуруҳга ажратиш каби муаммони ҳал қилади. В.П. Захаров, С.Ю.Богдановалар рус тили миллий корпусини тузишда семантик разметкалаш-нинг ўз вариантини таклиф қилади [1]. Бу корпусда ҳар бир сўзга уч хил – сўз разряди, лексик-семантик тавсиф, деривацион изоҳ берилади.

4. Анафорик разметка. Матнга ишлов беришда катта қийинчилик туғдирадиган туркум бу олмош, чунки матндаги қайси сўзга ишора қилишига қараб турли маънони билдиради. Ишора сўзнинг матндаги маъносини ажратиб олиш учун разметканинг алоҳида кўринишига эҳтиёж туғилади. Анафорик разметка шу хилдаги маълумотни қўшиш учун керак бўлади. Маъноси олмош билан ифодаланаётган сўз алоҳида кодланиб, кейинги ўринда шу сўзга ишора қилаётган олмош ёнига шу код бириктирилади. Натижада олмошнинг матндаги маъноси аниқланади ҳамда керакли тадқиқотларда корпусга асосланиш имконияти пайдо бўлади.

5. Просодик разметка. Овоз транскрипция қилинган корпусда урғу ва оҳангни ифодаловчи изоҳ мавжуд бўлади. Разметканинг дискурс деб ҳам аталувчи ушбу тури шарҳ, изоҳ, эслатма, такрорлардаги тўхтамларни билдириш учун ишлатилади.

Корпусни разметкалаш (аннотациялаш) дастурлаштирилган йўллар билан амалга оширилади. Бунда, аввало, вақтни тежаш ва меҳнатни камайтириш назарда тутилса, иккинчидан, матнга автоматик ишлов бериш муаммосига ечим топилади. Ҳозирча анафорик ва просодик разметка қийинлигича қолиб кетяпти ҳамда разметка фақат қўлда бажариляпти, кейинчалик бу ҳам дастурлаштирилади, албатта. Морфологик ва синтактик разметка эса теггер ва парсинг ёрдамида амалга оширилса ҳам, бу дастурларнинг ҳам аксарияти автоматик разметкадан кейинги тузатишни талаб қилади. Чунончи, морфологик омонимия (кўпроқ флектив тилларга хос) ва синтактик кўп маънолилик ҳолатида дастур хулосанинг бир неча кўринишини таклиф қилади, тадқиқотчи эса кераклисини танлайди. Янги авлод корпуслари ҳажмининг фавқулодда катталашгани мутахассислар олдига разметканинг тўлиқ автоматлаштирилган турига ўтиш, янги теггер ва парсинглар яратиш вазифасини қўяди. Автоматик морфологик таҳлил (теггер) ёрдамида ҳар бир лексик бирликка (сўз туркуми, лемма, граммема гуруҳи) алоҳида грамматик характеристика (шахс-сон, келишик ва бошқа грамматик категория) берилади. Масалан, Браун корпусида сўзнинг частотасини аниқлаш осон. Фақат бу сўз шаклнинг (корпус тилида токен) частотаси бўлади. Лексеманинг частотасини аниқлаш учун эса ҳар бир сўзга унинг леммаси бириктирилган бўлиши керак. Корпусни автоматик разметкалашнинг оддий усули сўзнинг лексик категорияси кўрсатилган ҳажман катта электрон луғатни разметкаланган корпус билан бирлаштириш кифоя. Шунда электрон луғатдаги изоҳ (грамматик категория тавсифи) разметкаланмаган корпусдаги сўзга тег сифатида ўзлаштирилади. Масалан, корпус ва электрон луғатда ахборот ва сиёсат сўзлари мавжуд бўлса, луғатдаги “от” теги автоматик тарзда корпусга кўчади. Лекин бу усул билан корпусни тўлиқ разметкалашнинг имкони йўқ. Чунки баъзи сўз ва бирикмалар бир вақтнинг ўзида бир неча категорияга мансуб бўлиши мумкин. Бу ҳолат морфологик кўпмаънолилик (ambiguity) муаммоси ҳисобланади. Олма, математик, этик, сурма, сузма, бўлмоқ, қўллар, боғлар каби сўз шакллар бирдан ортиқ грамматик категорияга тегишли. Шу сабабли бундай сўзларни разметкалаш фақат электрон луғат ёрдамида амалга ошмайди. Табиийки, контекстда сўз шакл фақат битта категорияга тегишли бўлиб

қолади. Шунинг учун разметканинг яна ҳам мукаммалроқ кўриниши: морфологик разметка учун синтактик разметка, синтактик разметка учун семантик разметка қилиш корпусни тўлиқ ва тўғри разметкалашга олиб келади. В.П.Захаров фикрича, лингвистик разметканинг морфологик, синтактик, семантик, анафорик, просодик каби турлари қуйидаги тамойиллар асосида амалга оширилади [1]:

1) разметка схемасини тавсифлаш (асослаш);
2) умумий лингвистик тушунчалар тизимини аниқлаш;
3) фойдаланувчи учун маълум бўлган таҳлил схемасини шакллантириш ;
4) разметка схемасининг назарий анъанавийлигига эришиш;
5) халқаро андозаларга амал қилиш.

Хулоса ўрнида шуни таъкидлаш лозимки, разметканинг корпусдаги аҳамияти уни матнларнинг оддий электрон йиғиндисидан фарқловчи, матнни лингвистик амаллар бажаришга хослантирувчи корпуснинг асосий иш қуроли, воситаси бўлиб хизмат қилади.

REFERENCES

1. Zaxarov V.P., Bogdanova S.Yu. Korpusnaya lingvistika. - Irkutsk: IGLU, 2011.
2. Sharov S.A. Predstavitelniy korpus russkogo yazika v kontekste mirovogo opita. - https//lamb.viniti.ru.
3. Severyuxina L.V. Modelirovanie logiko-ponyatiynoy oblasti korpusnoy lingvistiki. - https//lamb.viniti.ru.
4. Kurs “Korpusnaya lingvistika”/ Kutuzov A.B. / Litsenzia Creative commons Attribution Share-Alike 3.0 Unported. - http://www. ruscorpora.ru.
5. http://www.unikoeln.de/philfak/englisch/bald/corpora.