

ӘОЖ 81'42; 801.7; МҒТАР 16.21.33
<https://doi.org/10.47526/2025-4/2664-0686.274>

Л.Т. АЛИМТАЕВА¹, Ж.Б. САТКЕНОВА², С.Б. БЕЙСЕМБАЕВА³

^{1,2}филология ғылымдарының кандидаты

әл-Фараби атындағы Қазақ ұлттық университеті (Қазақстан, Алматы қ.)

¹e-mail: alymtaeva.lazzat@gmail.com; ²e-mail: jaina1978@mail.ru

³филология ғылымдарының кандидаты

Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті

(Қазақстан, Түркістан қ.), e-mail: saltanat.beysembayeva@ayu.edu.kz

СИНТАКСИСТІК БЕЛГІЛЕНІМНІҢ ТЕОРИЯЛЫҚ НЕГІЗДЕРІ

Аңдатпа. Мақалада синтаксистік белгіленімнің теориялық негіздері жан-жақты қарастырылды. Зерттеу аясында түрлі грамматикалық мектептердің синтаксистік құрылымға қатысты ұстанымдары салыстырмалы түрде талданды. Атап айтқанда, Н. Хомскийдің өзгеріс тудырушы, туындаушы теориясы (трансформациялық, генеративтік), түпкі және беткі синтаксис деңгейлері сараланып, олардың синтаксистік талдау бағдарламаларының аннотация үлгілеріне негіз болғаны көрсетілді.

Л. Теньердің бағыныңқылық грамматикасы дәстүрлі бастауыш-баяндауыш дихотомиясын жоққа шығарып, сөйлемді граф үлгісінде бейнелеуді ұсынды. Оның актанттар теориясы предикаттық етістікке негізделген модель ретінде қарастырылып, қазіргі компьютерлік тіл өңдеу тәжірибесіне ықпалы анықталды. И. Мельчуктың МаТТ теориясы синтаксистік валенттілік пен семантикалық актанттарды формалды түрде белгілеудің маңызын дәлелдеп, машиналық аударма мен компьютерлік лексикографияда кеңінен қолданылатыны аталды.

Зерттеудің мақсаты – қазақ тілін синтаксистік тұрғыдан аннотациялаудың ғылыми-теориялық базасын қалыптастыру. Бұл орайда салыстырмалы-сипаттамалы әдіс, формалды грамматика модельдері мен құрылымдық-синтаксистік талдау әдістері басшылыққа алынды. Құрамдық және бағыныңқылық грамматикалардың ерекшеліктері сараланып, қазақ тілінің грамматикалық жүйесіне бағыныңқылық грамматика моделінің икемді екендігі дәйектелді.

«Universal Dependencies» жобасындағы қазақ тілі синтаксистік белгіленімінің мүмкіндіктері мен шектеулері сараланды. «UD» үлгісі тілдерді біріздендіру бағытында тиімді платформа болғанымен, қазақ тілінің синтаксистік белгіленім тәжірибесі бір ғана мақсатпен шектеліп қалмау керек, оның феномендік ерекшеліктері айқындалатын ұлттық корпус мәтіндерінің белгіленімі жасалу керек.

Зерттеу тұжырымдары қазақ тіл біліміндегі синтаксистік аннотацияның теориялық негіздерін қалыптастыруға үлес қосып, ұлттық корпустар үшін практикалық тұрғыдан маңызды болып табылады.

* Бізге дұрыс сілтеме жасаңыз:

Алимтаева Л.Т., Саткенова Ж.Б., Бейсембаева С.Б. Синтаксистік белгіленімнің теориялық негіздері // *Ясауи университетінің хабаршысы*. – 2025. – №4 (138). – Б. 64–79. <https://doi.org/10.47526/2025-4/2664-0686.274>

*Cite us correctly:

Alimtaeva L.T., Satkenova Zh.B., Beisembayeva S.B. Sintaksistik belgilenimnin teorialyq negizderi [Theoretical Foundations of Syntactic Annotation] // *Iasau universitetinin habarshysy*. – 2025. – №4 (138). – Б. 64–79. <https://doi.org/10.47526/2025-4/2664-0686.274>

Мақаланың редакцияға түскен күні 11.09.2025 / жарияланған күні 30.12.2025

Кілт сөздер: синтаксистік белгіленім, формалды грамматика, бағыныңқылық грамматикасы, сөйлем модельдері, тіл корпусы.

L.T. Alimtaeva¹, Zh.B. Satkenova², S.B. Beissembayeva³

^{1,2}*Candidate of Philological Sciences*

Al-Farabi Kazakh National University (Kazakhstan, Almaty)

¹*e-mail: alymtaeva.lazzat@gmail.com; ²e-mail: jaina1978@mail.ru*

³*Candidate of Philological Sciences*

Khoja Akhmet Yassawi International Kazakh-Turkish University

(Kazakhstan, Turkistan), e-mail: saltanat.beysembayeva@ayu.edu.kz

Theoretical Foundations of Syntactic Annotation

Abstract. The article provides a comprehensive analysis of the theoretical foundations of syntactic annotation. The study presents a comparative overview of different grammatical schools and their approaches to syntactic structure. In particular, N. Chomsky's transformational-generative theory, the concepts of deep and surface syntax, and their role in the development of syntactic parsing programs and annotation models are examined. L. Tesnière's dependency grammar is highlighted as an alternative to the traditional subject–predicate dichotomy, proposing to represent the sentence as a graph. His theory of actants, centered on the predicate verb, is analyzed as a model that significantly influenced modern computational language processing practices. I. Mel'čuk's Meaning–Text Theory (MTT) emphasizes the importance of formally representing syntactic valency and semantic actants, and it has been widely applied in machine translation and computational lexicography.

The main aim of the research is to establish a scientific and theoretical foundation for syntactic annotation of the Kazakh language. The methodological framework includes comparative-descriptive analysis, formal grammar models, and structural-syntactic analysis. The advantages of constituency and dependency grammars are discussed, with the conclusion that the dependency grammar model is more adaptable to the grammatical system of Kazakh.

The paper also evaluates the possibilities and limitations of the Kazakh syntactic annotation within the Universal Dependencies (UD) project. While UD provides an effective platform for cross-linguistic unification, the syntactic annotation of Kazakh should not be restricted to this framework. A national corpus-based annotation is required to adequately capture the language's unique features, such as its agglutinative structure, case multifunctionality, possessive constructions, and specific properties of complex sentences.

The study contributes to strengthening the theoretical foundations of syntactic annotation in Kazakh linguistics and offers significant practical implications for the development of national corpora and further advancement of language technologies.

Keywords: syntactic annotation, formal grammar, dependency grammar, sentence models, language corpus.

Л.Т. Алимтаева¹, Ж.Б. Саткенова², С.Б. Бейсембаева³

^{1,2}*кандидат филологических наук*

Казахский национальный университет имени аль-Фараби (Казахстан, г. Алматы)

¹*e-mail: alymtaeva.lazzat@gmail.com; ²e-mail: jaina1978@mail.ru*

³*кандидат филологических наук*

Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави

(Казахстан, г. Туркестан), e-mail: saltanat.beysembayeva@ayu.edu.kz

Теоретические основы синтаксической разметки

Аннотация. В статье всесторонне рассмотрены теоретические основы синтаксической разметки. В рамках исследования проведён сравнительный анализ взглядов различных грамматических школ на синтаксическую структуру. В частности, проанализирована трансформационно-генеративная теория Н. Хомского, уровни глубинного и поверхностного синтаксиса, показано их влияние на формирование моделей аннотации и программ синтаксического анализа. Синтаксическая концепция Л. Теньера отвергла традиционное деление на подлежащее и сказуемое и предложила представлять предложение в виде графа. Его теория актантов, основанная на предикативном глаголе, оказала заметное влияние на современные практики компьютерной обработки языка. Теория MaTT И. Мельчука доказала важность формального представления синтаксической валентности и семантических актантов, что нашло широкое применение в машинном переводе и компьютерной лексикографии.

