THE ROLE OF THE MORPHOLOGICAL ANALYZER IN CORPUS LINGUISTICS

E'zoza F. Sharipova*

* Gulistan State Pedagogical Institute (Republic of Uzbekistan)

Abstract **–** In Uzbek linguistics, the fundamental basis of the science has been formed, and today it is increasing the potential of computer linguistics, which is widely heard on the Internet, that is, it is recognized as a novelty. The main goal of the article is that it is necessary to automatically analyze the text of thousands of scientific researches in linguistics, as well as philosophical or virtual innovations that should be made in the future.

Index Terms – morphological analyzer, automatic analysis, graphematic analysis, morphological analysis, syntactic analysis, semantic analysis, lemma, stem, tag, prosodic tag, anaphoric tag, semantic tag, morphological tagging, syntactic tag.

КОРПУС ЛИНГВИСТИКАСИДА МОРФОЛОГИК АНАЛИЗАТОРНИНГ РОЛИ

Шарипова Эъзоза Фазлиддиновна*

* Гулистон давлат педагогика институти (Ўзбекистон Республикаси)

Аннотация **–** Ўзбек тилшунослигида ҳам фанннинг фундаментал асоси шакллантирилган бўлиб, бугунги кунда интернет тармоқларида баралла янграётган, яъни янгилик сифатида тан олинаётган компютер лингвистикасининг салоҳиятини янада оширмоқда. Мақолада кўзланган асосий мақсад ҳам шундаки, тилшуносликда тадқиқ қилинган минглаб илмий изланишлар ҳамда келажакда қилиниши лозим бўлган фалсафий ёки вертуал янгиликларнинг ўқувчи оммасига етказиш учун матининг автоматик таҳлили жуда зарур эканлиги ёритиб берилди.

Калит сўзлар – морфологик анализатор, автоматик таҳлил, графематик таҳлил, морфологик таҳлил, синтактик таҳлил, семантик таҳлил, лемма, стем, тег, просодик тег, анафорик тег, семантик тег, морфологик теглаш, синтактик тег.

I. Кириш

Морфологик анализатор муаллифларнинг фикрича, анализатор вазифалари сирасига сўзнинг барча грамматик шаклларини тавсифлаш, унинг айнан маълум бир матндаги истисно ҳолатлари ҳақида маълумот бериш кабилар кириши таъкидланган. Анализаторнинг энг охирги, кўп вақт сарфланадиган вазифаси нотаниш сўзнинг эҳтимолий грамматик маъносини аниқлашдан иборат бўлади.

Морфологик анализаторнинг муҳим томони шундаки, у битта тилга мослаштирилмаган, балки бир неча тилни таҳлил қилишда фойдаланиш, маълум бир тилнинг қоидалари мажмуи луғат базасига жойлаштирилиб, морфологик таҳлил алгоритми турли тилларга жойлаштирилиши мумкин.

Морфологик анализаторда морфологик маълумотлар базасини яратиш ҳам муҳим. Бунда фойдаланувчи эҳтиёжи талабига жавоб берадиган кенг кўламли маълумотлар мажмуи назарда тутилади. Ушбу маълумотлар сараланади, маълумотлар базасини бошқариш тизими томонидан бошқариладиган жадвал шаклида сақланади. Ҳозирги вақтда маълумотлар базаси билан ишлашга мўлжалланган кўплаб тизим мавжуд: SQL, MySQL, Oracle, Accss. Катта ҳажмдаги маълумотлар билан ишлаш доим ҳам қийин кечади, ваҳоланки, ҳар бир тизимнинг ўзига хос афзаллиги мавжуд.

II. НАТИЖАЛАР

Морфологик таҳлилни, амалга ошириш учун, табиий тилни синаб кўриш ва таклиф қилинган алгоритмни текшириш мақсадида Embarcadero RAD Studio синов дастури яратилди, ушбу анализаторда Access маълумотлар базаси билан ишлаш тизимидан фойдаланилган. Мазкур анализатор маълумотлар базасининг уч жадвали (асос, қўшимча ва сўз туркуми) ҳамда ушбу жадваллар ўртасидаги муносабатдан иборат [18].

