

Л.Т. Алимтаева* , Д.О. Токмырзаев , К.Қ. Пірманова 

Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

*e-mail: alimtayeva.lazzat@gmail.com

ТҮРКІ ТІЛДЕРІНІҢ СИНТАКСИСТІК БЕЛГІЛЕНІМ ТӘЖІРИБЕСІ

Мақалада түркі тілдерінің синтаксистік құрылымын формалды грамматика тұрғысынан және заманауи аннотациялық модельдер негізінде сипаттаудың тәжірибесі қарастырылады. Синтаксистік аннотация тілдің грамматикалық жүйесін формалды түрде сипаттайтын және оны автоматты өңдеуге мүмкіндік беретін маңызды құрал ретінде танылады. Зерттеу барысында «Universal Dependencies» (UD), «MaTT» (Multilingual Aligned Treebank of Turkic) және «Kazakh Dependency Treebank» (KazDT) сияқты жобаларға сүйеніп, түркі тілдеріне тән морфологиялық және синтаксистік ерекшеліктер сипатталды. Синтаксистік белгіленім модельдері: «құрамдық», «аралас», «басыңқы-бағыныңқылық грамматикасы» т.б. тәсілдердің сипаты, ерекшеліктері, түркі тілдері үшін ұтымды тұстары мен кемшіліктері сараланды. Нәтижесінде басыңқы-бағыныңқы қатынастар грамматикасы негізінде жасалған синтаксистік аннотация моделі түркі тілінің құрылымын тиімді сипаттауға мүмкіндік беретіні дәлелденді. Басыңқы-бағыныңқы грамматикасының (басыңқы-бағыныңқы қатынастар) теориялық негіздері, синтаксистік аннотацияның форматы мен стандарттары сараланды. Түркі тілдерінің жалғамалы табиғаты мен еркін сөз тәртібінің «UD» сияқты әмбебап жобаларға бейімделуі талдауға түсті.

Сонымен қатар қазақ тілінің аннотацияланған корпустарын жетілдіру, автоматты парсинг, тілдік білім беру жүйесіне енгізу секілді болашақтағы бағыттары көрсетілді. Мақала түркі тілдерінің синтаксистік белгіленім тәжірибесі негізінде қазақ тілін цифрлық кеңістікке енгізудің маңызды қадамдарының бірі ретінде синтаксистік аннотацияны ғылыми тұрғыда негіздеуді мақсат етті.

Түйін сөздер: түркі тілдері, синтаксистік аннотация, басыңқы-бағыныңқы грамматикасы, «UD», KazDT, формалды модельдер, парсинг.

L.T. Alimtayeva*, D.O. Tokmyrzaev, K.K. Pirmanova

Al-Farabi Kazakh National University, Almaty, Kazakhstan

*e-mail: alimtayeva.lazzat@gmail.com

Experience of Syntactic Annotation of Turkic Languages

The article examines the experience of describing the syntactic structure of Turkic languages from the perspective of formal grammar and based on modern annotation models. Syntactic annotation is recognized as an important tool for formally describing a language's grammatical system and enabling its automatic processing. Relying on projects such as Universal Dependencies (UD), MaTT (Multilingual Aligned Treebank of Turkic), and the Kazakh Dependency Treebank (KazDT), the study describes morphological and syntactic features characteristic of Turkic languages. Various syntactic annotation models – “phrase structure grammar,” “hybrid,” and “dependency grammar,” among others are analyzed in terms of their characteristics, differences, as well as their advantages and disadvantages for Turkic languages. The results demonstrate that a syntactic annotation model based on head-dependent relations grammar provides an effective means of describing the structure of Turkic languages. The article outlines the theoretical foundations of dependency grammar (head-dependent relations), and the formats and standards for syntactic annotation. It also discusses the adaptation of the agglutinative nature and free word order of Turkic languages to universal projects such as UD.

In addition, future directions are identified, including the enhancement of annotated corpora for the Kazakh language, automatic parsing, and integration into language education systems. The article aims to scientifically substantiate syntactic annotation as one of the key steps in integrating the Kazakh language into the digital space, building on the syntactic annotation experience of Turkic languages.

Keywords: Turkic languages, syntactic annotation, dependency grammar (head-dependent relations), UD, KazDT, formal models, parsing.

А.Т. Алимтаева*, Д.О. Токмырзаев, К.К. Пирманова
 Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
 *e-mail: alimtayeva.lazzat@gmail.com

Опыт синтаксической разметки тюркских языков

В статье рассматривается опыт описания синтаксической структуры тюркских языков с точки зрения формальной грамматики и на основе современных аннотационных моделей. Синтаксическая аннотация признаётся важным инструментом, позволяющим формально описать грамматическую систему языка и обеспечивающим возможность её автоматической обработки. В ходе исследования, опираясь на проекты «Universal Dependencies» (UD), «MaTT» (Multilingual Aligned Treebank of Turkic) и «Kazakh Dependency Treebank» (KazDT), были описаны морфологические и синтаксические особенности, характерные для тюркских языков. Были проанализированы модели синтаксической разметки: «грамматика фразовой структуры», «смешанная» и «грамматика зависимостей» и др., их характеристики, различия, а также преимущества и недостатки для тюркских языков. В результате было доказано, что модель синтаксической аннотации, основанная на грамматике главно-зависимых отношений, позволяет эффективно описывать структуру тюркских языков. Описаны теоретические основы грамматики зависимостей (главно-зависимых отношений), форматы и стандарты синтаксической аннотации. Рассмотрена адаптация агглютинативной природы и свободного порядка слов тюркских языков к универсальным проектам, таким как «UD».

Кроме того, обозначены будущие направления, такие как совершенствование размеченных корпусов казахского языка, автоматический парсинг, интеграция в системы языкового образования. Статья ставит целью научное обоснование синтаксической аннотации как одного из важных шагов по введению казахского языка в цифровое пространство на основе опыта синтаксической разметки тюркских языков.

Ключевые слова: тюркские языки, синтаксическая аннотация, грамматика зависимостей, «UD», «KazDT», формальные модели, парсинг.

Кіріспе

XXI ғасырдағы тіл білімінің басым бағыттарының бірі табиғи тілді өңдеу (Natural Language Processing – NLP) саласының дамуына ықпал ететін формалды синтаксистік белгіленіммен құрылымдалған тілдік ресурстарды жасау. Сөз тәртібі біршама еркін, морфологиясы бай, жалғамалы сипаттағы қазақ тілі үшін синтаксистік құрылымды формалдау, көптілді модельдермен сәйкестендіру және автоматты өңдеуге бейімдеу – әлі де өзекті мәселелердің бірі.

Қазір синтаксистік белгіленімге арналған әмбебап формат ретіндегі халықаралық ауқымдағы «Universal Dependencies» (UD) жобасы кең қолданысқа ие (Universal Dependencies). Алайда бұл жүйе үнді-еуропа тілдеріне ыңғайланғандықтан, «UD» шеңберінде түркі тілдерінің, соның ішінде қазақ тілінің синтаксистік ерекшеліктерін белгілеуде әлі де шешімін таппаған мәселелер баршылық. Қазақ тіліндегі сөйлем мүшелерінің орын тәртібінің еркіндігі, жалғаулардың полисемантикалы сипаты (мәнмәтіндегі), есімді және етістікті тіркестердің лексика-грамматикалық күрделілігі – осы мәселелер қатарында.

Қазақ тіліне арналған «UD» платформасындағы алғашқы синтаксистік корпус – «Kazakh Dependency Treebank» жобасы (Makazhanov et

al., 2015; Universal Dependencies). Бұл жоба қазақ тілін цифрландыру және тілдік деректерді формалдау бағытында маңызды бастама болғанымен, оның құрылымы мен аннотациялау қағидаттары жетілдіруді қажет етеді. Сонымен қатар түркі тілдеріне ортақ морфосинтаксистік ерекшеліктерді қамтыған «Turkish Treebank» (Atalay, 2003), UniTurk (Sharipbay, 2018: 78-87) «Tatar Treebank» (Taguchi et al., 2022) тәрізді жобалар қазақ тілінің «UD» негізінде аннотациялануына әдіснамалық тұрғыдан үлгі бола алады.

Бүгінгі күні «Қазақ тілінің ұлттық корпусында» толыққанды синтаксистік белгіленім мен формалды модельдерге негізделген құрылымдық тіл ресурстарының болмауы – осы бағыттағы зерттеулердің маңыздылығын арттыра түседі. Синтаксистік аннотация тек лингвистикалық талдау үшін ғана емес, қазақ тілін цифрлық кеңістікте дамыту, тілдік технологиялар жүйесінде қолдану үшін де негізгі құралға айналуға мүмкіндік береді. Мәтінді автоматты талдау, нейрожелімен оқыту, қазақ тілінде іздеу жүйелері мен ақпараттық сүзгілер жасау – осының барлығы синтаксистік құрылымды формалды түрде сипаттауды қажет етеді.