Цель исследования – формирование научно-теоретической базы синтаксической аннотации казахского языка. Для этого использованы сравнительно-описательный метод, модели формальной грамматики и структурно-синтаксический анализ. Рассмотрены преимущества составляющей и зависимой грамматик, обоснована адаптивность модели зависимой грамматики к грамматической системе казахского языка. Проанализированы возможности и ограничения синтаксической разметки казахского языка в проекте «*Universal Dependencies*». Несмотря на эффективность «UD» как платформы унификации языков, синтаксическая разметка казахского языка не должна ограничиваться лишь этим форматом: необходимо создание национального корпуса, который выявит его феноменальные особенности.

Результаты исследования способствуют формированию теоретических основ синтаксической аннотации в казахском языкознании и имеют важное практическое значение для развития национальных корпусов.

Ключевые слова: синтаксическая разметка, формальная грамматика, грамматика зависимостей, модели предложения, языковой корпус.

Кіріспе

Синтаксистік қатынас теориясының өзектілігі жаңа ғылыми технологияға негізделген тілдік корпустар құрастыру ісінде айқын байқалады. Ғылым-білімнің бүгінгі жетістігі оның барлық салаларына компьютерлік технологияның енгізілуін талап етеді. Ғылыми-зерттеу жұмыстарындағы дәстүрлі қалыптасқан материал жинауға қатысты қол жұмысы бүгінде тілдік корпустарды жасаумен байланысты автоматтандырылып, адам атқаратын жұмысты жеңілдетуге баса назар аударылуда. Бұл ретте компьютер жадына ғылыми-зерттеу жұмыстарына қажетті материалды салып, үлкен көлемді мәтіндерді лингвистикалық талапқа сәйкес өңдеп шығару қажеттілігі өзекті мәселелердің бірі.

Синтаксистік белгіленім – концептуалды-құрылымдық модельге негізделген, лингвистикалық модельдерді әзірлеудің прагматикалық бағытталған технологияларын қолдану арқылы жоғары тиімділікке қол жеткізетін функционалды компьютерлік жүйе [1, 129-б.].

Тілді синтаксистік тұрғыдан аннотациялаудың өзіне тән базалық ұғымдары, терминдері қалыптасқан. Сөйлемді формалды талдау дегеніміз – сөйлем құрылымын мазмұндық тұрғыда емес, нақты ережелер мен формальды модельдерге сүйене отырып сипаттау. Бұл тәсіл бойынша сөйлемнің грамматикалық ұйымдасуы математика-логикалық немесе алгоритмдік

үлгілер арқылы бейнеленеді, яғни оның «қалай айтылғанын» емес, «қалай құрылымданғанын» нақты ережелер арқылы көрсету.

XX ғасырдың басында тіл білімінде тілдік құрылымдарды бейнелеудің әртүрлі теориялық үлгілері ұсынылды. Солардың ішінде ерекше маңызға ие болған ұғымдардың бірі – «ағаш» моделі. Бұл терминнің қалыптасуы кездейсоқ емес, ол тілдің сатылы (иерархиялық) табиғатын көрнекі әрі жүйелі түрде сипаттауға мүмкіндік беретін сызба.

XX ғасырдың 60-жылдары американдық құрылымдық мектеп өкілдері тіл біліміне туындаушы грамматика (генеративная грамматика) (Лат. *generative* – туындаушы (пораждающая), өзгеріс тудырушы грамматика (трансформационная) [2, 75-б., 386-б.; 3, 45-б.] терминдерін енгізді. 1950–1960 жылдары Н. Хомский туындаушы грамматика теориясын қалыптастырып, сөйлемді талдаудың жаңа моделін ұсынды. Оның «*Syntactic Structures*» [4, 117-б.] және «Аспекты теории синтаксиса» [5, 117-б.] еңбектерінде құрамдық ағаш (коституенттік) үлгілері көрсетілді. Құрамдық грамматикаға (ҚГ) негізделген талдау әдісі сөйлемнің жасалу, туындау тетіктерін түсіндіруге жол ашты.

Осы кезеңде француз ғалымы Л. Теньер сөйлем құрылымын басқа қырынан түсіндіруге тырысты. Оның «*Éléments de syntaxe structurale*» (1959) атты еңбегінде бағыныңқылық грамматикасы (БГ) теориясы ұсынылды [6, 690-б.]. Л. Теньердің көзқарасы бойынша, сөйлемнің орталығы – етістік, ал басқа сөздер осы орталықтың бағыныңқылары болып табылады. Сөйлем құрылымын бейнелеу үшін «бағыныңқылық ағашы» (синтаксистік белгіленім сұлбасы) енгізілді. Бұл үлгі бойынша негізгі сөз түбір, қалған сөздер одан тараған бұтақтар түрінде бейнеленді.

1970–1990 жылдары компьютерлік лингвистика дамып, тілдің автоматты өңделуі өзекті салаға айналды. Осы кезеңде «синтаксистік ағаш» моделі кеңінен қолданыс тапты, себебі ол сөйлем құрылымын формалды және көрнекі түрде бейнелеудің ең қолайлы тәсілі еді. Парсинг алгоритмдері – құрамдық және бағыныңқылық бағытта – табиғи тілдерді автоматты талдау үшін белсенді қолданылды [7, 428-б.].

Қазіргі кезеңде бұл идеялар «Universal Dependencies» («UD») жобасында жалғасын тапты. 2014 жылы басталған бұл жоба әлем тілдеріндегі мәтіндерді ортақ стандартпен таңбалауды мақсат етті [8]. «UD» жүйесінде барлық тілдердегі сөйлемдер «бағыныңқылық ағашы» түрінде аннотацияланады: әрбір сөздің грамматикалық қызметі (бастауыш, толықтауыш, пысықтауыш және т.б.) және оның басыңқы-бағыныңқылық қатынасы нақты көрсетілді. Қазақ тілі де «UD» жобасына енгізіліп, ағаш құрылымды деректері жасалуда [9].

Осылайша, синтаксистік «ағаш» ұғымының қалыптасу тарихы Ф. Соссюрдің тіл жүйесі туралы жалпы идеяларынан бастау алып, Н. Хомскийдің құрамдық моделінде теориялық тұрғыдан дамыды, Л. Теньердің бағыныңқылық грамматикасында балама үлгі ретінде орнықты және қазіргі таңда «UD» жобасы арқылы тәжірибе жүзінде қолданылуда. «Ағаш моделі» қазіргі лингвистикада тек метафора емес, сөйлем құрылымын талдаудың әмбебап ғылыми құралы болып қалыптасты.

Қазақ тілін цифрлық кеңістікке бейімдеудің ең өзекті мәселелерінің бірі – синтаксистік аннотация жасау. Бұл сала бойынша алғашқы қадамдар жасалғанымен, атап айтқанда, «UD» платформасында белгілі бір көлемде аннотациялары бар болғанымен, ұлттық корпустарымыздағы мәтіндер базасының синтаксистік белгіленімін жасау әлі күнге дейін күн тәртібінде тұрған мәселелердің бірі. «UD» жүйесінің басты мақсаты – әлем тілдерін біріздендіру, әмбебап үлгі қалыптастыру, машиналық аударма мен автоматты өңдеуді жеңілдету. Мұндай ұстаным тиімді болғанымен, қазақ тілінің ішкі табиғатын, ұлттық-типологиялық ерекшеліктерін толық бейнелей алмайды.