Тадқиқотчи Ш.Ҳамроеванинг таъкидлашича, мукаммал морфологик анализатор яратиш қийин. Шундай бўлса-да, морфологик анализатор тузиш учун қуйидаги алгоритмни амалга ошириш нисбатан аниқ таҳлилни амалга оширувчи анализатор яратишга асос бўла олади:

1) сўзнинг грамматик шакли фрейм-моделини яратиш. Бунда кўмакчи морфема категориясининг тўғри аниқланиши учун туркумга оид қўшимчалар жадвали тузиш назарда тутилади;

2) анализатор тузишда компьютер хотирасига мурожаат сонини камайтиришга мўлжалланган алгоритмлар ишлаб чиқиш;

3) синов дастури ишлаб чиқиш, натижаларни унда тажрибадан ўтказиш.

Яна бир тадқиқотчимиз К.Шадмановнинг таъкидлашича, тизимли дастурнинг назарий асосларидан бири – формал тил назарияси. Бу назария ўзининг математик тушунчалари билан оддий дастурчига мураккаб туюлади, лекин тизимли дастур шундай назария асосида яратилади. Ҳар қандай дастурий интерфейс бирор-бир формализмга, формализм эса формал тил назариясига таянади. Умуман олганда, ихтиёрий тилда (форма ёки табиий) тузилган гап мантиқан тўғри бўлиши учун у қуйидаги шартларни қаноатлантириши керак:

1) сўзлар алифбо талабига мос ёзилган бўлиши (морфология, лексика);

2) гап қурилишида грамматик хатога йўл қўймаслик (грамматик, синтактик);

3) гап мазмуний қурилиши (семантика) нинг тўғрилиги [16].

Туркий тилларнинг морфологик анализаторларини ишлаб чиқиш ўтган асрнинг 60-йилларида бошланган [15], дастлабки анализаторларнинг асосий хусусияти шунда эдики, улар айнан битта тилга мўлжалланмаган: бир морфоанализаторни бошқа тилга мослаштириш мумкин. Чунки дастурга морфотактик қоидалар бириктирилган, қоидалар тизими тилнинг лексикон ҳамда аффиксал морфемалари билан тўлдирилса, бошқа тилнинг анализатори сифатида ишлатиш имкони пайдо бўлади; бошқа тилнинг анализаторини яратиш учун алоҳида кодлар ёзишга зарурат мавжуд эмас. Ўшандан ҳозирги даврга қадар технологиялар ўзгарди, туркий тилларнинг универсал морфологик анализаторлари яратилди, луғатлар ҳажми, маълумотни қайта ишлаш тезлиги ошди. Туркий тиллар морфоанализаторини ишлаб чиқиш ҳаракати бошланганига 50 йилдан ошганига қарамасдан, бу соҳа ўсиши ҳалигача туркий тилларнинг барчасида турли даражада. Татар [13], бошқирд [11], қозоқ [17], чуваш [7], турк [2], хакас [6] тиллари ҳамда универсал [4] морфологик анализатор ҳақидаги ишлар фикримизни қўллаб-қувватлайди.

Адабиётларда автоматик таҳлилнинг стемминг, сўзшаклнинг луғат асосидаги таҳлили, мантиқий ёндашув асосидаги таҳлил, жадвал [9] асосида, луғатсиз таҳлил каби турлари кўрсатилади. Мутахассислар автоматик морфологик таҳлил стемматизация, лемматизация, граммемалаш каби асосий блоклардан ташкил топишини уқтиришади [5].