Қазақ тілінің синтаксистік табиғатына сай келетін аннотациялық модельдер мен оларды «UD» шеңберінде бейімдеу үшін «MaTT»

(Meaning and Text Theory) және «Universal Dependencies» жүйелеріне сүйене отырып, қазақ, орыс және ағылшын тілдерінің белгіленімдерін салыстырмалы-типологиялық аспектіде сараптау маңызды болып табылады. Бұл қазақ тілінің синтаксистік құрылымдарын дұрыс аннотациялауға мүмкіндік береді.

Қазіргі тіл білімінде тілдік деректердің құрылымдық негізін жүйелі түрде сипаттау мен өңдеу мәселесі өзекті ғылыми-тәжірибелік міндеттердің бірі болып саналады. Соңғы он жылдықтарда компьютерлік лингвистика мен корпустық тіл білімінің серпінді дамуы нәтижесінде табиғи тілдердің синтаксистік құрылымын формалды модельдеу мен автоматты талдау тәсілдеріне деген қызығушылық артып отыр. Бұл үрдістер тілдің терең құрылымын, грамматикалық заңдылықтарын және олардың қолданыстағы көріністерін формалды деңгейде сипаттауға негізделеді.

Соның ішінде синтаксистік белгіленім – тіл бірліктерінің сөйлем ішіндегі қызметі мен байланысын белгілеу – көкейкесті мәселелердің бірі. Синтаксистік аннотацияланған корпустар (ағашбанктер) табиғи тілді өңдеудің көптеген қосымшаларында: автоматты аударма, сұранысты өңдеу, сөйлеуді тану, мәтінді синтездеу, грамматикалық талдаушыларды оқытуда кеңінен қолданылады. Бұл ресурстар тілдік бірліктердің морфологиялық және синтаксистік ерекшеліктерін құрылымдық және формалды түрде сипаттайтын маңызды құрал болып табылады.

Ағылшын, неміс, француз сияқты тілдер үшін мұндай ресурстар әлдеқашан қалыптасып, жетілдірілген, алайда ерекше құрылымға ие, жалғамалы түркі тілдері үшін бұл бағыттағы зерттеулер мен тәжірибелер соңғы он жылдықтардың еншісінде. Түрік, татар, башқұрт және қазақ тілдерінің алғашқы синтаксистік белгіленімдері «UD» платформасында әзірленді. Соның ішінде қазақ тілінің синтаксистік белгіленіміне арналған жобалар – отандық тіл біліміндегі жаңа ғылыми-әдістемелік бағыттардың бірі ретінде танылады.

Материалдар мен әдістер

Талдау нысаны ретінде «Turkish Universal Dependencies Treebank», «Kazakh Dependency Treebank» (KazDT) корпустары алынды. «Turkish Universal Dependencies Treebank» корпусы түрік тілінің синтаксистік құрылымын «UD» стандарттарына сәйкес белгілеу негізінде Анкара университеті мен Google Research ынтымақ-

тастығымен әзірленген. Ал «KazDT» корпусы Назарбаев университеті жанындағы «National Laboratory Astana» зертханасында жүзеге асырылып, «Universal Dependencies» 2.0 стандарты негізінде аннотацияланған. Корпустың соңғы нұсқасы бірнеше мың сөйлемнен тұрады және әрбір сөйлемнің морфологиялық, лексикалық және синтаксистік белгіленімдері жасалған. Түрік тілінің жалғамалы сипаты, біршама еркін сөз тәртібі және көмекші сөздер жүйесі синтаксистік белгіленім жасауда күрделі формалды шешімдерді талап ететін мәселелер. «KazDT» құрамында морфологиялық, лексикалық және синтаксистік белгіленімі жасалған шамамен 61 000 сөйлем бар. Зерттеу барысында осы корпустар құрылымы мен белгілеу жүйесі, көрінісі сараланды.

Салыстырмалы-типологиялық әдіс – «UD» форматындағы басқа түркі тілдерінің (түрік, татар, тува, хакас) синтаксистік ағашбанктерімен салыстыру негізінде қазақ тілінің ерекшеліктерін айқындау; KazDT корпусындағы формалды синтаксистік құрылымдарды сипаттап, олардың «UD» стандарттарына сәйкестігін анықтау сияқты тәсілдер әдістемелік негіз ретінде басшылыққа алынды.

Түркі тілдеріне ортақ «UD» нұсқаулықтары мен олардың қазақ тіліне бейімделу аспектілері де әдістемелік тұрғыдан негізделіп отыр. Бұл әдістер қазақ тіліндегі күрделі синтаксистік қатынастарды дәлірек сипаттап, саралап олардың аннотациялау тәжірибесін біріздендіруге мүмкіндік береді. Сонымен қатар зерттеу барысында «Turkish IMST Treebank», «BOUN Treebank» (BOUN Treebank) секілді қосымша ресурстар қарастырылып, аннотация тәжірибесіндегі ерекшеліктері мен «UD» модельдеріне бейімделу үлгілері зерделенді. «UD» жүйесінде түрік тілінің жалғамалы құрылымына сәйкес case, obl, nmod, advcl, aux сияқты қатынастардың қолданылуы, оларды сөз тәртібіне тәуелсіз түрде сипаттау жолдары көрсетілді.

Әдебиеттерге шолу

Түркі тілдерінің синтаксистік белгіленімін жасау және формалды грамматикаға бейімдеу мәселесі қазіргі тіл ғылымындағы өзекті мәселелердің бірі. Бұл салада түрік тілінің синтаксистік құрылымы терең сипатталып, басыңқы-бағыныңқы қатынастар грамматикасы бойынша үйлестірілген алғашқы жүйелердің бірі ретінде – «METU-Sabancı Turkish Treebank» жобасын атай

аламыз. Аталған жоба негізінде құрылымдық бірліктердің белгіленімдерінің көп деңгейлі үлгісі жасалып, жобаның өзі кейінгі түркі тілдеріндегі синтаксистік ағашбанктерді (Treebank – нақты тілдік деректердің (корпустың) синтаксистік құрылымдарын белгілеу нәтижесінде жасалған **мәліметтер қоры**) жасау тәжірибесінде теориялық және әдістемелік база қызметін атқарды. Жобаның ғылыми негізі туралы Бурга Октемнің зерттеулерінен таныса аламыз (Burga, Öktem, Wanner, 2017: 32-41).

А. Мақажанов, У.А. Тукеевтердің жетекшілігімен жүргізілген зерттеулер – қазақ тілінің синтаксистік белгіленімін «UD» жобасы шеңберінде жүзеге асыруға бағытталған маңызды еңбектердің бірі. Онда қазақ тілінің жалғамалы сипаты «UD» форматына бейімделіп, тиісті синтаксистік қатынастар жүйесі ұсынылды. «Kazakh Dependency Treebank» корпусы аталған зерттеу нәтижелеріне негізделіп жасалды (Makazhanov, Sultangazina, Makhambetov, 2015: 338-350).

Tyers F.M., Washington J. сынды ғалымдар түркі тілдеріне тән құрылымдық ерекшеліктерді белгілеудің принциптері нақтыланған «UD» нұсқаулығын ұсынды (Tyers, Washington, 2017: 276-297). Бұл нұсқаулық морфо-синтаксистік ерекшеліктерді, сөз тіркестерінің байланысын және сөзтұлғалық ерекшеліктерді біршама дұрыс белгілеуге негіз болды.

Hakkani-Tür D., Oflazer K., Tür G. еңбектерінде түркі тілдеріне ортақ морфологиялық проблемалардың алдын алуға арналған статистикалық модельдер ұсынылып, белгіленім сапасын арттыру жолдары көрсетілді (Hakkani-Tür, Oflazer, 2002: 381-410).

«KazDT» жобасында А. Шарипбаевтың жетекшілігімен әзірленген «UniTurk» метатілі қолданылған. Түркі тілдерінің морфологиялық бірліктерін біріздендіру мақсатымен әзірленген бұл ережелер қазақ тілін аннотациялау барысында кеңінен қолданыс тапты (Sharipbay, 2018: 78-87).

Көп тілділік және код-ауысым мәселелерін ескерген С. Taguchi бастаған ғалымдар тобы татар тілінің «NMCTT» синтаксистік ағашбанкін әзірлеу барысында ішкі сөздік ауысым жағдайындағы аннотациялау жолдарын сипаттайды (Taguchi, Iwata, Taro, 2022: 95-104).