Қазақ тілінің синтаксистік жүйесі өзінің ерекшелігімен дараланады: жалғамалы сипаты, меншіктілік қатынастың матасу арқылы жүзеге асуы, зат есімнің бағыныңқы мүше ретінде

қабыса байланысып атрибуттық қатынас жасауы. Күрделі, аналитикалық етістіктердің ерекше қызметтері, көлемдік септіктердің (көмектес, шығыс, барыс, т.б.) пысықтауыш қызметінде жұмсалуды – түркі оның ішінде қазақ тіліне тән айрықша ерекшеліктері болып табылады. Сонымен қатар, қазақ тілінің номинативтілік сипаты, баяндауыштың орын тәртібінің тұрақтылығы (басқа сөйлем мүшелерінің де орын тәртібі тұрақты дегенімізбен, түрлі екінші дәрежелі коммуникативтік талаптарға сай, орындарының жылжымалы болуы, эллипсистенуі), сөз таптарының сөйлем ішінде конверсиялану, оның ішінде заттанудың сөйлеуді ықшам, әсерлі, үнемді жеткізу қызметінің ерекшелігі; басқа тілдерде құрмалас құрылымдар арқылы берілетін ойдың, пікірдің қазақ тілінде жай сөйлеммен берілуі, іргелес құрмаластардың өзіндік ерекшеліктері – синтаксистік белгіленімде міндетті түрде көрініс табуға тиіс ерекшеліктер.

Осы тұрғыдан алғанда, қазақ тілінің ұлттық корпустарына негізделген синтаксистік аннотация жасау тек техникалық қажеттілік емес, сонымен қатар тілдің табиғи ерекшеліктерін ғылыми тұрғыда дәл бейнелеудің құралы болып табылады. Мұндай белгіленім қазақ тілінің типологиялық ерекшеліктерін айқындап қана қоймай, оның бай мүмкіндіктерін, құрылымдық икемділігін, ұлттық реңктерін сақтауға жол ашады. Сондықтан синтаксистік аннотация жасауда бір жағынан әлемдік тәжірибені, соның ішінде «UD» жобасының қағидаттарын игеру қажет болса, екінші жағынан қазақ тілінің ішкі заңдылықтарын терең ескерген ұлттық үлгіні қалыптастыру – бүгінгі тіл білімінің ең өзекті міндеттерінің бірі.

Зерттеу материалдары мен әдістері

Зерттеу қазақ тілін синтаксистік тұрғыдан аннотациялаудың ғылыми-теориялық негіздерін қалыптастыруға бағытталды. Зерттеу міндеттері: қазақ тілінің синтаксистік құрылымын сипаттау үшін қай ғылыми бағыттардың теориялық негіздері тиімді екенін анықтау; Н. Хомскийдің өзгеріс тудырушы-туындаушы грамматикасы, Л. Теньердің бағыныңқылық грамматикасы мен И. Мельчуктың «MaTT» теориясының қазақ тілінің синтаксисін формалды түрде бейнелеуге қаншалықты мүмкіндік беретінін сараптау; «UD» жобасының қазақ тілінің табиғатын толық қамту деңгейін бағалау; сондай-ақ қазақ тілінің феномендік ерекшеліктерін сақтап қалатын ұлттық аннотация жүйесін құрудың әдіснамалық және техникалық принциптерін айқындау.

Қазақ тілінің синтаксистік белгіленімін жасауда бағыныңқылық грамматика үлгісінің ана тіліміздің құрылымдық табиғатына барынша сәйкес келетіні, оның ұлттық корпус үшін негіз бола алатындығы – зерттеудің негізгі гипотезасы. Сонымен қатар, «UD» жобасы тілдерді халықаралық деңгейде біріздендіру тұрғысынан тиімді болғанымен, қазақ тілінің синтаксистік құрылымын толық бейнелей алмайды деген тұжырым жасалды.

Зерттеу бірнеше кезеңнен тұрды. 1. Тіл біліміндегі негізгі синтаксистік теориялар сараланып, олардың басты ұстанымдары мен айырмашылықтары талданды. Мұнда салыстырмалы-сипаттамалы әдіс пен аналитикалық ой қорыту тәсілі қолданылды: әрбір теорияның өзегін құрайтын ұғымдар айқындалып, олардың артықшылықтары мен әлсіз тұстары анықталды. 2. Аталған теориялар қазақ тілінің құрылымдық ерекшеліктерімен салыстырылып, практикалық материалдар негізінде олардың қолдану мүмкіндіктері бағаланды. Бұл тұста құрылымдық-синтаксистік талдау, формалды модельдеу және индуктивті пайымдау әдістері басшылыққа алынды. 3. «UD» жобасының қазақ тіліне арналған аннотация үлгілері сарапталып, корпустық талдау әдісі қолданылды. «UD» жобасының қазақ тіліне арналған деректер базасы зерттеу материалы ретінде қарастырылды. Мұндай материалдар жиынтығы тілдің синтаксистік құрылымын әртүрлі қырынан талдауға, формалды модельдердің қазақ тіліне сәйкестік деңгейін бағалауға мүмкіндік берді.

Талдау мен нәтижелер

Қазақ тіл білімінде алғаш синтаксистік категориялар мен сөйлем мүшелерінің анықтамаларын А. Байтұрсынұлының еңбегінен көреміз. Ғалымның сөйлемге берген анықтамасы: «Ауыздан шыққан сөздің бәрі сөйлем бола бермейді, айтушының ойын тыңдаушы ұғарлық даражада түсінікті болып айтылған сөздер ғана сөйлем болады» [10, 257-б.].

Е. Омаров «Синтаксистік жаттықтырғыш» атты еңбегінде сөйлемді мүшелеу қағидаларын ұсынды. Ғалым: «Сөйлемнің бір мүшесіне ерген жетек өзіне тағы бір сөзді бағындырып соған жетекші бола алады; оған бағынған жетектің өзіне ерткен тағы жетегі болуы да мүмкін... сөйлемнің мүшелері осылай: бір сөз өзіне екінші сөзді бағындырып, екінші сөзге үшінші бір сөз, бұған тағы бір сөз бағынып, біріне бірі үйтіп тіркесе (шұбыра) жетектесіп болса, сөйлемнің ондай мүшелері – шұбырма жетек делінеді», – деп сөйлемдегі сөздердің байланыстарына қатысты анықтама береді Сонымен қатар, «жетектесудің түр-түрін айырғу үшін сөйлемді мынадай қылып ысқемалауға болады...» деп сөйлем ішіндегі сөздердің өзара байланысын түсіндіреді [11, 60-б.]. Бұл – қазіргі синтаксистік белгіленім жасауда қолданылатын жүйелердің, теориялардың негізі. Е. Омаров – XX ғасырдың басында-ақ қазақ тіл білімінде сөйлемді модельдеудің үлгісін жасаған ғалым.

Қ. Жұбанов мақалаларында аударма мәселесіне байланысты көркем әдебиет, ғылыми, публицистикалық стильдегі сөйлемдерді схемамен көрсеткен. сөйлем мүшелерінің саны, оның атаулары, «айқындауышты өзінің айқындап тұрған сөзімен байланысын», «қай сөз қай сөзді жетектеп тұрғанын», «сөйлемдегі сөздердің қалайша қиындасып тұрғанын», «тіркестердің логикалық екпінге тәуелділігін», т.с.с. 12 түрлі сөйлем сызбасын көрсеткен [12, 535–533-бб.]. Бұлар қазақ сөйлемдерінің синтаксистік модельдерінің алғашқы үлгілері. Профессор А.Қ. Жұбанов: «XX ғасырдың 60-шы жылдарынан бастап жүйелі, ғылыми тұрғыда қолданыла бастаған статистикалық әдістің алғашқы бастамасы Қ. Жұбанов еңбектерінде кездеседі. Ғалым сөз тіркесінің формалары, сөйлем мүшелерінің орын тәртібі, буындар және ондағы дыбыс саны, буындарды оқыту әдістемесі, т.б. мәселелері сөз болған тұста көптеген сандық (статистикалық) деректер мен мәліметтер келтірген. Қазақ тілінің нормативті жүйелеу жұмысына проф. Қ. Жұбановтан кейінгі белгілі тілші-ғалымдар: Н.Т. Сауранбаев, С.К. Кеңесбаев, М.Б. Балақаев, А.Ы. Ысқақов және т.б. ғалымдар атсалысты, бұл еңбектер қазіргі қолданбалы лингвистикадағы компьютерлік лингвистикамен айналысушы зерттеушілерге қайнар көз болып отыр» - деп атап көрсетеді [13, 119-б.].