Стемматизация (ёки стемминг инг. stemming сўзидан) – қидирилиётган сўзнинг асосини топиш жараёни, бунда ҳар қандай стем (сўзшакл асоси) қидирилаётган сўзшаклнинг грамматик асосига тенг бўлиши шарт эмас: таҳлил жараёнида бир парадигмага мансуб сўзшаклларнинг битта стем сифатида кўрсатилиши маълум стем(асос)ни аниқлаш учун етарли. М.Абжалова ушбу жараённи шундай тавсифлайди: "Стемминг, асосан, фойдаланувчининг сўрови бўйича изланишни кенгайтириш мақсадида қидирув тизимлари учун қўлланади, матнни меъёрлаштириш жараён қисми. Сўз асосини топишнинг муайян усули стемминг алгоритми, унинг амалга оширилиши, яъни дастурнинг ўзи стеммер дейилади [3].

Лемматизация сўзшаклни леммага келтириш жараёни бўлса, лемма сўзнинг асосий (луғатда бериладиган) шакли, лексема. Демак, лемматизация жараёни кўпроқ флектив тиллар учун (масалан, рус тили) аҳамиятли, чунки агглютинатив тилларда (жумладан, ўзбек тили), одатда, сўзнинг "стем"и лексемага тенг бўлади. Фақат флекцияга учраган сўзшаклда лемма ва стемнинг мос келмаслик ҳолатлари кузатилади. Масалан, от, олмош учун лемма

бирлик сон, бош келишикдаги шакл: бола, мен; сифатнинг оддий даража кўриниши - катта, ёруғ; равиш, сон ҳамда ёрдамчи сўзларнинг леммаси уларнинг луғатдаги шаклига тенг келишини айтишимиз мумкин.

Граммемалаш (теглаш инг. gavetg сўзидан) сўзшаклга грамматик характеристика (грамматик белги) ёзиш демак. Граммема (грамматик характеристика) – сўзшаклнинг маълум морфологик синфга мансублигини кўрсатувчи содда морфологик кўрсаткич [5].

Граммема атамасига рус тилининг ўзлашма сўзлар луғатида шундай таъриф берилади: "граммема (инг.tagging) - синтактик шакл ясовчи морфема ёки ёрдамчи сўз воситасида (мас., кўмакчи) ифодаланадиган грамматик маъно; икки (ҳатто ундан ҳам ортиқ) граммема бир морфема орқали ифодаланиши ҳам мумкин. Масалан, замон кўрсаткичи хабар майлини ҳам билдиради: келди (ўтган з., хабар м., аниқ н.). Фақат битта белгиси билан фарқ қилувчи граммема (масалан, бирлик ва кўплик) битта грамматик категорияни шакллантиради" [12]. В.А.Плунгяннин граммемани қуйидагича таърифлайди: “Граммема (инг. grammeme) грамматик категория элементларидан бири бўлган грамматик маъно бўлиб, бир грамматик категориянинг турли граммемалари бир-биридан фарқланади ҳамда бир пайтда ифодалана олмайди. Рус тилида бирлик ва кўплик сон категориясининг граммемаси, албатта, сўзда ёки униси, ёки буниси воқеланади, лекин иккаласи эмас. Шунингдек, грамматик кўрсаткич ҳам граммема дейиладики, граммема морфологик категорияни ташкил этувчи морфологик шакллар билан ифодаланади, шунингдек, синтактик шакл билан ифодаланувчи граммема ҳам учрайди. К.Пайк томонидан таклиф қилинган граммема атамаси, А.А.Зализняк томонидан анъанавий лингвистик терминга айлантирилди [10].

Олимларнинг фикрларидан шундай хулосага келиш мумкинки, граммема грамматик маъно, у грамматик категориянинг элементи, грамматик шакл билан ифодаланади. Масалан, стол леммасининг стол сўзшаклига қуйидаги граммемалар йиғиндиси бириктирилиши мумкин: (мр, ед, им, неод). Бироқ туркий тилларда граммема бириктиришнинг бу тартиби амал қилмайди, чунки битта морфема, кўпинча, бир граммемага тўғри келади; айнан бир морфема бир сўзшаклда қайта-қайта келишини тил материаллари қўллаб-қувватлайди.

