МРНТ16.31.21

А.Қ.Жұбанов

А.Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері,
филология ғылымдарының докторы, профессор Алматы қаласы, Қазақстан

ТАБИҒИ ТІЛДІҢ МӘТІН МОРФОЛОГИЯСЫН АВТОМАТТЫ ТҮРДЕ ТАЛДАУ ЖАЙЫНДА

Аннотация: Морфологиялық талдаудың логикалық көбейту әдісі, көбінде, флективті тілдерде қолданылады және негіз сөздер сөздігінің болуын талап етеді. Қарастырып отырған автоматтандырылған морфологиялық талдаудың соңғы түрі – сөздіксіз талдау немесе «тәуелсіз» талдау. Ол аффикстер кестелері көмегімен және арнайы түзілген грамматикалық мағынасы жоқ сөздердің тізімі негізінде іске асады. Морфологиялық талдаудың бұл түрі зерттеу тәжірибесінде өте сирек қолданыс тауып жүр.

Тірек сөздер: табиғи тіл, сөз, сөз тіркесі, морфология, мәтін

А.К. Жубанов

главный научный сотрудник Института языкознания имени
А. Байтурсынова, доктор филологических наук, профессор
Алматы, Казахстан

ОБ АВТОМАТИЧЕСКОМ МОРФОЛОГИЧЕСКОМ АНАЛИЗЕ ТЕКСТА ПРИРОДНОГО (ЕСТЕСТВЕННОГО) ЯЗЫКА

Аннотация: Метод логического умножения морфологического анализа, в большинстве случаев, используется в флективных языках и требует наличия словаря базовых слов. Вид автоматизированного морфологического анализа, который нами рассматривается, – это анализ без словаря или «независимый» анализ. Он реализуется с помощью таблиц аффиксов и на основе списка слов, не имеющих специального грамматического значения. Данный вид морфологического анализа находит очень редкое применение в исследовательской практике.

Ключевые слова: естественный язык, слово, словосочетание, морфология, текст.

A.K. Zhubanov

Chief researcher of the Institute of Linguistics named after A. Baitursynov,
Doctor of philological sciences, professor
Almaty, Kazakhstan

ON THE AUTOMATIC MORPHOLOGICAL ANALYSIS OF THE NATURAL LANGUAGE TEXT

Annotation. The method of logical multiplication of morphological analysis, in most cases, is used in inflexional languages and it requires availability of a dictionary of basic words. The

type of automated morphological analysis that we are considering is dictionary-free analysis or "independent" analysis. It is implemented by means of tables of affixes and on the basis of a list of words that do not have a special grammatical meaning. This type of morphological analysis is very rarely used in research practice.

Keywords: natural language, word, word combination, morphology, text

«Компьютерлік лингвистиканың» айналысатын мәселелерінің ең маңыздыларының бірі – табиғи тілдің мәтін морфологиясын автоматты түрде талдау. Егер дәстүрлі түсінік бойынша «сөз морфологиясы» деген ұғым сөздің тұлғалық құрылымын, яғни сөз түбірін және оған жалғанатын қосымшалардың түрін анықтайтын болса, компьютерлік тіл біліміндегі «морфология» терминінен басқаша ұғым туындайды. Дәлірек айтсақ, ол сөздің мәтін бойындағы сыртқы пішініне (тұрпатына) қарай, тілдік құрылымның әркилы деңгейлері бойынша тілдік мәліметтер алу мүмкіндігін білдіреді. Алғашында «морфологиялық талдау» ұғымы айтылған мағынада машиналық аударма саласында пайда болды [3].

«Морфологиялық талдаудың» жаңаша түсінігі бойынша алынатын ақпарат біздің дәстүрлі ұғымымыздағы морфологияға қатысы жоқ көптеген амалдардан тұруы мүмкін. Осының салдарынан компьютерлік тіл біліміндегі «морфологиялық талдау» ұғымы – амалдар ұғымы (операционное понятие), яғни сөздің сыртқы пішіні бойынша тілдік ақпараттарды танып-білу үшін жүргізілетін іс-әрекет деп ұққан жөн сияқты. Басқаша айтсақ, егерде дәстүрлі тіл білімі ұғымы бойынша «морфологиялық талдау» кезінде «нені талдаймыз?» деп сұрақ қоятын болсақ, компьютерлік тіл білімінде «қалай талдаймыз?» деген сұрақ қойылады, яғни ол сөз тұлғасы бойынша қажетті тілдік ақпараттарды қай жолмен алуға болатындығымен айналысады.