Мұндай тәсілдер қазақ тіліндегі аралас стильдегі мәтіндердің белгіленімін жасауда да бағдар бола алады.

Түрік тіліндегі «UD» жобасының аннотациялау тәжірибесі Atalay еңбегінде егжей-тегжейлі баяндалса (Atalay, 2003: 15-29), кейінгі жылдары

U. Türk зерттеулері бойынша осы ағашбанктің белгіленімін жетілдіру бағытында зерттеулер жүргізіліп, синтаксистік қатынас түрлері мен құрылымын дәл белгілеу қолға алынған.

Жоғарыда аталған еңбектер түркі тілдерінің синтаксистік аннотациясын жүйелеуде теориялық тірек, әдіснамалық бағдар және практикалық үлгі ретінде маңызды болып саналады.

Нәтижелер мен талқылау

Тіл деңгейлері бойынша түркі тілдерін аннотациялау кешенді түрде XX ғасырдың соңғы ширегінде басталды. Алғашқыда түркі тілдерінің морфологиялық белгіленімін автоматтандыру, олардың фонетика-орфографиялық ережелерін формалды сипаттау бағытында маңызды зерттеулер жүргізілді. Алайда синтаксистік белгіленім (яғни сөйлем құрылымының ағаш түріндегі бейнесін жасау) басқа деңгейлермен салыстырғанда кештеу қолға алынды.

Бұл тілдердің ішінде бай корпустық тәжірибесі бар тіл ретінде түрік тілін атай аламыз. Себебі бұл елде синтаксистік талдама мен компьютерлік өңдеу мәселелері ертерек қолға алынды. Түрік тілінде 1990 жылдардың соңынан бастап статистикалық және ережеге негізделген парсерлер (белгілі бір грамматика моделінің ережелеріне негізделген сөйлемнің құрылымын автоматты түрде талдау үдерісі) әзірленіп, синтаксистік талдау үшін қажетті корпустар жинақталды. Олардың ғылыми негіздері туралы түрік ғалымдары тарапынан бірнеше мақалалар жарияланды (Hakkani-Tür, Oflazer, Tür, 2002: 381-410).

Ал өзге түркі тілдері бойынша ұзақ уақыт бойы морфологиялық талдау құралдарына көңіл бөлініп, синтаксистік белгіленімдері кейінірек әзірленді. Мысалы, қазақ, өзбек, татар, қырғыз тілдерінің 2010 жылдардың ортасына дейін толыққанды синтаксистік белгіленімі жасалған корпустары болған жоқ. Тек соңғы он жылдың көлемінде ғана осы тілдер бойынша біртіндеп синтаксистік белгіленім ағашбанктерін әзірлеу жұмыстары қолға алына бастады. Жалпы алғанда, түркі тілдеріндегі синтаксистік белгіленімнің қалыптасуы морфологиялық аннотациялау дәстүрінің ізімен жүріп, «UD» сияқты ортақ платформалардың ықпалымен қарқын алуда.

XX ғасыр соңына қарай түрік тілінің ғалымдары басыңқы-бағыныңқы грамматикасы тәсіліне назар аудара бастады. Осы орайда түрік тілі корпус-аннотациясына Игорь Мельчуктың «мағына-мәтін теориясы» (Meaning ↔ Text Theory)

негіз болған бағыныңқылық грамматикасы басшылыққа алынды (Mel'čuk, 1988: 428).

Бұл синтаксистік талдау моделінің авторлары – Карнеги-Меллон университетінің ғалымдары Д. Темперли (Davy Temperley), Д. Слитор (Daniel Sleator) және Дж. Лаферти (John Lafferty). Аталған синтаксистік модель сөйлемдегі сөз жұптарының арасындағы байланыстарды анықтайды, бірақ дәстүрлі синтаксистік талдау модельдерімен салыстырғанда, оларды толық ағашбанк түрінде бейнелеу көзделмейді, тек нақты сөздер арасындағы байланыстар көрсетіледі. Бұл талдау моделінің негізгі өлшемшарттары – байланыстардың бағыты және сөз жұптарының арақашықтығы (Sleator, Temperley, 1991: 91-196). Түрік тілінің морфосинтаксистік ерекшеліктері ескеріле отырып, синтаксистік белгіленімін жасау барысында бағыныңқылық грамматикасы тандалды, яғни сөздердің синтаксистік байланысы морфологиялық қосымшалар арқылы танылатындықтан, ағылшын тілі тәрізді қатаң сөз тәртібі жоқ тілдерде синтаксистік қатынастарды басыңқы-бағыныңқы байланыстар сызбасы арқылы белгілеу тиімдірек саналды. Дегенмен алғашқы синтаксистік корпустар кемшіліксіз болды деп айта алмаймыз.

Түрік тілі үшін алғашқы синтаксистік ағашқұрылысты корпустар 2000 жылдары әзірленді. Олардың алғашқысы – 7262 сөйлемнен тұратын «METU-Sabancı» түрік ағашбанкi, онда әр сөйлемнің морфологиялық және синтаксистік құрылымы қолмен белгіленді. Бұл – түркі тілдері бойынша ең ірі синтаксистік корпус болды. Кейінірек 2016 жылы аталған корпусты жетілдіру мақсатымен «İTÜ-METU-Sabancı Treebank» (IMST) нұсқасы ұсынылды (Burga, Öktem, Wanner, 2017: 32-41). Онда алдыңғы нұсқадағы кемшіліктер түзетіліп, жаңа белгіленім үлгі-сызбасы енгізілді. Нәтижесінде түрік тілінің синтаксистік белгіленімі біріздендіріліп, ағашбанк деректері жаңартылды.

«Universal Dependencies» халықаралық жобасы аясында түркі тілдерінің синтаксистік белгіленімі жетіле түсті. 2014-2017 жылдар аралығында «UD» бастамасымен әлем тілдері бойынша бірыңғай белгіленім стандарттары әзірленіп, соларға сәйкес корпустар жасала бастады. 2017 жылы «UD» 2.0 нұсқасында алғаш рет үш бірдей түркі тілдері – түрік, қазақ және ұйғыр тілдерінің синтаксистік аннотацияланған корпустары жарияланды.

Бұл корпустарда сөз таптары, морфологиялық ерекшеліктері және сөздер арасындағы ба-

сыңқы-бағыныңқы қатынастар бірыңғай «UD» стандарты бойынша белгіленді. Зерттеушілердің айтуынша, аталған тілдердің белгіленімдерін жасау барысында «UD-ның» әмбебап нұсқаулығы басшылыққа алынғанымен, тілдердің типтік ерекшеліктеріне орай, кейбір сәйкессіздіктер орын алды (Makazhanov, Sultangazina, Makhambetov and Yessenbayev, 2015: 338-350). Алғашқы кезеңде түрік, қазақ, ұйғыр тілдерінің корпустарының белгіленім стандарттары толық болмады, дегенмен бұл – түркі тілдерінің синтаксистік белгіленімінің дамуының бір белесі еді.

Түркі тілдерінің синтаксистік белгіленімін жасау бағытында бірнеше ғылыми мектептер мен жобалар қалыптасты. Ғалымдар екі басты мәселеге назар аударды: біріншіден, түркі тілдеріне ортақ аннотация стандартын жасау, екіншіден, тиімді құралдар мен алгоритмдер әзірлеу (мағыналық сәйкестік, парсинг өнімділігі, т.б.)

2017 жылы Ф.М. Тайерс, Дж. Вашингтон сынды ғалымдардың жетекшілігімен түркі тілдеріне арналған «UD» нұсқаулығы әзірленді (Tyers, Washington, 2017: 276-297). Авторлар бұл тілдердің морфологиялық жүйесі мен синтаксистік құрылысындағы ортақ ерекшеліктерді ескере отырып, «UD» нұсқаулығымен барынша сәйкестендіруге тырысты. Нәтижесінде «UD» ағашбанктерінде түркі тілдерінің негізгі басыңқы-бағыныңқы қатынастары (мысалы, бастауыш – баяндауыш байланысы, толықтауыштық, анықтауыштық, пысықтауыштық қатынас, т.б.) ұқсас қағидалар бойынша белгіленетін болды.

Түрлі ғылыми орталықтарда түркі тілдері бойынша жеке синтаксистік аннотация жобалары жүзеге асырыла бастады. Мысалы «BOUN Treebank» (Боғазичи университеті, Түркия) және «Turkish Treebank» (ITU & METU) сынды жобаларда стильдері әртүрлі (веб мәтіндер, жаңалықтар, т.б.) сөйлемдер қолмен белгіленді. Бұл жобалар түрік тіліндегі бірнеше тәуелсіз синтаксистік ағашбанктердің пайда болуына себеп болды. Кейіннен олар «UD» стандартына автоматты және жартылай автоматты түрде көшіріліп, қазіргі таңда «UD Turkish» деген атаумен бірқатар біріктірілген (жалпы көлемі 733 мың сөзден астам) корпустардан тұрады.