Морфоанализаторда ишлатиладиган асосий тил бирликлари сифатида қуйидагилар ажратилади:

1) морф – элементар сегмент белги: шаклан тасодифан ўхшаш бўлган айнан бир морфологик хусусиятга эга бўлган минимал ҳодиса;

2) алломорф – бир морфеманинг бир хил фонетик таркибга эга бўлган морфлар йиғиндиси. Алломорфнинг асосий хусусиятларидан бири алломорф контекстини тавсифловчи хусусияти, у иккига: алломорфнинг сўзшакл чегарасидаги контекси ҳамда алломорфнинг сўзшакл чегарасидан ташқаридаги контекстига бўлинади;

3) сўзшакл – таркибий қисмлари орасида морфотактика орқали аниқланадиган алоқа мавжуд бўлган алломорфлар кетма-кетлиги [5];

4) морфема – тилнинг сўзни ташкил этувчи энг кичик маъноли бирлиги; у битта вазифани бажарувчи, турли умумий, ўхшаш хусусиятларга эга бўлган морф (алломорф)лар йиғиндисидан иборат бўлади;

5) бирикиш қоидалари – морфоанализатор ишини ташкил этувчи асосий элементлардан бири. Бу қоидалар оддий элементдан нисбатан мураккаб бирликларни келтириб чиқаради.

Умуман олганда, туркий тилларнинг кўптилли полифункционал интернет-сервис дастури морфоанализатор бирикишлар қоидасининг икки типи: бир сўзшакл ичидаги бирикиш қоидалари; "аналитик шаклларни келтириб чиқарувчи бирикиш қоидалари билан "иш кўради” [5], Бундай қоидаларнинг мавжудлиги морфоанализатор иш жараёнида алгоритмлар кетма-кетлигини камайтиради, жараённи соддалаштиради. Демак, ўзбек тили морфоанализаторини ишга тушириш учун морф, алломорф, сўзшакл, морфема ва бирикиш қоидалари каби муҳим бирликлар ажратилиши талаб этилади.

Яна бир нарсага алоҳида эътибор қаратиш лозимки, туркий тиллар морфологик таҳлилини амалга оширишда жаҳонда мавжуд уч асосий ёндашувнинг парадигматик [14], автоматик [1], генератив каби турларидан фойдаланиш мумкин.

Парадигматик ёндашувда икки типдаги луғатдан: асослар луғати ҳамда парадигмалар луғатидан фойдаланилади. Бундай усулда ишлайдиган морфологик анализаторнинг асосий хусусияти шундаки, луғатдаги ҳар бир леммага парадигмага ҳавола қилувчи индекс ёзилади. Парадигматик ёндашувдан, одатда, флектив тиллар (масалан, рус тили) морфологик таҳлилида фойдаланилади. Флектив тилларда парадигмалар ҳажми катта эмас, аммо

парадигмалар сони кўп. Анализаторнинг маълумотлар омборида ҳар бир типдаги асос учун парадигманинг тўлиқ шакли сақланади. Бундай усул UniParser анализаторида қўлланган [4]. UniParserнинг маълумотлар омбори қуйидаги файлларни ўз ичига олади:

1) асослар ажратиб кўрсатилган лексемалар рўйхати, синтактик шакл ясовчилар синфи ҳамда мазкур лексеманинг – сўзшаклларига бириктирилиши талаб қилинадиган мавжуд лексик ахборотлар;

2) турли синтактик шакл ясовчи кўрсаткичлар рўйхати [5].

Бундан шундай хулосага келиш мумкинки, туркий тилларда ҳинд-европа тилларидан фарқланиб турадиган бир қатор структур хусусиятлар бор, бу парадигматик ёндашув асосида ишлаш қатор ноқулайликлар келтириб чиқаради. Бундай хусусиятлар сифатида автомат морфологиянинг ўнг томонлама амал қилиши; агглютинация; парадигматик синфлар орасида қатъий чегаранинг мавжуд эмаслиги; парадигма ҳажмининг потенциал чегараланмаганлиги; лексик қатламнинг грамматик синф ва сўз туркумлари бўйича аниқ таснифланмаганлиги.