Машиналық (компьютерлік) аударманың алғашқы тәжірибелерінде мұндай амалдардың саны көптеп кездесетін, ал бүгінгі таңда орыс және еуропа тілдері бойынша бұл мәселе біршама шешілді деуге болады [4].

Қазақ тілі мен басқа да түркі тілдері бойынша мұндай ізденістер айтарлықтай қолға алынбай жүр.

Аталған түсініктегі «морфологиялық талдау» бірнеше бағытта құрастырылуда. Одардың біріншісі – сөзтұлғаны оның негізі мен оған жалғануы мүмкін болатын қосымшаларға ажырататын классикалық талдау негізінде модельдеу.

Екінші бағыт сөзтұлғалардың соңғы әріптерінің тіркесімдік заңдылықтарының жиілік сөздіктердегі статистикалық мәліметтеріне сүйенеді.

Үшінші бағыт – соңғы кездердегі ізденістердің нәтижесі. Бұл бағыт бойынша теңдеулердің ашық жүйелерінің пішіні ретінде морфологияның әмбебап математикалық моделі жасалады. Модельдің есептегіш мүмкіндігі негізінде сөзтұлғаларды нормалау және қажетті грамматикалық ақпарат алумен қатар, сөзтұлғаларды жинақтау (синтездеу) да іске асады [5, 44].

Сөзтұлғалардың әріптік құрамының өзгеруін анықтайтын сөздерді өзара топтауды (сыныптарға бөлуді) негіз етіп, автоматтандырылған морфологиялық талдаудың алгоритмі құрастырылады. Мұндай топтау «морфологиялық топтау» деп аталып жүр [6]. Сөзтұлғалардың жазылуындағы әріптік өзгеріске ұшырау флективті және аглютинативті тілдерде біркелкі емес. Мәселен, орыс тіліндегі: «сижу – сидишь» сөзтұлғаларының жазылуындағы сөз негізіндегі әріптер өзгеріске ұшыраса, сол сөздердің қазақша баламасы – «отырмын – отырсың» сөздерінің жазылуында қосымшадағы әріптер өзгерген.

АМПАР атты машиналық аударма жүйесінде сөздердің синтаксистік қызметі мен септік және т.б. жалғаулардың негізінде сөздердің морфологиялық кластары екі түрге бө-

ліниіп қарастырылды: 1) негіздерінде өзгеріс болатын сөздер тобы, 2) сөздердің флективті тобы.

Соңғы топ белгілі сөздерге ғана тән қасиеттегі белгілер жүйесі арқылы немесе осындай қасиеттерді бойына сақтаған дерек сөздер арқылы сипатталады.

Енді морфологиялық талдаудың түрлеріне қысқаша тоқталайық. Олар төмендегідей:

– негіз сөздердің сөздігі арқылы морфологиялық талдау жүргізу;

– сөзтұлғалар сөздігі көмегімен морфологиялық талдауды іске асыру;

– логикалық көбейту әдісін қолданып, морфологиялық талдау жасау;

– морфологиялық талдауды сөздіктердің көмегінсіз (сөздіксіз) арнайы кестелер арқылы іске асыру.

Еуропа тілдерін зерттеуде көп тараған морфологиялық талдау түрі – негіз сөздердің сөздігі мен кейбір көмекші кестелер жүйесін пайдалану. Аталған сөздіктерде ішкі флексиясыз, жәй және күрделі сөздердің негіздері мен олардың негіз тұлғалары (формалары) толық беріліп отырады. Сөздіктегі әрбір негіз сөзге морфологиялық кластардың екі түрін ажырататын шартты белгі – «код» қойылады. Ал омонимдік негіздерге шартты белгілердің (кодтардың) тіркесімдік түрлері беріледі.

Морфологиялық талдаудың екінші түрі, яғни сөзтұлғалар сөздігі арқылы жүргізілетін талдау да зерттеушілер тәжірибесінде көп қолданыс тапқан, кең тараған талдау түрі болып саналады. Олай дейтініміз, морфологиялық талдаудың компьютерлік алгоритмінде морфемаларға бөлшектеу мен оларды сөздік бойынан іздестіру әрекеттері негіз сөздер сөздігіне қарағанда жеңіл іске асады. Бірақ талдаудың бұл түрінің өзіне тән осал жерлері де жоқ емес. Мысалы, егер іздестіретін сөз сөзтұлға сөздігінде кездеспеген жағдайда, біз оның грамматикалық ақпараты жайлы да ештеңе біле алмаймыз. Сондықтан сөзтұлға сөздігі көмегімен талдау жүргізу жүйесінде аффикстер мен түбір сөздер тізімдері берілуі қажет. Осымен бірге бірнеше сөзтұлғаларды бір лексикалық бірлікке сәйкестендіруге қажетті, оларға тән қасиеттерге ие белгілер (атрибуттары) толық берілгені талап етіледі. Бұл айтылғандардан туындайтын қорытынды – морфологиялық талдаудың негіз сөз сөздіктері арқылы жүргізілгені ұтымды.