Қазақстан мен Татарстанда түркі тілдерінің ортақ метатілі мен онтологиясын құру мақсатында «UniTurk» атты бастама көтерілді. А.А. Шәріпбайдың жетекшілігімен түркі тілдеріндегі морфологиялық белгіленім модельдері ұсынылды (Sharipbay, 2018: 78-87). Мұндай бас-

тамалар түркі тілдерін бір платформада өңдеп, көп тілді құралдар жасауға септігін тигізді.

Қазақ тілі корпустары бойынша да шартты белгілердің бірізділігі негізгі мәселе болып отыр. Мемлекеттік тілдің және түркі тілдерінің корпустарында немесе түркі тілдерінің метатілін жасауды мақсат тұтқан отандық компьютер лингвистикасы саласындағы мамандар ұсынып жүрген морфологиялық шартты белгілерде бірізділік жоқ немесе әлі де толықтыруларды қажет ететін мәселелері баршылық (Абдикарим, 2023: 8-21).

Беріге дейін татар, башқұрт секілді кейбір тілдердің синтаксистік ағашбанктері болмады. 2022 жылы жапондық ғалымдар «UD» жобасы аясында татар тілінің алғашқы ағашбанк жобасын әзірлеп, теориялық негіздерін саралады (Taguchi, Iwata, Tago, 2022: 95-104). Бұл корпус татар және орыс тілдерінің код-ауысымдық (аралас) қолданыстарын да белгілейтін ерекше үлгісызбамен толықтырылды. Ғалымдар мәтіндердегі орыс тілінен енген сөздерді және морфема деңгейіндегі код-ауысымды құбылыстарды арнайы белгілермен таңбалап, сол арқылы татар тілінің синтаксистік құрылымын дәлірек көрсетуді мақсат етті. Мұндай жобалар түркі тілдерін аннотациялау кезінде қостілділік, диалектілік ерекшеліктерді көрсетуде де үлгі бола алады.

«UD»-да жалпы түркі тілдерінің синтаксистік белгіленімінің ортақ стандартын жасау (үйлесімді аннотация метатілі мен ережелері) – ең өзекті мәселелердің бірі. Себебі жалғамалы тілдердің ерекшеліктерін ескеретін бірізді анықтамалық болмайынша, әр тілге бөлек, жеке ереже жасау машақаты көп мәселелердің бірі болып табылады. Қазіргі кезде «UD» аясында құрылған «Turkic UD Working Group»-тың (түркі тілдері «UD» жұмыс тобы) жұмысы осы бағытқа негізделген. Бұл топ түркі тілдерінің басыңқы-бағыныңқы байланыстарының айырмашылықтарын талдап, оларды барынша үйлестіруге тырысуда. Сонымен қатар, ғалымдар кейбір дереккөзі аз тілдер үшін синтаксистік талдаушыларды (парсерлерді) жасау барысында трансферлік оқыту, мультимодельді нейрондық желілерді қолдану сияқты бағыттар бойынша зерттеулер жүргізуде. Бұл түркі тілдерінің синтаксистік белгіленімін дамытудың жаңа ғылыми кезеңі қалыптасқанының белгісі.

Синтаксистік белгіленім жасау барысында тілдердің типтік ерекшелігіне орай оңтайлы модель таңдау маңызды мәселелердің бірі болып табылады. 1. «Құрамдық грамматикасы» моделі:

өзара сатылы байланысқан фразалардан құралады, сөз тіркестерінің басыңқы сыңарының сөз табы ескеріліп, есімді, етістікті деп ажыратылып көрсетіледі, мұнда сөйлемнің фразалық құрылымы толық қамтылғанымен, шұбалыңқы сөйлемдердің сатылы байланысын көрсету біршама қиындау мәселелердің бірі, түркі тілдеріндегі сөйлем мүшелерінің орын тәртібінің біршама еркіндігі де фразалық бұталарды тым күрделендіреді. 2. «Аралас» модель: мұнда құрамдық және бағыныңқылық грамматика әдістері тоғыстырылады, яғни сөздер арасындағы бағыныңқы байланыспен қатар әр бір синтаксистік қатынасқа құрылған сөз тіркестерінің арасындағы байланыс та қамтылады, белгіленім ақпараты толық болғанымен, ережелері күрделі және уақыт тұрғысынан тиімсіз болып келеді. 3. Басыңқы-бағыныңқы грамматикасы: сөз тәртібі біршама еркін болса да, тәуелділік сызбасы арқылы байланыстар дұрыс көрсету, күрделі тіркестердегі бағыныңқылықты икемді граф арқылы бейнелеу мүмкіндігі бар. Негізгі предикат мәніндегі сөздің сөйлемнің соңында келуі басқа модельдер үшін қолайсыз болса, бұл модель үшін оңтайлы болып табылады. Сондықтан аннотация мақсатына орай, егер сөздердің арасындағы басыңқы-бағыныңқы қатынастарды көрсету маңызды болса – бағыныңқылық грамматикасы; фразалық топтар мен сатылы байланыс маңызды болса – «құрамдық грамматика»; семантикалық ролдер маңызды болса – «функционалды-семантикалық құрылым»; көп деңгейлі белгіленім маңызды болса – «аралас модельдерді» *басшылыққа алған жөн*.

У.А. Тукаев бастаған ғалымдар тобы да машиналық аударма үшін синтаксистік аннотацияны бағыныңқылық грамматикасы әдісі бойынша жасау анағұрлым нәтижелі болатынын дәлелдейді (Тукаев, Жуманов, Картбаев).

Link Grammar Parser – синтаксический анализатор, разработанный Дэниелом Слитором и Дэви Темперли, который основан на грамматике связей. Анализируя предложение, система определяет синтаксическую структуру, которая состоит из множества маркированных звеньев, соединяющих пары слов. Основная идея грамматики связей позволяет наравне с синтаксической структурой предложения работать и с морфологией (Бакиева, Батура, Еримбетова, Митьковская, Семенова, 2016: 5-14)

Синтаксистік белгіленім жасау барысында түркі тілдерінің ортақ типологиялық ерекшеліктерін ескерусіз қалдыруға болмайды. Ең ал-

дымен, түркі тілдері – жалғамалы тілдер. Салыстырмалы түрде сөздердің морфологиялық құрылымы күрделі, грамматикалық қосымшалар бір сөзге тізбектеле жалғанады. Мысалы, түрік тіліндегі бір ғана етістікке 5-6 морфема қосылып, ол сөйлемге тең ақпаратты бере алады, сондықтан морфологиялық белгіленімнің нақтылығы – синтаксистік қатынастарды дұрыс белгілеудің алғышарты. Морфемалардың мәнмәтіндегі көп мағыналық сипатынан да туындайтын мәселелер баршылық. Бір сөздің мәнмәгін арқылы танылатын бірнеше талдау нұсқасы болуы мүмкін. Таңбалардың ықтимал үйлесімдерінің саны өте көп болғандықтан, мұндай ерекшелік статистикалық модельдер үшін белгіленім үдерісін біршама қиындатады. Сондықтан жалғамалы тілдерді аннотациялау барысында алдын-ала морфологиялық дисамбигуация (бір мағыналандыру) жүргізіледі. Мысалы «UD» жобасындағы қазақ, қырғыз, т.б. тіл корпустарында әр сөздің морфологиялық белгіленімінің қоса берілуі талап етіледі. Түркі тілдеріне арналған көптеген парсерлер біріздендірілген морфологиялық талдағыштарды пайдаланады немесе нейрондық желі негізінде мәнмәтіннен морфологиялық ерекшеліктерді үйренуге тырысады.