С. Аманжолов сөйлем мүшелерін «жаңаша жіктеу» деп, бірнеше сөйлемдердің құрылысын көрсетеді. Мұнда автор үлгі-сызбаны сөйлем мүшелерін және құрмалас сөйлемдердің сыңарларын ажырату үшін, сөйлемнің мағыналық құрылымын көрсету мақсатында қолданған [14, 243-б.; 15, 78–80-бб.].

Х. Арғынов бастауышы бар жақты сөйлем мен бастауышсыз жақты сөйлемдерді ажыратып көрсету мақсатында сөйлемдерді үлгі-сызба түрінде берген [16, 100-б.]. Сөйлемдердің осыған ұқсас сызбаларын, сөйлем мүшелерінің орнын көрсететін кестелерді М. Балақаев, Т. Қордабаев, Т. Сайрамбаевтың еңбектерінен көре аламыз. Қазақ тіліндегі коммуникативтік синтаксистің негізін салушы ғалымдардың бірі профессор Р.С. Әмір «Жай сөйлем синтаксисі» оқулығының логика-грамматикалық типтері деген тарауында сөйлем семантикасының бірнеше деңгейлерден тұратындығын, оның біріншісі субъект (S) пен предикаттың (P) абстрактыланған мағынасы мен грамматикалық тұлғаларына байланысты деп шартты белгілермен атап көрсетеді [17, 23-б.]. Аталған еңбектерде қазақ тіліндегі сөйлемдердің құрылымын модельдеудің алғышарттары жасалған.

Ж. Сәдуақасұлы сөйлемнің құрылымдық модельдерін беруде: Б-бас мүше, Т-толықтауыш, Тж – жанама толықтауыш, А – анықтауыш, 31234567 – септік тұлғалары [18,

41-б.]; А.Б. Қайыров N – кез келген есім сөз, S – зат есім, Adj. – сын есім, Num – сан есім, Pron. – есімдік, Adv. – үстеу, Inf. – тұйық етістік, Advlok., Adv.temp., Advmod., – үстеудің мағыналық түрлері, N1234567 – есімді компоненттің септік тұлғалары, V – етістік, т.с.с. шартты таңбаларды пайдаланған [19, 121-б.]. Аталған авторлар сөйлемнің құрылымдық және мағыналық сипаттарын бірлікте қарастырып, бір құрамды және екі құрамды хабарлы жай сөйлемдерді модельдеу арқылы олардың типтерін анықтаған.

Г. Мұсаеваның «Жай сөйлемдердің меңгеру динамикасы» атты зерттеуінде S – сөйлем, N – сөйлемнің субъектісі (бастауыш), V – предикат, сөйлемнің баяндауышы, O – сөйлемнің объектісі (толықтауыш), O2 – ілік септік, O3 – барыс септік, O4 – табыс септік, O5 – жатыс септік, O6 – шығыс септік, O7 – көмектес септік, A – анықтауыш, A1 – сапалық анықтауыш, A2(O2) – меншікті анықтауыш, Adv – пысықтауыш, sing – жекеше түр pl – көпше түр, poss. – тәуелдік жалғау және қиысу, меңгеру, матасу, қабысу байланысу формаларының сызба белгілерін (стемдерін) бере отырып, бір негізді және екі негізді сөйлемдердің алгоритмін көрсетеді [20, 137-б.]. Бұл еңбектердің қазақ тілін формалды аспектіде зерттеу, сөйлемдерді модельдеу, тіл корпустарында синтаксистік белгіленімін жасау сияқты аспектілер бойынша маңызы зор.

Қазақ тіліндегі мәтіндердің тууы мен жасалуының негізгі формалдық, семантикалық ұстанымдарын тану мақсатында А. Жұбановтың «Основные принципы формализации содержания казахского текста» және тілдің синтаксистік, семантикалық, статистикалық модельдерін құрастыру мен қолданбалы лингвистиканың фонетикалық аспектілері мен терминтанудың негізгі бағыттары жөнінде баяндалған «Қолданбалы лингвистика: формалды модельдер» атты еңбектері жарық көрді [21, 213-б.; 22, 34-б.].

Әлемдік тіл білімінде синтаксистік қатынастардың формалды моделін жасауда И.А. Мельчуктың «Dependency Syntax: Theory and Practice» [7] еңбегі мен Л. Теньердің «Éléments de syntaxe structurale» [6] атты классикалық еңбегі бағыныңқылық грамматикасының теориялық платформасын қамтамасыз етті. Сонымен қатар, синтаксистік құрылымдардың сатылы ұйымдасу ережелерін түсіндіруге мүмкіндік берген Н. Хомскийдің өзгеріс тудырушы, туындаушы (трансформациялық-генеративтік) грамматиканы негіздеген «Syntactic Structures» [5] атты еңбегі. Зерттеулердің қазіргі тілдік модельдермен сабақтастығы «UD» [23] және «Surfa Syntactic Universal Dependencies» («SUD») [24] жүйелерінің теориялық ережелері арқылы қамтамасыз етілді.

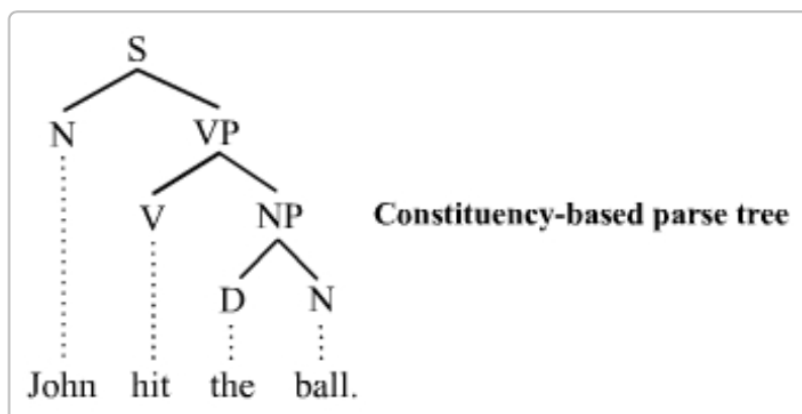
Синтаксистік аннотациялаудың теориялық негізін басты екі грамматикалық бағыт құрайды: құрамдық грамматика (ҚГ) және бағыныңқылар грамматикасы (БГ). ҚГ сөйлемді сатылы құрылымға жіктеп, құрамдас бөліктердің арақатынасын айқындайды. Ал БГ сөйлемдегі сөздердің өзара бағыныңқылық байланыстарын көрсетіп, сөйлемді сөздер арасындағы синтаксистік тәуелділіктер жүйесі ретінде қарастырады.

Құрамдық пен бағыныңқылық тәсілдер арасындағы айырмашылықтар олардың теориялық негізінен туындайды. Алғашқы тәсіл бойынша синтаксистік қатынастар грамматикалық категориялар арқылы көрсетілсе (бастауыш тобы, баяндауыш тобы деп бөлу), басыңқы-бағыныңқылық тәсілде тікелей сөйлемдегі сөздердің арасындағы байланыс көрсетіледі. Құрамдық тәсіл бойынша сөйлем құрылымы көпсатылы деңгейде сипатталып, әр деңгей дербес бірлік ретінде талданады, ал екінші тәсіл бойынша тек нақты сөздер және олардың арасындағы бағыныңқы байланыстар анықталады. Сөйлемдерді модельдеу барысында ҚГ аналитикалық тілдерге (мысалы, ағылшын) икемді болып келсе, ал БГ копармалы немесе жалғамалы, сөз тәртібі біршама еркін, тілдердің (қазақ, орыс, т.б.) тиімді әдісі саналады. Сонымен қатар, БГ құрылымда кейбір синтаксистік құбылыстарды бейнелеу үшін қосымша ережелер қажет болуы мүмкін (сөйлемдегі ауыспалы компоненттер немесе эллипсис), ал құрамдық құрылымның кемшін тұсы – артық фразалық түйіндер көбейіп,

әсіресе ұзақ әрі күрделі сөйлемдерді талдағанда, ағаш құрылымы тым шырмалып кетуі мүмкін. Осындай тиімсіз тұстары болғанымен, қазіргі кезде бұл екеуі бірін-бірі толықтыратын әдістер ретінде танылады. Синтаксистік аннотациялау тәжірибесінде де бастапқыда ағылшын тілді корпустар ҚГ-на негізделсе, кейін әлем тілдерінің көпшілігі үшін басыңқы-бағыныңқылық үлгілер тиімді деп танылды. ҚГ және БГ тіл құрылымын формалды талдаудың екі түрлі қыры, олардың әрқайсысының тарихи қалыптасқан дәстүрі мен тілдік негізі бар.