Морфологиялық талдаудың логикалық көбейту әдісі зерттеу аясынан айрықша орын алады. Бұл әдісте «сөздік функциясы» деген ұғымға ерекше көңіл бөлген жөн. Әрбір «сөздік функциясына» сөзтұлға функциясы және оған тән ақпарат сәйкестендіріледі. Осының негізінде әр сөзтұлғаның өзіне ғана тән ақпараттық деректер арқылы функция мәнін арнайы кестемен беруге мүмкіндік туады. Бірақ аталған функцияның мәнін басқаша, төменде көрсетілетіндей тиімді амалдар арқылы беруге де мүмкіндік бар:

1) сөзтұлға – әріптер тізбегі ретінде морфемдік сегменттерге бөліктенеді;

2) сөзтұлға – морфемдік сегменттер тізбегі ретінде, басқа морфемдік элементтердің реттелмеген жиынымен ауыстырылады;

3) сөзтұлғаға, морфемдік жиын ретінде, белгілі-бір ақпарат сәйкестендіріледі;

4) бұл ақпарат сөзтұлға жайындағы қажетті деген ең соңғы ақпарат түрінде қайтадан өзгертіліп беріледі [5, 46].

Әрбір морфемаға құрылымында сондай морфемасы бар сөзтұлғалардың жиынтық ақпаратын сәйкес қоюға болады. Логикалық түсінік бойынша мұндай ақпараттар жиынтығы дизъюнкциямен, яғни сөзтұлғаның құрылымдық белгілерінің бір-біріне қарсы қойылуымен сәйкес келеді. Сөзтұлға жайлы ақпарат осы сөзтұлғаның табиғатына тән қиылысудан немесе логикалық конъюнкция сияқты морфемалық ақпараттардан тұратынына көз жеткізуге болады.

Морфологиялық талдаудың логикалық көбейту әдісі, көбінде, флективті тілдерде қолданылады және негіз сөздер сөздігінің болуын талап етеді.

Қарастырып отырған автоматтандырылған морфологиялық талдаудың соңғы түрі – сөздіксіз талдау немесе «тәуелсіз» талдау. Ол аффикстер кестелері көмегімен және арнайы түзілген грамматикалық мағынасы жоқ сөздердің тізімі негізінде іске асады. Морфологиялық талдаудың бұл түрі зерттеу тәжірибесінде өте сирек қолданыс тауып жүр.

Енді автоматтандырылған морфологиялық талдаудың қазіргі күйі қандай деген сұрақтың жауабына қысқаша тоқталайық.

Қазіргі кезде кез келген мәтіннен компьютер арқылы сапалы ақпарат алу жүйесіне қойылатын талап заман сұранысына қарай өсті. Енді компьютер көмегімен морфологиялық талдау жасау алгоритмін іске қосу мәселесіне, негізгілерін алғанда, мынадай талап қойылуда:

– кез келген тақырыптағы мәтіннің 98-99% қамтитын көптақырыпты (политематический) «аса күшті» сөздік морфологиялық талдау жүйесінің негізін құруы тиісті;

– автоматты түрдегі талдау алгоритмі кез келген сөзтүрленістерін ескеретіндей мүмкіндікке ие болуы қажет. Осы жағдай ескерілсе ғана сөзтұлғаны жанжақты тану мүмкіндігі артады және сан жағынан алғанда ондай мүмкіндіктің нәтижесі сөздікте қамтылған лексикалық бірліктен бірнеше есе артық болуы мүмкін;

– автоматтандырылған морфологиялық талдау жүйесінде «жаңа» сөздер, сөздікте қамтылған басқа сөздермен бірдей дәрежеде қарастырылуы қажет және оларды дұрыс танудың ықтималдығы 90-95% кем болмауы қажет;

– мәтінді компьютер арқылы өңдеу жылдамдығы тұтынушыны қанағаттандыратындай болуы қажет және оған енгізілетін мәтін көлеміне шек қойылмауы керек;

– компьютер арқылы жүргізілетін морфологиялық талдау жүйесінде қажеттікке қарай өзгерістер жүргізу (сөздікті толықтыру, өңдеу және т.б.) мүмкіндігі алдын-ала ескерілуі қажет, яғни бұл аталған жүйені «үйретуге» барынша мүмкіндік тууы керек [5, 48-49].