Екінші үлкен ерекшелік – үндіеуропа тілдеріне тән қатаң сөз тәртібімен салыстырғанда (мысалы ағылшын тіліндегі SVO – бастауыш-баяндауыш-толықтауыш реті) түркі тілдерінде сөйлем мүшелерінің орын тәртібі біршама еркін. Қазақ, қырғыз, түрік тілдерінде сөздердің базалық тәртібі SOV (бастауыш – толықтауыш – баяндауыш) түрінде болғанымен, коммуникативтік талап, мақсаттарға сай өзгеруі мүмкін. Мысалы қырғыз тіліндегі «Айдана Нүрбекті көрдү» деген сөйлемді кемінде алты түрлі тәртіпте құрауға болатыны анықталған. Солардың ішінде ең жиі кездесетіні SOV тәртібі 48%, ал сирек кездесетін орын тәртіптері OSV (8%), SVO (25%), т.б. үлестерге ие екен. Мұндай еркін сөз тәртібі синтаксистік аннотациялауды күрделендіруі мүмкін, өйткені сөздердің арасындағы грамматикалық байланыстарды анықтау үшін формалды ережеге сүйену жеткіліксіз болғандықтан, мәнмәтінді және морфологиялық белгілерді ескеру қажет болады. Осы мәселенің шешімін табу барысында басыңқы-бағыныңқы қатынастар грамматикасы тиімді құрал ретінде танылады, себебі ол – сөздердің орын тәртібін ескермей-ақ сөздердің байланысын көрсетуге икемді тәсіл. Түркі тілдеріндегі әдетте предикат мүше, баяндауыштың орын тәртібі сөйлемнің

соңы болғандықтан: сөйлем құрамындағы тұрлаусыз мүшелер онымен бағыныңқылық қатынасқа құрылады. Мұндай жағдайда да дәстүрлі құрамдық ағаштардан гөрі, dependency ережелері сөздердің тікелей байланысын көрсетуге икемді келеді.

Сонымен бірге түркі тілдерінде жиі кездесетін құрылымдардың бірі – күрделі етістіктер. Мысалы, қазақ тіліндегі «алып келе жатыр еді» деген баяндауыш құрамында негізгі етістікпен бірге бірнеше көмекші етістік бар. Олардың синтаксистік белгіленімін жасау барысында қатынас түрін анықтау – оңай мәселе емес. «UD» нұсқаулығында мұндай жағдайларда бір етістік басыңқы, қалғаны оның бағыныңқылары деп (auxiliary қатынаста) белгілеу келісілген.

Қазақ тіліндегідей түрік тілінде де негізгі етістікке көмекші етістік тіркесіп, тұтас бір ұғымды береді: мысалы, *fark etmek* – байқау, аңғару, *yardım etmek* – көмектесу, жәрдемдесу, *yola çıkmak* – жолға шығу т.б. Ағылшын тілінде бұл ұғымдар бір негізгі сөзбен берілуі мүмкін (*to notice, to help*), сондықтан «UD» аннотациясында мұндай тіркестерді күрделі етістік ретінде қарастырып, көбіне бір негізгі етістікке тіркескен compound немесе fixed қатынастарымен белгіленеді. Мысалы *fark etmek* талданғанда *fark* сөзі етістігінің құрамдас бөлігі ретінде *compound:svc* қатынасымен көрсетіледі. Сол сияқты *tamir etmek* тіркесінде *tamir* зат есімі *etmek* етістігінің мағынасын толықтыратын сөз ретінде белгіленеді. Бұл ерекшеліктер қосымша ережелер мен нұсқауларды қажет етеді.

Кейбір түркі тілдерінде (мысалы татар, башқұрт) көмекші етістік пен есім сөздің тіркесі арқылы жасалған предикат жиі кездесетіндіктен, олардың белгіленімін жасау үшін басқа да қосымша шартты белгілердің қажет болуы мүмкін. Бұл – зерттеушілер шешуге тиіс негізгі мәселелердің бірі.

Түркі тілдерінде бастауыштың түсірілуі (*pro-drop*) – кең таралған құбылыс. Есімдік бастауыштың берілуі міндетті емес, өйткені етістіктің жіктік жалғауы арқылы субъекті кім екені айқын көрінеді (мысалы, «*geldim*» – «(мен) келдім»). Синтаксистік аннотациялау барысында түсіп қалған бастауышқа арнап бөлек түйін (*node*) қоюдың керегі жоқ, онсыз да байланыс айқын. Бірақ бұл тәсіл бастауышы мүлде жоқ сөйлемдерді автоматты талдауда қиындық туғызуы мүмкін, өйткені модель субъектіні айқын ақпаратсыз өңдеуі керек болады. Қазіргі «UD» ағашбанктерінде мұндай жағдайларға арнайы Null элементтері қойылмай-

ды, орнына етістіктің морфологиялық сипатына сүйенеді. Жасырын бастауышты да белгілеу зерттеушілердің күн тәртібінде тұрған мәселелердің бірі (Makazhanov, Sultangazina, Makhambetov, 2015: 338–350).

Аталған ерекше типологиялық мәселелерден өзге: түркі тілдерінде септеулік шылаулардың көпшілігінің жеке бірлік емес, жалғау *morph* ретінде көрінуі (барыс септік, шығыс септік, т.б.). Бұл жағдайды «UD» тәрізді жалпы стандартта белгілеу үшін кейде сөздерді шартты түрде белгілеу орын алады. Сонымен қатар, кейбір тілдерде болымсыздық немесе сұраулық мағыналардың қосымшалармен берілуі (қазақ тіліндегі болымсыздық мәннің морфологиялық тәсілдері), осындай жағдайларда белгіленім ережелерін жасау үшін, түркітанушылар мен компьютерлік лингвистердің тығыз ынтымақтастығы қажет.

Түркі тілдерін синтаксистік аннотациялау барысында бірқатар практикалық мәселелер бар. Осыған байланысты зерттеушілер тарапынан түрлі әдістемелік және технологиялық шешімдер ұсынылады:

1) Ортақ аннотация нұсқаулығының жеткіліксіздігі. Түркі тілдерінің синтаксистік қатынастарын формалды белгілеудің ортақ дәстүрі болмағандықтан, алғашқы жобаларда басшылыққа алынған ережелер әртүрлі болды. Мысалы түрік тіліндегі алғашқы «METU-Sabancı» ағашбанкінде басыңқы-бағыныңқы қатынастар белгілі бір схема бойынша таңбаланса, кейін «UD» жобасына өткенде кейбір байланыстардың белгілеу тәртібі өзгерді. Сол сияқты қазақ, ұйғыр тілдерінің «UD» аннотациясын жасаған топтардың арасында да бірізділік болмады. Бұл – бір тілді екінші тілге модель ретінде қолдануда немесе салыстырмалы зерттеулерде қиындық тудыратын жайт. Бұның шешімі ретінде жоғарыда айтылған «Turkic UD» нұсқаулығы әзірленді, онда барлық түркі тілдеріне ортақ синтаксистік қатынастар (*subj, obj, obl, mod, comp* т.б.) және олардың нақты тілдердегі көріністері ережелермен бекітілді. Сонымен бірге, түркі тілдерінің өкілдері «UD» қауымдастығы аясында даулы құрылымдар мәселелері бойынша тұрақты түрде талқылаулар өткізіп тұрады. Мысалы пассив құрылымдардың агенсін (көбіне *obl:agent* деп белгілеу ұсынылған), күрделі етістіктерді (мысалы көсемшелер тіркесі) *compound ne aux* ретінде белгілеу мәселелері қарастырылып, ортақ шешімдер қабылданды.

2) Корпус көлемінің шағындығы және деректердің аздығы. Кейбір түркі тілдері үшін

аннотацияланған корпустардың көлемі әлі де мардымсыз. Мұндай деректердің жеткіліксіздігі статистикалық парсерлердің сапасын шектейді: модельдер грамматиканы толық үйреніп үлгермейді. Бұл мәселені шешімін табу үшін кросс-лингвистикалық оқыту қолданылуда, яғни үлкендеу корпусы бар туыстас тілден (көбіне түрік тілінен) алынған модельді кіші тілге бейімдеу тәсілі. Зерттеулер көрсеткендей, түрікше тренингтен өткен нейрондық парсер қазақ немесе өзбек тіліне нөлден үйретілгендерден анағұрлым жақсы нәтиже берген. Сондай-ақ, біріктірілген көп тілді модельдер құру тәсілі де сынақтан өткізілді. Бұл бағыттағы тағы бір шешім – корпус көлемін көбейту үшін автоматты аударма мәтіндерді және параллель корпустарды пайдалану. Мысалы, «UD» әзірлеушілері ағылшын мәтіндерін қазақшаға аударып, одан алынған жасанды корпус негізінде парсер сапасын жақсартуға тырысты, екінші жағынан, мұндай әдістерді қолдандудың сапасыз аударма мен тіл ерекшеліктерінің толық айқындалмауы сияқты проблемалары да жоқ емес.