1950 жылы Ноам Хомский тіл білімінде төңкеріс жасаған жаңа синтаксистік теориясын ұсынды. Ғалым: «Грамматика дегеніміз – белгілі бір тілдің шектеулі ережелерінен шексіз дұрыс құрылған сөйлемдерді тудыратын құрал» – деп түсіндірді [4, 40-б.]. Бұл теорияға сәйкес, сөйлем белгілі бір формалды ережелер негізінде құрылады. Мұндай ережелер сөйлемнің синтаксистік ағаш түріндегі құрылымын анықтаудың негізі, мысалы, ағылшын тіліндегі қарапайым сөйлем құрылымын мынадай ережелер арқылы көрсетуге болады: $S \rightarrow NP + VP$ (сөйлем бастауыш пен баяндауыштан тұрады); $NP \rightarrow Det + N$ (бастауыш – анықтаушы мен зат есімнен құралады); $VP \rightarrow V + NP$ (баяндауыш – етістік пен толықтауыштан тұрады).

Осы ережелердің көмегімен «*John will hit the ball*» сияқты сөйлемдердің грамматикалық құрылымын синтаксистік ағаш немесе фразалар түрінде төмендегідей бейнелеуге болады (1-сурет).



1-сурет – Ағылшын тіліндегі «John will hit the ball» сөйлемінің синтаксистік ағашы
(ҚГ бойынша талдау). «Universal Dependencies» сайтынан алынған.

Бұл ағаштың жоғарғы түйіні – S (сөйлем бастауыш), ол N (бастауыш зат есім) және VP (баяндауыш етістікті тіркесі) деген екі құрылымнан тұрады. Әрі қарай NP құрамы «The ball» (артикль + зат есім), VP өз ішінен «will hit the ball» бөліктеріне мүшеленген. Әрбір түйін ішінде тиісті грамматикалық категориялары таңбаланып, жапырақ түйіндер нақты сөздерді көрсетеді.

Н. Хомский грамматикасының маңызды жаңалығы – өзгеріс тудырушы, «трансформациялар» ұғымын енгізуі болды. Бұл теорияның маңызды категорияларының бірі – түпкі (deep structure) және беткі құрылым (surface structure) ұғымдары. Түпкі құрылым сөйлемнің мағыналық-семантикалық деңгейін, яғни ойдың логикалық қаңқасын көрсетеді. Мұнда бастауыш, баяндауыш, толықтауыш сияқты синтаксистік қызметтердің арасындағы қатынастар нақты әрі абстрактілі түрде беріледі. Түпкі құрылым сөйлем мазмұнының базалық негізін құрайды (пропозиция), ал беткі құрылым сөйлемнің нақты айтылған немесе жазылған формасы болып табылады (иллокуция). Ол түпкі құрылымның түрленуінен

(трансформациялануынан) туындайды. Тілдік жүйеде бір түпкі құрылымның түрлі беткі формалары болуы мүмкін. Негізгі етісті сөйлемді ырықсыз етіс түріне айналдыру, сұраулы сөйлем жасау, сөйлемдерді ықшамдау, актуалдау – тема-реманың ауысуы сияқты түрлілік өзгеріс тудырушы (трансформациялық) грамматиканың ережелері арқылы түсіндіріледі. Бұл өзгерістер нәтижесі – беткі құрылымды да синтаксистік ағаш түрінде бейнелеуге болатындығы. Бұл екі деңгейдің де сөйлем синтаксисін зерттеудегі ролі зор. Түпкі құрылым сөйлем мазмұнын жалпы грамматикалық үлгі-сызба түрінде бейнелесе, беткі құрылым сол мазмұнның тілдік көрінісін айқындайды. Осындай жіктеу сөйлемнің мағынасы мен формасы арасындағы байланысты түсіндіруге, парафраздық құбылысты сипаттауға мүмкіндік береді.

«John will hit the ball» сөйлемінің бастауышы (John), баяндауышы (hit), толықтауышы (the ball) арасындағы абстрактілі қатынас көрсетілген. Бұл түпкі құрылымның тілдік көрінісі. Сөйлем формасы нақты. Трансформациялық ережелер арқылы ол түрлі беткі формаларға айналуы мүмкін:

- *The ball was hit by John* (ырықсыз етіс)
- *Did John hit the ball?* (сұрау мәнді форма)
- *It was the ball that John hit* (актуалданған форма).

Н. Хомскийдің түпкі және беткі құрылым теориясы тілдің ішкі ұйымдасуын, сөйлем құраудың әмбебап заңдылықтарын түсіндіруге бағытталды (1-кесте). Ол синтаксистік модельдеудің жаңа тәсілдерін қалыптастырып, қазіргі компьютерлік лингвистика мен когнитивтік тіл білімінің дамуына мейлінше ықпал етті. Соның нәтижесінде тіл білімінде туындаушы грамматика тілдік білімді формалды түрде сипаттаудың негізгі әдісі ретінде басты орынға шықты. Синтаксистік ағаштар ұғымы да кең таралып, кез келген сөйлемнің құрылымын көрнекі түрде бейнелеудің стандарт әдісіне айналды.

1-кесте – Түпкі құрылым және беткі құрылымдардың сипаты

№	Белгісі	Түпкі құрылым	Беткі құрылым
1	Деңгейі	Абстрактілі, семантикалық-мағыналық деңгей.	Нақты, айтылым/жазылым деңгейі.
2	Мақсаты	Сөйлемнің логикалық-мағыналық қаңқасын көрсету.	Сөйлемнің нақты тілдік формасын беру.
3	Құрамдық байланысы	Синтаксистік қызметтердің (NP, VP, т.б.) қатынасын көрсету.	Сөздердің орын тәртібі, формасы мен трансформация нәтижесін көрсету.
4	Рөлі	Сөйлем мазмұнының базалық (түпкі) негізін жасау.	Сөйлемнің көрінісін, қолданыстағы түрін анықтау.
5	Құрылым түрі	Базалық, түпкі құрылым.	Түпкі құрылымының өзгеруі нәтижесінде пайда болатын құрылым.
6	Қолданылуы	Теориялық, абстрактілі синтаксистік модельдеу.	Қолданыстағы сөйлемді (иллокутивті акт деңгейіндегі сөйлем) модельдеу.
<i>Ескерту:</i> Кесте деректер негізінде құрастырылған [4; 5; 6; 7].			

Н. Хомскийдің өзгертуші-туындаушы теориясының (ӨТТ) тіл біліміне қосқан үлесі: а) синтаксистік құрылымды математикалық дәлдікпен сипаттау мүмкіндігі; ә) әмбебап грамматика идеясының ұсынылуы – құрылымдық қағидалар барлық тілдерге ортақ және сөйлеуші бұрын естіп көрмеген сөйлемді де грамматикалық тұрғыдан дұрыс құрастыра алатынын – осы модель арқылы дәлелдеуге болады деген тұжырымдардың қалыптасуы.

Сонымен қатар, ӨТТ адам миындағы тіл механизмін модельдеудің бір әдісі ретінде қарастырылып, психоллингвистика мен когнитивтік бағыттарға ықпал етті. Н. Хомскийдің теориясы құрылымдық грамматиканың дамуындағы ең маңызды кезең болды және қазіргі синтаксистік зерттеулердің едәуірінің әдіснамалық негізін қалады.