Туркий тиллар учун морфологик анализатор тузишнинг автоматик ёндашуви кўпроқ тўғри келади. Автоматик ёндашувга асосланган морфоанализатор FST (finite state transducer) ва WFST (англ. weighted finite state transducer) қайта ишлаш тизимига эга, кириш ва чиқиш (анализ-синтез) таҳлилни амалга ошира олади. Бундай анализаторларнинг моҳияти шундаки, улар «грамматик кетма-кетлик» қоидасига амал қилади: жараёнда морфологик бирликнинг сўзшаклдаги кетма-кетлиги қоидаси асос қилинади. Улар қандай бирликлар ишлатилиши билан фарқ қилади:

1) морфемалар кетма-кетлиги ҳамда зарурий алломорф изчиллигини кўрсатиш;

2) алломорфлар изчиллиги қоидалари.

Ўзбек тилининг морфологик анализаторини яратишда ҳам парадигматик ҳамда автоматик усуллардан фойдаланиш мақсадга мувофиқ.

Демак, морфоанализаторни ишга тушириш учун, одатда, таркибий қисми аффиксал морфемалар базаси; асос морфемалар базаси; таснифлаш қоидалари; алломорфлар мослиги қоидаларидан иборат бўлади [5], Н.А.Исраилова, П.С.Бакасовалар морфологик маълумотлар омбори тузилишини қуйидагича таснифлашади:

1) морфологик маълумотлар базаси морфологик анализ ва синтез жараёни учун талаб қилинадиган барча ахборотларни қамраб олиши зарур;

2) тилда мавжуд флекция ҳолатлари, шундан келиб чиқадиган ўзак ва қўшимча чегарасидаги фонетик ўзгаришлар ҳам маълумотлар омборида акс этиши керак;

3) морфологик маълумотлар омбори одатий синтактик шакл ясовчилар билан бирга супплетив, ўзгармас лексемалар ҳақидаги маълумотларни ҳам қамраб олиши лозим;

4) маълумотлар омбори омоним лексемалар, шунингдек, тўлиқ ва грамматик шакл таъсирида ҳосил бўлувчи омонимлар базасига ҳам эга бўлиши лозим [9].

III. Хулоса

Барча ҳодисалар ўзига хос хусусиятга эга бўлгани каби, лексема ҳам нутқда сўз шаклида юзага чиқади, бунда у грамматик жиҳатдан тугал шаклланган, яъни грамматик морфема билан бириккан ҳолда намоён бўлади.

Албатта, лексема ҳар қандай грамматик морфема билан бирика олмайди. Шунинг учун грамматик категория ҳамда шакл (форма)га хилма-хил муносабатига кўра, лексеманинг яна бир қирраси аниқланади. Лексема нутқда сўз шаклида гап таркибида маълум бир синтактик боғланиш, қуршов ва гап бўлаги вазифасида келади. Лексемага “яхлитлик” (бутунлик, субстанциаллик) назарияси асосида ёндашсак, лексеманинг гап қурилишида тута оладиган ўрни ҳам унинг қиррасидан бири саналади. Лексема серқирралигини изчил ва қарама-қаршиликдан холи таснифнинг назарий асоси юзасидан билдирилган фикр билан бирлаштирсак, биз унинг ҳар бир қирраси бўйича алоҳида-алоҳида таснифни беришимиз ҳамда лексема гуруҳини аниқлашимиз лозим бўлади. Шу йўл билан лексеманинг ҳар бир таснифда тутган ўрни асосида кашф этган белгининг йиғиндиси, унинг (лексеманинг) нисбий моҳиятини ташкил этишини тасаввур қилсак бўлади.