3) Аннотация құралдарының жетіспеуі. Түркі тілдерінде қолмен белгілеуді оңтайландыратын арнайы бағдарламалардың саны аз. Көп жағдайда жалпыға ортақ құралдар (мысалы, «Brat», «INCEPTION» сияқты аннотация платформалары) бейімделіп қолданылады. Мұнда да кейбір платформаларда типтік ерекшеліктерге байланысты, мысалы, көлемді сөздерді тасымалдау, арнайы қаріптер (араб жазулы ұйғыр, татар мәтіндерінде) қате берілуі мүмкін. Мұның шешімі жаңа бағдарламалар жазу немесе бар құралдарға қосымша модульдер енгізу. Мысалы, 2023 жылы әзірленген өзбек тілі корпустарын белгілеу жобасында «INCEPTION» платформасына өзбек тілінің морфологиялық талдағышы жалғанып, ғалымдардың жұмысы жеңілдетілген. Сол сияқты, қазақ тілінің орфографиялық ерекшеліктерін (*ң, ғ, ө, ұ, ү, ә, қ* таңбалары) дұрыс таңбалау үшін «Brat» жүйесіне қосымша қаріптерді орнату қажет болды.

4) Тегтеуші-белгілеуші мамандардың тапшылығы. Синтаксистік белгіленім дұрыс жүзеге асырылу үшін, жоғары білікті тіл мамандары мен бағдарламашылар қажет және олардың еңбегі дұрыс қаржыландырылуы керек (білікті мамандарды немесе студенттерді тарту). Кейбір тілдердің (башқұрт, түрікмен, қарақалпақ, т.б.) әлі күнге дейін синтаксистік белгіленімдерінің болмауы – адам ресурсының жетіспеуінің салдары. Бұл мәселенің шешімін табу барысында

түркі мемлекеттері деңгейінде ортақ жобалар (консорциумдар) құру ұсынылды. Мысалы 2021 жылы түркі елдерінің бірлестігімен ұсынылған «Turkic Languages Treebank Consortium» идеясы бойынша әр тілге жауапты топ бекітіліп, «UD» стандартымен корпустар жасау жоспарланған.

5) Тілдердің типтік ерекшеліктері. Сирек кездесетін бірақ маңызды синтаксистік құбылыстар бар. Мысалы, якут (саха) тілінде синтаксистік қатынастар сөйлем мүшелерінің орын тәртібі арқылы берілуімен қатар, бұл тілдің қопармалы (бірнеше сөздің бір сөзге бірігіп кетуі) сипаты да бар екені белгілі. Чуваш тілінде басқа түркі тілдеріне ұқсамайтын синтаксистік құрылымдар бар. Осындай жағдайларда біріздендірілген «UD» ережелері жеткіліксіз болғандықтан, сол тілдерге арнайы шағын нұсқаулықтар жазылып, «UD» қауымдастығының талқылауына ұсынылады. Мысалы, 2020 ж. «UD» конференциясында якут тіліне тән басыңқы-бағыныңқы байланыстар бойынша жаңа ережелер ұсынылған болатын (масдарлық құрылымды белгілеуге арналған тег және т.б.). Олардың бірқатары «UD 2.10» нұсқасына енгізілді. Осылайша бұл жоба жан-жақты кеңейіп, әр тілдің қажетіне бейімделу үстінде.

Түркі тілдері бойынша ресурстар мен білімді орталықтандыру стратегиясы жоғарыда айтылған мәселелердің шешімі болар еді. Бар корпустарды көпшілікке қолжетімді ету, оларды өндеуге арналған код, құралдарды «GitHub» сияқты алаңдарға жүктеу тәжірибесі бұл тілдер бойынша да етек алуда. Мысалы, «UD»-ға енген ағашбанктердің барлығы «GitHub» репозиторийлерінде қолжетімді, оларды кез келген зерттеуші жүктеп алып, пайдалана алады. Бұл өзара тәжірибе алмасуды жеңілдетеді және жаңа кіші тіл корпустарының жылдам дамуына ықпал етеді.

Саралау барысында синтаксистік белгіленімі кең қамтылған тіл – түрік тілі екеніне көз

жеткіздік. Бірнеше дербес ағашбанктері жасалып, олар кейін «UD» құрамында біріктірілген, жалпы саны жүз мыңдаған сөзді құрайды. Қазақ, ұйғыр тілдері бойынша алғашқы шағын ағашбанктер 2017 жылы «UD» аясында жарық көрді (қазақша ~10 мың сөз, ұйғырша ~40 мың сөз). Әзербайжан, қырғыз, өзбек тілдерінің шағын корпустары кейінірек жасалды (2020-2025ж., бірнеше жүз сөйлем). Татар тілі де «UD» жобасына 2022 жылы қосылды (2 мыңдай сөзді шағын корпус). Ал башқұрт, түркімен сияқты әдеби ірі тілдердің әзірге синтаксистік белгіленімі жасалған корпустарының болмауы көңіл аударарлық мәселе. Сонымен қатар, түркі тілдерінің құрамындағы біраз кіші тілдер (шор, хакас, тува, қарақалпақ, гагауз т.б.) бойынша да әзірге синтаксистік аннотация жобалары қарастырылмаған.

Төменде түрік тілінен алынған күрделі құрмалас сөйлемнің «Universal Dependencies» стандарты бойынша толық синтаксистік талдауын қарастырайық. Сөйлем әдеби стильге тән, бірнеше бағыныңқы бөліктері бар:

“Kitabın raftan düştüğünü ve kapağının yırtıldığını fark eden kütüphane görevlisi, sabahın erken saatlerinde henüz kimse gelmemişken, sessiz adımlarla salona girip hasar görmüş cildi tamir etmeye koyuldu” / «Кітаптың сөреден құлағанын және қабының жыртылғанын байқап қалған кітапханашы ертесіне ешкім келмей тұрып, аяғын ақырын басып залға кіріп, бүлінген мұқабаны қалпына келтіруге кірісті».

Сөйлем құрамында үш негізгі клауз бар: бірінші – кітапханашының байқап қалуы (бағыныңқы анықтауыш – *relative clause*), екінші – ешкімнің келмеуі (мезгіл бағыныңқы), үшінші – негізгі әрекет (залға кіріп, жұмысты бастау). 1-кестеде «UD» форматын пайдаланған талдау нәтижесінің таңбалануы ұсынылған.

1-кесте – «UD» талдауының негізгі бағандары

№	Сөз (FORM)	Лемма (LEMMA)	UPOS	HEAD	DEPREL
1	Kitabın	kitap	NOUN	3	nsubj (бастауыш)
2	raftan	raf	NOUN	3	obl (пысқтауыш)
3	düştüğünü	düş	VERB	8	ccomp (нысан бағыныңқы)
4	ve	ve	CCONJ	6	cc (жалғаулық)
5	kapağının	kapak	NOUN	6	nsubj (бастауыш)
6	yırtıldığını	yırtıl	VERB	3	conj (ыңғайлас)
7	fark	fark	NOUN	8	compound (күрделі ет.)

№	Сөз (FORM)	Лемма (LEMMA)	UPOS	HEAD	DEPREL
8	eden	et	VERB	10	acl (анықтауыш бағыныңқы)
9	kütüphane	kütüphane	NOUN	10	nmod (қатысты толықтауыш)
10	görevlisi	görevli	NOUN	28	nsubj (бастауыш)
11	sabahın	sabah	NOUN	14	nmod (тәуелдік)
12	erken	erken	ADJ	14	amod (анықт.)
13	saatlerinde	saat	NOUN	16	obl (мезгіл)
14	henüz	henüz	ADV	16	advmod (үстеу)
15	kimse	kimse	PRON	16	nsubj (бастауыш)
16	gelmemişken	gel	VERB	28	advcl (мезгіл бағыныңқы)
17	sessiz	sessiz	ADJ	19	amod (анықт.)
18	adımlarla	adım	NOUN	22	obl (қалай?)
19	salona	salon	NOUN	22	obl (бағыт)
20	girip	gir	VERB	28	advcl (қимыл бағыныңқы)
21	hasar	hasar	NOUN	23	obj (толықтауыш)
22	görmüş	gör	VERB	25	acl (анықтауыш)
23	cildi	cilt	NOUN	27	obj (толықтауыш)
24	tamir	tamir	NOUN	27	compound (күрделі ет.)
25	etmeye	et	VERB	27	xcomp (көсемше)
26	koyuldu	koyul	VERB	26	ROOT (негізгі баяндауыш)
27	ROOT	—	—	—	—

Ескерту: Кесте «UD» жүйесі негізінде жасалды.