1960-70 жылдары Игорь Мельчуктың басшылығымен құрамдық және БГ-ның өзіндік синтезін ұсынған ауқымды тіл білімінің моделі – «мағына-мәтін» теориясы (Meaning-Text Theory, МТТ) қалыптасты. И. Мельчуктың көздеген мақсаты табиғи тілдің компьютерлік моделін жасау болғандықтан, тілдің бірнеше деңгейлерінің құрылымы – семантика, синтаксис, морфология және лексика-фразеология деңгейлері – «мағынадан мәтінге» және кері бағытта түрленетін көпқабатты модель ретінде ұсынылды. Аталған теория бойынша синтаксистік компонент толықтай БГ-на негізделді. Алайда БГ-дан ерекшелігі – мұнда түпкі және беткі синтаксис деңгейлері басшылыққа алынды [7, 23-б.].

Түпкі модельде мағыналық рөлдер басты орынға шығып, кейбір көмекші сөздердің қызметі ескерілмеді, ал беткі синтаксисте нақты сөздер арасындағы басыңқы-бағыныңқылық байланыстар толық көрініс тапты. И. Мельчук енгізген тағы бір ұғым – лексикалық функциялар және валенттілік: белгілі бір сөздің (етістіктің) семантикалық актанттары оның синтаксистік басыңқы-бағыныңқылық қасиеттерін де анықтайды, яғни қай сөздер басыңқы, қай сөздер бағыныңқы болатыны сол сөздердің мағынасына сәйкес анықталады. «МаТТ» шеңберінде сөйлемнің басыңқы-бағыныңқылық құрылымын белгілейтін арнайы белгілер жасалды (мысалы, 1-актант, 2-актант, атрибуттық байланыс, т.б.). Кейіннен бұл белгілер көптеген корпус-аннотация стандарттарына негіз болды, мысалы, И. Мельчук пен оның мектебі құрастырған орыс тілінің синтаксистік сөздігіндегі басыңқы-бағыныңқылық белгілері кейін ресейдегі ұлттық корпус жобасында да басшылыққа алынды. «МаТТ» теориясы жалпы тіл білімінде және машиналық аударма, корпус лингвистикасы салаларында кеңінен танылды. И. Мельчук еңбектерінің ықпалымен 1980–90 жылдары бірнеше тілдердің үлкен синтаксистік басыңқы-бағыныңқылық ағашбанктері (treebank) құрылды.

XX ғасыр ортасында түрлі мектептер синтаксисті өздерінің әдіснамалық мақсаттарына сай дамытты. Құрылымдық мектеп өкілдері тілдегі нақты категорияларды анықтауға және олардың сатылы байланысына мән берсе, Мәскеу-Монреаль мектебі (И. Мельчук) мағынаны туындату (генерациялау) үдерісіне көңіл аударды, дегенмен қазіргі әмбебап аннотация жүйелерінің қалыптасуына екі дәстүрдің де ықпалы ерекше болды. Бірыңғай стандартпен әлем тілдерінің мәтін корпустарының синтаксистік белгіленімін жасау мақсатымен, 2010 жылдары халықаралық «UD» консорциумы құрылды [20]. «UD» – компьютерлік лингвистика мен корпустық грамматикалар тоғысындағы жоба, оның басты ерекшелігі – барлық тілдер үшін ортақ грамматикалық категориялар мен басыңқы-бағыныңқылық қатынастардың жиынтығын ұсыну [20]. «UD» – түрлі тілдерге ортақ ереже бойынша морфосинтаксистік белгіленім жасауға арналған халықаралық стандарт. 2016 жылы жобаның бірінші нұсқасы ресми түрде жарық көрді. Қазіргі кезде «UD» үлігісі негізінде 100-ден астам тілдік корпустардың белгіленімі жасалып, көптілді ауқымды дерекқорлар қалыптасқан. Мақсаты – әртүрлі тілдердің сөйлем құрылымдарын бір жүйеге келтіріп, кросслингвистикалық (тіларалық) негізде үйлестірілген белгіленімін жасау. Мұндай үйлесімді белгіленім көптілді синтаксистік талдау жүргізуді, типологиялық салыстыруларды жүзеге асыруды және көптілді NLP қолданбаларын құруды айтарлықтай жеңілдетеді. «UD» негізінен «Станфорд бағыныңқылғы» деп аталатын ағылшын тілі грамматикасының бұрынғы үлігісіне сүйеніп жасалды. «UD» аннотациясының құрылымы бірнеше деңгейлі ақпаратты қамтиды: 1) сөз таптары – әр сөздің тиісті әмбебап сөз табы категориясы белгілері (барлығы 17 POS-тег, мысалы NOUN, VERB, ADJ, PRON, ADV, ADP, CCONJ, DET т.б.); 2) морфологиялық белгілері (сөздің грамматикалық формалар белгілері мысалы,

жекеше/көпше, жақ, шақ, рай, септік сияқты категориялар) стандарт қысқартылған мәндермен көрсетіледі; 3) синтаксистік басыңқы-бағыныңқылық қатынастар – сөйлемдегі әрбір сөздің қай басыңқы сөзге тәуелді екендігін және оның байланысу түрін көрсетеді. «UD» әр сөзді жеке бір жол ретінде көрсететін CoNLL-U форматындағы кестеде сақталады: онда сөздің реттік нөмірі, өзі, леммасы, сөз табы (UPOS), морфологиялық тегтері, басыңқы сөздің нөмірі және синтаксистік қатынасы көрсетіледі.

Бұл платформада грамматикалық категориялар (сөз таптары) мүмкіндігінше кең қамтылып, дәстүрлі топтарға сәйкес беріледі. Мысалы: зат есім (NOUN), етістік (VERB), сын есім (ADJ), үстеу (ADV), есімдік (PRON), жалғаулықтар (CCONJ, SCONJ), артикль (DET), сан есім (NUM), септік жалғаулары мен шылаулар (PART), т.б. (барлығы 17 топ). Жекелеген тілдердегі айрықша категориялар үшін (мысалы, корей немесе жапон тілдеріндегі белгілік сөздер) қосымша таңбалар енгізілуі мүмкін.

«UD» жүйесінде синтаксистік қатынастың шамамен 40-қа жуық арнайы атауы бар. Ең жиі қолданылатындары: *nsubj* (nominal subject) – бастауыш; *cop* / *xcomp* – баяндауыш құрамындағы предикативті зат есім немесе сын есім; *obj* – тура толықтауыш; *obl* (oblique nominal modifier) – жанама толықтауыш немесе пысықтауыш мәніндегі мүше; *amod* – зат есімнің сындық анықтауышы; *advmod* – етістіктің үстеу пысықтауышы; *advcl* – үстеу бағыныңқы сөйлем; *acl* – анықтауыштық бағыныңқы сөйлем; *cc* – жалғаулық шылау (координатор).

Осы қатынастардың 37-сі нақты басыңқы–бағыныңқы байланыстарды білдіреді, мысалы, бастауыш (*nsubj*) пен тура толықтауыш (*obj*) – дәстүрлі грамматикадағы бастауыш пен толықтауышқа сәйкес келеді. Ал *obl* (жанама толықтауыш/пысықтауыш) – етістікке бағыныңқы әртүрлі сөйлем мүшелерінің (мезгіл, мекен, амал, т.б.) ортақ қамтылған атауы.

Бұл жүйедегі барлық қатынастар тек қана дәстүрлі басыңқы–бағыныңқы байланыстарға сәйкес келе бермейді, кейбір қатынастар сөздердің салаласа байланысын да сипаттайды, Мысалы: *flat* және *fixed* – күрделі есімдердің (аты-жөні, географиялық атаулар) сыңарларын байланыстырады. Мұндай тіркестер тұтас бір сөз тіркесі ретінде қарастырылады. *-cc* – жалғаулық шылауларды (мысалы, *and* / *және*) белгілеу үшін қолданылады. Сондықтан «UD»-де қатынастардың басым бөлігі дәстүрлі грамматикадағы сабақтас байланысқа сәйкес келсе де, кейбірі салалас, бір-біріне тәуелсіз байланыстарды да сипаттайды.