Бүгінгі таңда, орыс тіліне және басқа да еуропа тілдеріне қатысты морфологиялық талдау жүйелеріндегі кейбір сөздіктер мен мәтіндердің көлемдері жайлы мынандай мәліметтерді айтуға болар еді: негіз сөздерден тұратын политематикалық сөздіктің көлем 100 мың лексикалық бірліктен кем емес; сөздік, көлемі 30 млн. сөзқолданыстан асатын политематикалық мәтін негізінде жасалғандықтан, оның кез келген тақырыптағы ғылыми-техникалық мәтінді қамту дәрежесі өте жоғары; көлемі 3 млн. сөзқолданыстан тұратын мәтіннің негізінде түзілген сөзтұлғалар сөздігі 46 мың лексикалық бірлікті құрайды. Сөз болып отырған екі сөздікте де сөздердің құрылымы мен синтаксистік қасиеті жөнінде толық түрдегі грамматикалық ақпарат берілген [7].

Автоматтандырылған морфологиялық талдау сөздіктердің айрықша (жаңа, арнайы) түрлерін өмірге келтірді. Оларда тілдік бірліктердің тұлғасына, туындау ережелеріне, сөзтудырғыш, сөзтүрлендіргіш топтарына қарай реттеу жақтары қарастырылды. Осындай типтегі сөздікке тілдік бірліктің соңғы жағынан бастап әліпбиге түсірілген кері әліпбилік (обратно-алфавитный) сөздік деп ата-лынатын сөздікті жатқызуға болады. Морфологиялық талдаудың талабына сай түзілген бұл сөздікте сөзтудырғыштық, сөзтүрлендіргіштік типтегі сөздер және соңғы қосымшалары бірдей болып келетін күрделі сөздер өзара жіктелген ретте көрініс табады. Кері әліпбилі сөздік бойынша сөздердің морфологиялық құрылымына қатысты көптеген мәселелерді шешуге болады. Мәселен, қосымшалары бірдей болып келетін сөздердің топталып берілуі олардың грамматикалық сипатын және қосымша мен қандай сөзтүрлендіргіш типке жататындығы арасындағы қатынастарға қатысты деректерді анықтауға мүмкіндік туады. Грам-

матикалық форманттардың синонимдік, омонимдік қатарларын жеңіл ажыратуға және олардың тіркесімдік, сандық мөлшерінің сипаттарын анықтауға мүмкіндіктер туындайды. Кері әліпбилі сөздік бойынша, аналитикалық тілдерге қарағанда флективті және аглютинативті тілдер жүйелерінен морфологиялық форманттарға байланысты ақпараттарды көптеп алуға болады.

Осы мақалада сөз болған компьютердің араласуымен ғана жүзеге асатын морфологиялық талдаулар мен басқа да зерттеу түрлерін мемлекеттік тілімізге пайдалана отырып, ана тіліміздің компьютерлік лингвистика саласын дамытуды және соның нәтижесінде қол жеткізген жетістігімізді күнделікті өмір қажетіне пайдалануды заман талабы деп түсінеміз.

ӘДЕБИЕТТЕР ТІЗІМІ:

[1] Городецкий Б.Ю. Актуальные проблемы прикладной лингвистики // Новое в зарубежной лингвистике. – Вып. XII. – 1983. – С. 84.

[2] Баранов А.Н. Введение в прикладную лингвистику. Эдиториал УРСС; – М., 2001. – С. 105.

[3] Василевский А.Л., Марчук Ю.Н. Вычислительная лингвистика. Учебное пособие для студентов отделения прикладной лингвистики. М.: МГПИИЯ им. М.Тореза, 1970. – 225 с.

[4] Лингвистические вопросы алгоритмической обработки сообщения. М., 1983. – 45 с.

[5] Марчук Ю.Н. Основы компьютерной лингвистики. Учебное пособие. – М. 2000. – 156 с.

[6] Белоногов Г.Г., Новоселов А.П. Автоматизация процессов накопления, поиска и обобщения информации. – М., 1979. – С.165.

[7] Зеленков Ю.Г. Морфологический анализ в системах автоматической обработки научно-технической информации: канд. дис. – М.: ВИНИТИ, 1988. – С.145.