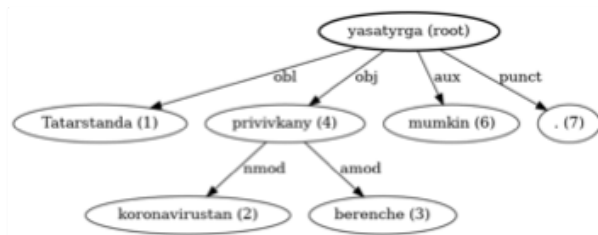
«UD» талдауының негізгі бағандары: әр сөздің ID нөмірі, оған сәйкес лемма (негізгі түбір), сөз табы (UPOS), (HEAD) және басыңқы-бағыныңқы байланыстары (DEPREL) берілген. Негізгі баяндауыш “*koyuldu*” етістігі – сөйлемнің ROOT (тірек) түйіні. Оған тікелей бағынышты үш үлкен бөлікті байқауға болады: біріншісі – бастауыш қызметіндегі “*kütüphane görevlisi*” (кітапхана қызметкері), екіншісі – мезгіл бағыныңқы “*gelmemişken*” (келмеген кезде), үшіншісі – амал бағыныңқы “*girip*” (кіріп). Бастауыштың “*görevlisi*” анықтауыш клаузасы (acl) бар, ол “*fark eden*” деген құрылым арқылы берілген («байқап қалған»). “*Fark etmek*” күрделі етістігі compound қатынасы арқылы белгіленіп, оның бағыныңқы толықтауыштары – “*düştüğünü*” және “*yırtıldığını*” – ccomp және conj ретінде берілген (екеуін *ve* жалғаулығы байланыстырып тұр). Мезгіл бағыныңқы “*gelmemişken*” ішінде өз бастауышы “*kimse*” және бірнеше пысықтауыштары (“*henüz*”, “*sabahın erken saatlerinde*”) бар. Ал “*girip*” көсемшесі негізгі әрекеттің қалай жүзеге асқанын білдіреді, оның жанында мекен және амал пысықтауыштары (“*salona*”,

“*sessiz adımlarla*”) орналасқан. Негізгі предикатқа “*koyuldu*” толықтауышы xcomp арқылы меңгерілген – “*tamir etmeye*” (жөндеуге кірісті). Бұл комплементарлы клаудың ішінде күрделі етістік “*tamir et*” compound ретінде және оның тура толықтауышы “*cildi*” obj ретінде белгіленген; “*cilt*” сөзіне “*hasar görmüş*” деген есімше анықтауыш тіркескен (бүлінген, зиян көрген деген мағынада). Осындай көпқабатты құрмалас құрылымды «UD» стандарты аясында осылайша формалды түрде көрсете аламыз.

Бұл мысалдан түрік тілінің морфологиясы мен синтаксисі өзара тығыз байланысты екендігіне көз жеткіземіз. Сөйлемнің мағыналық-құрылымдық тұлғасы қосымшалар арқылы танылып, тиісті басыңқы-бағыныңқы байланыстары көрсетілген. Басыңқы-бағыныңқы байланыстар ағашының құрылымы да көбінесе басыңқы етістік маңына топтасады (мысалы, жоғарыда *koyuldu* түйінінен көптеген доға байланыстар тарайды). Түрік тіліндегі бұл байланыстардың 95%-ы сөйлемнің негізгі мүшесінен (баяндауыш) кейін орналасатын (head-final) үлгіге жатады, мысалда берілгендей: негізгі етістік сөйлем соңында,

тәуелді сыңарлар оның алдынан кеп тіркескен. Осы тәрізді ерекшеліктерді ескеру түрік тілі корпусарын автоматты өңдеуін (парсингті) жетілдіруге септігін тигізеді.

Төменде басыңқы-бағыныңқы байланыстар ағашы түрінде белгіленген татар тіліндегі бір сөйлемнің мысалы берілді (1-сурет).



1-сурет – Татар тіліндегі сөйлемнің бағыныңқы грамматикаға негізделген синтаксистік ағашының мысалы («UD» жобасындағы «NMCTT» корпусынан алынған):
«Татарстанда коронавирустан беренче прививканы ясагырга мөмкин».

Ағаш сұлбасында әрбір сөздің басыңқымен байланысы жебе арқылы көрсетілген (Taguchi, Iwata, Taro, 2020: 95-104). Мысалы, *yasatyrga* (жасауға) етістігі – бүкіл сөйлемнің негізі (root), сондықтан ең жоғары позицияда орналасқан. Осы басты етістікке тікелей бағынатын төрт сөз бар: *privivkany* (екпені) – тура толықтауыш (*obj* қатынасы), *Tatarstanda* (Татарстанда) – мекен пысықтауыш (*obl*), *mumkin* (мүмкін/ болады) – көмекші модаль сөз (*aux*) және (*punct*) – нүкте сөйлемнің соңы. Толықтауыш «*privivkany*» екі сөздің басыңқысы: *berenche* (бірінші) – анықтауыш (*amod*), және *koronaviyruстан* (коронавирустан) – зат есімнің бағыныңқы толықтауышы (*nmod*, қандай екпе екенін нақтылайды). Осылайша басыңқы-бағыныңқы синтаксистік ағашы сөйлемдегі барлық сөздердің сатылы грамматикалық байланысын айқын көрсетеді: әрбір бұтақ өз басыңқысымен және бағыныңқысымен жұптасып, тұтас құрылымды құрайды.

Бұл мысалдан түркі тілдерін аннотациялаудағы кейбір ерекшеліктерді көре аламыз: *koronaviyruстан* сөзінің құрамында орыс тілінен енген «коронавирус» және түркі жалғауы *-tan* бар, корпус аннотациясында бұл құбылыс арнайы белгімен көрсетілген (кестеде *CSPoint=koro-naviyru\$man| LangID=MIXED[RU\$TT]* түрінде, яғни «аралас тіл сегменті» деген белгі). Бұл татар тіліндегі мәтіндерде жиі кездесетін код-ауысым жағдайын белгілеп отыру үшін енгізілген ереже-

лер. Сол сияқты *mumkin* сөзі «UD» белгілеуінде *AUX* (көмекші) ретінде таңбаланған, өйткені ол модальді мәнді (мүмкіндік) білдіреді; ал негізгі іс-әрекеттің өзі *yasatyrga* етістігімен берілген, ол *VERB* деп таңбаланды. Түркі тілдерінде модалды, шақтық көмекшілер жиі кездесетіндіктен, оларды дұрыс белгілеу – маңызды мәселелердің бірі (мысалы кей жобаларда *mumkin* сияқты сөздерді *Adj* немесе *Particle* деп те таңбалаған, бірақ «UD» стандарты бойынша ол көмекші етістікке жақын функция атқаратындықтан *AUX* деп алынған).

Мысалдардан тілдердің құрылымын түсінуге синтаксистік белгіленімнің қаншалықты ықпал ететіндігін көре аламыз. Осындай басыңқы-бағыныңқы байланыстар арқылы сөйлемнің мағыналық-құрылымдық тұлғасы айқындалып, компьютердің оны өңдеуі жеңілдейді. Баяндауыштың сөйлем соңында тұратыны, басқа мүшелер оған бағынып тармақталатыны, анықтауыштардың анықталатын сөзіне іргелес орналасатыны сияқты ерекшеліктер дәл осындай ағаштардан көрінеді. Синтаксистік ағашбанктер грамматикалық ережелерді формалды түрде тексеруге, әртүрлі тілдердің құрылымын салыстыруға мүмкіндік береді. Мысалы, қазақ және түрік тілдеріндегі толықтауыштардың орын тәртібін немесе бағыныңқы сөйлемдердің тіркесу ерекшеліктерін осы ағашбанктерден автоматты түрде іздеуге болады. Сол арқылы түркі тілдерінің грамматикалық ұқсастықтары мен айырмашылықтары туралы нақты деректер алынады.

Қорытынды

Түркі тілдерінің синтаксистік белгіленімін жасау бағытындағы жұмыстар қарқынды даму үстінде, бірақ әлі де шешімі табылмаған мәселелер баршылық. Жүзеге асырылған ағашбанктердің саны көбейіп, көлемі ұлғайған сайын, олардың ғылыми зерттеулер мен қолданбалы жобалар үшін пайдасы арта түспек. Қазіргі қол жетімді корпусардың өзінен түркі тілдерінің синтаксистік жүйесі туралы көптеген қызықты құбылыстар анықталды, алайда қамтылмаған тілдер бойынша жұмыс жандануы тиіс.