«UD» жобасы бірнеше қағидаларға сүйенеді. Солардың бірі – басыңқы ретінде атауыш сөздерді белгілеу, яғни, «UD» ережелеріне сәйкес сөйлемде белгілі бір грамматикалық қызмет атқаратын көмекші сөздер (шылаулар, жалғаулықтар, көмекші етістіктер, артикльдер, т.б.) толық мағыналы атауыш сөздердің бағыныңқысы болып белгіленеді. Бұл қағиданың уәжі мынада: әлем тілдерінде атауыш сөздердің табиғаты біршама әмбебап әрі тұрақты, сондықтан олардың қызметі де ұқсас келеді. Ал көмекші сөздердің көрінісі әртүрлі: кей тілде олар дербес сөз ретінде жұмсалады, кей тілде морфемаға айналады, ал басқа бір тілдерде мүлде кездеспеуі мүмкін.

Сондықтан атауыш сөзді басыңқы ретінде белгілеу – түрлі тілдердегі баламалы құрылымдарды параллель түрде танудың бір тәсілі. Мысалы, ағылшын тілінде болымсыздық *not* шылауы арқылы берілсе, түркі тілдерінде оның морфологиялық тәсілі, етістіктің *-ма/ме* болымсыздық жұрнағы арқылы жасалатындығы белгілі. «UD» жүйесінде ағылшынның *not* сөзі негізгі етістіктің бағыныңқы пысықтауышы (*advmod*) ретінде белгіленеді, ал қазақ тіліндегі болымсыздық жұрнағы етістіктің морфологиялық белгісі ретінде (Feature: Polarity=Neg) таңбаланады. Екі жағдайда да басыңқы ретінде негізгі етістік алынады.

Сол сияқты, предлогтар мен септеуліктер де өздері анықтап, толықтырып, пысықтап тұрған сөздің бағыныңқы бірлігі ретінде белгіленеді. Мысалы, ағылшын тіліндегі “*go to school*” тіркесінде *to* предлогы етістікке емес, *school* сөзіне бағынып, *case* қатынасы ретінде

таңбаланады. Ал *school* сөзі өз кезегінде *do* етістігіне бағынып, *obl* қатынасы арқылы белгіленеді, дегенмен кейбір басқа жобаларда («*SUD*») мұндай қатынастар басқа қағидалар бойынша түрліше рәсімделуі мүмкін.

Бүгінде «UD» аясында 100-ден астам тілдің синтаксистік ағашбанктері құрылып, кең қолданыс табуда [8]. Олардың құрамында қазақ тілі де бар (Kazakh UD Treebank). Қазіргі дереккөздер мен зерттеулер қазақ тілінің синтаксистік аннотациясын болашақта жетілдірудің мынадай негізгі бағыттарын айқындайды:

Біріншіден, «UD» аясында аннотация стандарттарын жетілдіру (жаңарту) қажеттігі (кейбір күрделі құрылымдар үшін жаңа субтегтер қосу).

Екіншіден, ресурстарды кеңейту – деректер көлемін ұлғайту (зиянды блоктарды, ауызша сөйлеу үлгілерін енгізу), жанрлық әртараптандыру; коруптар сапасын арттыру (аннотация дәлдігін қайта тексеру).

Үшіншіден, халықаралық стандарттарға үйлестіру – UD-ға қосымша «Penn Treebank» сияқты басқа стандарттармен синхрондау, көптілді модельдермен ықпалдастыру.

Төртіншіден, NLP құралдарымен ықпалдастыру: қазіргі синтаксистік талдағыштар, машиналық аударма мен сұраныс жүйелері қазақ тіліне бейімделіп, өнімділігін арттыру үшін осы аннотациялық корпустарды қолдану (тренинг және тест). Мысалы, UDPipe, spaCy сияқты тәуелділік талдағыштарға арналған қазақша модельдер даярлануы, олардың сапасын қазақ тілінің өз ерекшеліктерімен қайта баптау қажеттігі. Осы бағыттағы жұмыстар қазақ тілінің ақпараттық ресурстарын дамытудың, тілдің цифрлық экожүйесін байытудың бағдары бола алады.

Сонымен, «MaTT» теориясы БГ-ны семантикамен қабыстыра отырып, көпдеңгейлі лингвистикалық модель ұсынса, UD бастамасы бүкіл тілдерге ортақ тәуелді аннотация үлгісін жасап шығарды. Екеуіне де ортақ идея – синтаксистік құрылымды «сөз – бағыныңқы сөз» формасында көрсету және тіл аралық ортақтықты осы құрылым арқылы таныту.

Әр бағыттың өз артықшылықтары мен әлсіз тұстары бар. ӨТГ-де семантика еленбей, тек құрылымға назар аударылғаны үшін сынға алынған, себебі мағыналық деңгейдің ескерілмеуі талдаудың пәрменділігін төмендетуі мүмкін. Ал БГ-да аралық абстрактілі түйіндер қарастырылмайтындықтан, күрделі құрылымдарды сипаттау қиынға соғады. Соған қарамастан, тіл біліміндегі көптеген мәселелердің шешімін табу жолында бұл екі тәсіл өзара үйлесіммен қолданылады: бірінің әлсіз тұстарын екіншісі толықтырады. Қазіргі грамматикалық теориялар (мысалы, Head-Driven Phrase Structure Grammar, Lexical Functional Grammar) құрамдық және басыңқы-бағыныңқылық идеяларды ықпалдастыра отырып, тілдің құрылымдық әрі функционалдық қырларын бірдей ескеруге ұмтылады (2-кесте).

2-кесте – Синтаксистік белгіленімнің теориялық негіздерінің салыстырмалы кестесі

№	Атауы	ҚГ (Н. Хомский және ізбасарлары)	БГ (Л. Теньер, т.б.)	«MaTT» теориясы (И. Мельчук)	«UD» (эмбебап басыңқы- бағыныңқылық аннотация)
1	2	3	4	5	6
1	Талдау бірлігі	Құрамдық (конституент) – сөз тіркестері (фразалар)	Сөздер (лексемалар), басыңқы және бағыныңқы бірліктер	Сөздер (терең және беткі деңгей актанттары)	Сөздер (толық мағыналы сөздер – басыңқы, көмекші сөздер – бағыныңқы)

2-кестенің жалғасы

1	2	3	4	5	6
2	Құрылым репрезентациясы	Сатылы ағаш, ішкі түйіндері грамматикалық категориялар (NP, VP, т.б.)	Көлбеу ағаш, сөз түйіндерінің бір бағытты байланыспен берілуі	Екі деңгейлі басыңқы-бағыныңқылық ағаштар, семантикалық және синтаксистік белгіленім жасау мүмкіндігі	Басыңқы-бағыныңқылық ағаш, стандарт белгілермен көрсету мүмкіндігі (беткі синтаксиске негізделуі)
3	Тарихи негізі	Америкалық структурализм (Блумфилд) → Хомский (1950–60 жж.) туындаушы теориясы	Еуропалық құрылымдық грамматика (Теньер, 1959);	Кеңес Одағы (1960 жж.) → Канада (1970–80 жж.), Мельчук еңбектері	2014 ж. бастап халықаралық жоба (Нивр және шәкірттері)
4	Артықшылығы	Жүйелі құрылым; формалды ережелермен дәл сипатталуы; сөйлем құрамын кезеңдеп талдау мүмкіндігі	Сөйлемнің нақты байланыстарын көрнекі етіп көрсету мүмкіндігі; сөз тәртібі өзгергенімен мағынасының сақталуы	Мағыналық және құрылымдық деңгейлерді байланыстыру; модельдердің машинамен өңдеуге қолайлы болуы	Түрлі тілдік құрылымдарды біріздендірілген форматпен белгілеу мүмкіндігі, көптілді корпустарға икемді болуы
5	Кемшін тұстары	Сөз тәртібі еркін құрылымды тілдерге икемсіз болуы; сөйлем ағашының тым көлемді болуы	Аралық құрамдық түйіндер жоқ болғандықтан, кейбір күрделі құрылымдардың толымсыз сипатталуы	Теориялық мәнінің күрделі болуы	Синтаксистік талдаудың тым қарапайымдануы; кейбір тілдік ерекшеліктердің еленбей қалуы; даулы мәселелердің болуы
<i>Ескерту:</i> Кесте деректер негізінде құрастырылған [4; 5; 6; 7; 8].					