IV. Адабиётлар

[1] Antworth, E.L. PC-KIMMO: a two-level processor for morphologicalanalysis. Occasional Publications in Academic Computing No. 16. Dallas: Summer Institute of Linguistics, 1990. – 273 p.; Kemal Ofl azer. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, -Vol. 9, No 2, – 1994.
[2] Kemal Ofl azer. Two-level Description of Turkish Morphology. Literary and Linguistie Computing, – Vol. 9, No 2, – 994.
[3] Абжалова М.А. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун): Фил. фан. бўйича фалсафа доктори (PhD)... диссер. – Фарғона, 2019. – Б.29.
[4] Архангельский Т.А. Принципы построения морфологического парсера для разноструктурных языков: дисс... канд. Филол. наук: 10.02.21. – Москва, 2012.
[5] Гатауллин Р.Р., Гатиатуллин А.Р., Неврозова О.А., Мухамедшин Д.Р., Сулейманов Д.Ш., Хакимов Б.Э, А.Ф.Хусаинов. Формальные модели и програмные инструменты компьютерной обработки татарского языка. – Казань: Академии наук, 2019. – С.39–40.
[6] Дыбо А.В., Шеймович А.В. Автоматический морфологический анализ для корпусов тюркских языков //Филология и культура. 2014. №2.
[7] Желтов П.В. Морфологический анализатор чувашского языка. Материалы международной конференции студентов и аспирантов по фундаментальным наукам Ломоносов 2002. – Москва, 2002.
[8] Исраилова Н.А., Бакасова П.С. Морфологический анализатор кыргызского языка // Пятая Международная конференция по компьютерной обработке тюркских языков «ТигКГапя 2017». – Труды конференции. В2-х томах. Т2. – Казань: Издательство Академии наук Республики Татарстан, 2017. – С.100–117
[9] Марчук Ю. Компьютерная лингвистика. – Москва: МГУ, 2006. – С. 65.
[10] Плунгян В.А. Классификация морфологических значений и обшая морфология: Введение в проблематику: Учебное пособие. – Изд. 2-е, исправленное. -Москва: Едиториал УРСС, 2003. – С.107.
[11] Сиразитдинов З.А. Алгоритмическая грамматика словоизменения башкирского языка // (Электронный ресурс). URL: http://mfbl.ru/bashdb/algram/algram.htm; Орехов Б.В., Слободян Е.А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка //Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28–31 октября 2010г.) /отв. ред. В.А.Баранов. Уфа: Ижевск-Вагант, 2010. – С. 167–171.
[12] Словарь иностранных слов русского языка / dic. Academic/ru>dic>nsf>die_fwords> граммема/
[13] Сулейманов Д.Ш., Гильмуллин А.А., Гильмуллин Р.А. Базаморфотактических правил для татарского глагола как основа двухуровневого морфологического анализатора // Сборник трудов Международного семинара «Диалог». - Казань, 1998. – С. 597–609.
[14] Тузов В.А. Морфологический анализатор русского языка //Вестник СПБГУ, сер. 1. 1996. Вып. 1 (№15).
[15] Халитова Н.А., Закирова Р.А., Гимадудтинова Р.У. Морфологический анализ при машинном переводе с татарского языка на русский // Версятностные методы и кибернетика. Сборник работ НИИММ им. Н.Т. Чеботарева при Казанском университете, Учен. зап. Казан.ун-та, 122, № 4. – Казань: Изд-во Казанского ун-та, 1962. – С. 98–105.
[16] Шадманова К. Табиий тиллар учун лексико-семантик луғат маълумотлар базаси яратиш тамойиллари //http://buxdu.uz/index.php/uz/.
[17] Шарипбаев А.А., Бекманова Г. Ергеш Б.Ж., Бурибаева А.К., Карабалаева М.Х. Интеллектуальный морфологический анализатор, основанный на семантических сетях // Материалы международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем» (ОЗТ1$-2012). – Минск, БГУИР, 16–18 февраля 2012 г. – С. 397–400.
[18] Sharipov F. Hozirgi o‘zbek adabiy tili. – Toshkent: BOOKMANY PRINT, 2023. – 100 b.

Reviewer:
F. G. Sharipov
DSc, Professor
Gulistan State University

Auhtor(s):

Author – E’zoza F. Sharipova, ezoza_sharipova@gmail.com
Corresponding Author – E’zoza F. Sharipova, ezoza_sharipova@gmail.com