Синтаксистік құрылымдарды белгілеудің теориялық негіздерін меңгеру – қазақ тілінің грамматикалық құрылысын терең түсінуге жол ашады. Синтаксистік қатынастар мен байланыс түрлерін тану арқылы сөйлемді дұрыс құрау, тыныс белгілерін сауатты қою, ойды анық жеткізу дағдылары қалыптасады. Қазір де бұл мәселелер

әлі де өзектілігін жоя қойған жоқ: жұмсалымды грамматика, құрылымдық тіл білімі, мәтін синтаксисі салаларында синтаксистік құрылымдар жаңа әдістер, бағыттар бойынша зерттелуде. Алайда олардың бәрі классикалық синтаксистік теорияның – сөз тіркесі мен сөйлем синтаксисі ілімінің – берік іргетасына сүйенеді. Сондықтан қазақ тілінің синтаксистік құрылымдарын әрі қарай зерттеу мен белгіленімін жасауда аталған негіздер басты бағдар болмақ.

«UD» үлгісіндегі әмбебап үлгілер ғылымға ортақ платформа берсе, «SUD» сияқты баламалар теориялық дәлдік пен тілге бейімделу жағын күшейтеді. Екі бағыттың да ықпалы нәтижесінде қазіргі тіл білімі грамматикалық құрылымдардың әрі әмбебаптылығын, әрі бірегейлігін тереңінен түсінуге мүмкіндік алып отыр. Бұның практикалық пайдасы да зор – көп тілді ақпарат өңдеу жүйелері жақсарып, тілдердің цифрлық ресурстары арасында алшақтық азаяды. Ал теориялық тұрғыда, құрамдық/бағынықтылық сияқты әдістер мен платформалар тілдердің негізгі категорияларын жаңаша пайымдауға түрткі болып, олардың әмбебап және айрықша қасиеттерінің шекарасын нақтырақ анықтауға септігін тигізеді. Сондықтан белгіленім жүйелерінің типологиясын меңгеру – қазіргі тіл маманы үшін

өте маңызды: ол грамматикалық талдаудың мүмкіндіктері мен шектеулерін түсінуге, таңдаған әдістің салдарын бағамдауға жол ашады.

Қорыта айтқанда, түрік тілі синтаксистік белгіленімі бойынша жинаған тәжірибе басқа туыстас тілдер корпусын дамытуға негіз болуда. Түркі тілдерінің барлығына ортақ морфосинтаксистік қасиеттер «UD» сияқты әмбебап стандарттар арқылы бірізді белгіленіп келеді. Ал жеке тілдік ерекшеліктер әр тілдің өз корпусында нақтылай көрсетіліп, қажет болғанда «UD» стандартына ұсыныстар енгізуге сеп болады. Түрік, қазақ, өзбек, татар тілдері корпустарын салыстыра зерттеу арқылы тілдік әмбебаптар мен ерекшеліктерді айқындау мүмкіндігі артып отыр. Бұл бағыттағы жұмыстар алдағы уақытта түркі тілдерінің үлкен көлемді, белгіленімі бай корпустарын жасап, оларды тіл білімінде және жасанды зият саласында тиімді қолдануға жол ашпақ.

Зерттеуді ҚР Жоғары білім және ғылым министрлігі, ғылым комитеті қаржыландырады. Мақала ЖТН BR21882334 «Қазақ поэтикалық корпусы әзірлемесі: Абай өлеңдерінің морфологиялық және поэтикалық белгіленімі» БНҚ негізінде орындалды.

References

- Abdikarim, N. (2023). Sintaksistik-semantikalıq belgilenim zhasaıdyń alǵy sharttary [Preconditions for Developing Syntactic-Semantic Annotation]. L.N. Gumilev atyndaǵy Eýrazyalyq ulttyq ıńversitetiniń Habarshysy. Filologiya seriiasy [Bulletin of L.N. Gumilyov Eurasian National University. PHILOLOGY Series]. Vol. 3 (144), P. 8-21. <https://doi.org/10.32523/2616-678X-2023-144-3-8-21> (In Kazakh)
- Atalay, N.B., Oflazer K., Say B. (2003). The Annotation Process in the Turkish Treebank. Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003. P. 15-29. [Electronic resource]. URL: <https://aclanthology.org/W03-2405.pdf> (Date of use: 15.04.2025)
- Bakieva, A.M., Batura, T.V., Yerimbetova, A.S., Mit'kovskaia, M.V., Semenova, N.A. (2016). Issledovanie grammatiki sviazei na primere kazakhskogo i turetskogo iazykov [Investigation of the Grammar of Relations on the Example of the Kazakh and Turkish Languages]. Vestnik NGU. Seriya: Informatsionnye tekhnologii [Novosibirsk State University Bulletin. Series: Information Technology]. Vol. 3 (14). P. 5-14. (In Russian)
- BOUN Treebank. [Electronic resource]. URL: https://github.com/UniversalDependencies/UD_Turkish-BOUN (Date of use: 10.05.2025)
- Burga, A., Öktem, A., Wanner, L. (2017). Revising the METU-Sabancı Turkish Treebank: An Exercise in Surface-Syntactic Annotation of Agglutinative Languages. Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), P. 32-41, Pisa, Italy. Linköping University Electronic Press. [Electronic resource]. URL: <https://aclanthology.org/W17-6506> (Date of use: 15.04.2025)
- Sleator D., Temperley D. (1991). Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU October.
- Hakkani-Tür, D., Oflazer, K., Tür, G. (2002). Statistical Morphological Disambiguation for Agglutinative Languages. Computer Speech & Language. P. 381-410.
- KazDT. kazdet: NLA-NU Kazakh Dependency Treebank. [Electronic resource]. URL: <https://github.com/nlacs-lab/kazdet> (Date of use: 10.05.2025)
- Makazhanov, A., Sultangazina, A., Makhambetov, O., Yessenbayev, ZH. (2015). Syntactic Annotation of Kazakh: Following the «Universal Dependencies» Guidelines. A report. Proceedings of the 3rd International Conference on Turkic Languages Processing. P. 338-350.

- Mel'čuk I.A. (1988). Dependency Syntax: Theory and Practice. Albany. State University of New York Press.
- Sharipbay, A. et al. (2018). UniTurk: Unified Annotation Meta-language for Turkic Morphology. Journal of Mathematics, Mechanics, Computer Science, Vol. 4(100), P. 78-87.
- Taguchi, C., Iwata, S., Taro W. (2022). Watanabe «Universal Dependencies «Treebank for Tatar: Incorporating Intra-Word Code-Switching Information. Proceedings of the EURALI Workshop @LREC2020, P. 95-104. [Electronic resource]. URL: <https://aclanthology.org/2022.eurali-1.17.pdf>. (Date of use: 03.05.2025)
- Tukeev, U.A., Zhýmanov, Zh.M., Kartbaev, A.Zh. Kazahsko-angliiskii perevod s ispol'zovaniem grammatiki svyazi i pravil semanticheskikh situatsii [Kazakh-English Translation Based on Connection Grammar and Semantic Situation Rules]. [Electronic resource]. URL: <https://pps.kaznu.kz/ru/Main/faculty/5> (Date of use: 01.04.2025)
- Tyers, F.M., Washington, J. et al. (2017). «UD» Annotation Guidelines for Turkic Languages. Proceedings of 5th Intl. Conference on Turkic Processing. P. 276-297.
- Universal Dependencies. [Electronic resource]. URL: (https://universaldependencies.org/treebanks/tr_imst/index.html) (Date of use: 01.05.2025)

Авторлар туралы мәлімет:

Алимтаева Лаззат Тургыновна – филология ғылымдарының кандидаты, Әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы қ., e-mail: alimtayeva.lazzat@gmail.com);

Токмырзаев Дархан Онталәулы – IT бағдарламашы, Әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы қ., e-mail: dark.han@mail.ru);

Пирманова Күнсұлу Қамбарбекқызы – PhD, постдокторант, Әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы қ., e-mail: kunsulu.pirmanova@mail.ru).

Information about authors:

Alimtaeva Lazzat – Candidate of Philological Sciences, Al-Farabi Kazakh National University (Kazakhstan, Almaty, e-mail: alimtayeva.lazzat@gmail.com);

Tokmyrzaev Darkhan – Programmer, Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: dark.han@mail.ru);

Pirmanova Kunsulu – PhD, Postdoctoral Researcher, Al-Farabi Kazakh National University (Kazakhstan, Almaty, e-mail: kunsulu.pirmanova@mail.ru).

Сведения об авторах:

Алимтаева Лаззат Тургыновна – кандидат филологических наук, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: alimtayeva.lazzat@gmail.com);

Токмырзаев Дархан Онталәулы – IT-программист, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: dark.han@mail.ru);

Пирманова Күнсұлу Камбарбекқызы – PhD, постдокторант, Казахский национальный университет имени аль-Фараби (Алматы, Казахстан, e-mail: kunsulu.pirmanova@mail.ru).

Келін түсті: 15 шілде 2025 жыл
Қабылданды: 25 қыркүйек 2025 жыл