Өртүрлі грамматикалық мектептердің өкілдері синтаксистік құрылымды әр қырынан бағалайды. Н. Хомский тілді адам миындағы туа біткен құрылым деп қарастырып, грамматиканың міндеті – сол ішкі тілдік қабілеттің ережелерін модельдеу деп санайды. Кез келген тілдің шегінде шектеулі сандағы ережелер арқылы шексіз сөйлемдер құрауға болады, демек грамматика – сөйлеушінің ішкі интуициясын формалды түрде сипаттайтын тетік [18, 24-б.]. Ғалым бастауыш-баяндауыш мүшелеріне бөлуді синтаксистік құрылымның өзегі деп білді, оның пайымдауынша әр сөйлем міндетті түрде NP және VP құрамдарына ажыратылады. Ол енгізген құрылымдық деңгейлер (түпкі және беткі құрылымдар) және оларды түрлендіретін ережелер бастапқыда корпустарда тікелей қолданыс таппағанымен, 1960-70 жылдары бұл идеялар компьютерлік тіл өңдеуде ескеріле бастады. Мысалы 1960 жылдардың соңында алғашқы синтаксистік талдау бағдарламалары ғалымның мәнмәтіндік-еркін грамматикалары теориясына сүйенді. Трансформациялық қағидалар ағылшын тіліндегі синтаксистік ағаштарда құрамдас бөліктерді (constituents) деңгей-деңгейімен белгілеу тәжірибесін нығайтты. Бұл идеялар «Penn Treebank» сияқты жобаларда қолданылған конституенттік аннотацияның теориялық негізі болды.

Л. Теньер керісінше, дәстүрлі бастауыш-баяндауыш дихотомиясын сынға алып, сөйлемді «бағыныңқы тақтайша» (граф) түрінде бейнелеуді ұсынды. Л. Теньер үшін сөйлемнің басты мүшесі – предикаттық етістік, ол тікелей бірнеше «актанттарды» (мысалы, қазақ тілінде бастауыш, толықтауыш) өзіне бағындырады [6, 89-б.]. Л. Теньер субъект пен нысанды дәрежелері тең синтаксистік рөлдер ретінде қарап, оларды тек етістікке қатысы арқылы анықтау керек деген көзқарасты ұстанды. Л. Теньердің бұл идеясына 1970–80 жылдары компьютерлік аударма мен ақпараттық технологиялардың дамуына орай, ғалымдар тарапынан қызығушылық арта бастады.

И. Мельчук – өз еңбектерінде тілдік бірліктер арасындағы мағыналық және синтаксистік қатынастарды формальды түрде көрсетуге көп көңіл бөлген ғалым. И. Мельчуктың пайымдауынш, грамматика тек құрылым емес, сөздердің мағыналық өзара әрекеті моделіне айналуы қажет [7, 72-б.]. Оның «MaTT» теориясында әр сөздің лексикалық мағынасы синтаксистік валенттілік арқылы белгілі болады: етістіктің «семантикалық актанттары» оның сөйлемдегі тәуелді сөздерін анықтайды [17, 93-б.]. И. Мельчуктың жұмыстары БГ-ны қолданбалы бағытта дамытуға, әсіресе машиналық аударма мен компьютерлік лексикография саласында синтаксистік басыңқы-бағыныңқылық байланыстарды белгілеу әдістерін жасауға зор үлес қосты. Ғалымның енгізген синтаксистік байланыстың әмбебап типтері қазіргі көптеген корпус-аннотация стандарттарына негіз болды.

Жоғарыда аталғандардан бөлек БГ-ның дамуына чех ғалымдары (В. Матезиус бастаған Прага мектебі), неміс зерттеушілері (У. Энгель, т.б.), қазіргі заман тіл технологиясы мамандары (Ж. Нивр және басқалар) өз үлестерін қосты. Әрқайсысы өз ұстанымдары арқылы синтаксистік құрылымды тереңірек түсінуге ықпал етті.

Синтаксистік аннотация тәжірибесі үшін құрамдық/бағыныңқылық модельдерді таңдау көбіне нақты тіл мен нақты мақсатқа байланысты. Егер мақсатымыз белгілі бір тілдің құрылымын терең түсіндіру болса, ҚГ-ның категорияландыру күші тиімді; ал егер көптілді мәтіндердің жылдам парсингін жасау не машиналық аударма болса, БГ-ның қарапайымдылығы тиімді.

Қорытынды

Қорыта айтқанда, синтаксистік аннотация мен формалды грамматика – тіл білімінің іргелі құралдары. Құрылымдық тіл білімі синтаксистік аннотацияның категориялық аппаратын (сөз таптары мен синтаксистік деңгейлер) қалыптастырса, формалдау мен генеративтілік оның ережелік-құрылымдық негізін қалады. Осы теориялық алғышарттар синтаксистік белгілерді жүйелеуге философиялық негіз болды.

Қазақ тілін синтаксистік тұрғыдан аннотациялаудың ғылыми іргетасы да – формалды грамматика теориясы. Олардың қағидаттарын ұлттық тіл ерекшеліктерімен ұштастыра қолдану қазақ тілінің синтаксистік табиғатын толық әрі дәл сипаттауға жол ашады. Қазіргі таңда «UD» платформасында қазақ тілінің синтаксистік белгіленімі жасалған. Бұл – әлем тілдерін біріздендіруге, әмбебап үлгі құруға бағытталған тиімді жоба екендігін айттық. Дегенмен аталған жүйе қазақ тілінің ерекшелігін түгел қамти алмайды. Біздің тіліміздің жалғамалық сипаты, матасу арқылы берілетін меншіктілік, сапалық анықтауыштардың қабысуы, септіктердің пысықтауыштық қызметі, сөз таптарының конверсиялануы, номинативтілік сипаты, субстантивтенудің феноменділігі және құрмалас сөйлемдердің өзіндік табиғаты сияқты ерекшеліктер арнайы ұлттық синтаксистік белгіленім арқылы ғана айқын көрініс таба алады.

Сондықтан қазақ тілінің синтаксистік аннотациясы әлемдік тәжірибеге сүйеніп қана қоймай, формалды грамматика қағидаларына негізделген ұлттық үлгі ретінде қалыптасуы

тиіс. Сонда ғана ол тілдің бай мүмкіндіктерін, құрылымдық икемділігін және табиғи реңктерін толық танытып, ұлттық корпустар үшін сенімді ғылыми-әдіснамалық база бола алады.

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.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика. – Алматы: Қазақ тілі, 2017. – 336 б.
2. Словарь по языкознанию. Тіл білімі сөздігі / под общ. ред. Э.Д. Сулейменовой. – Алматы: Ғылым, 1998. – 540 с.
3. Салқынбай А., Абақан Е. Лингвистикалық түсіндірме сөздік. – Алматы: Сөздік-словарь, 1998. – 330 б.
4. Chomsky N. Syntactic Structures. – The Hague: Mouton, 2002. – 117 б.
5. Хомский Н. Аспекты теории синтаксиса. – М.: Изд. МГУ, 1972. – 117 с.
6. Tesnière L. Éléments de syntaxe structurale. – Paris: Librairie C. Klincksieck, 1959. – 690 p.
7. Mel'čuk I.A. Dependency Syntax: Theory and Practice. – Albany: State University of New York Press, 1988. – 428 p.
8. Universal Dependencies [Electronic resource]. URL: https://universaldependencies.org/treebanks/tr_imst/index.html (date of access 01.05.2025)
9. KazDT. kazdet: NLA-NU Kazakh Dependency Treebank [Electronic resource]. URL: <https://github.com/nlacsilab/kazdet> (қаралған күні 10.05.2025)
10. Байтұрсынов А. Тіл тағлымы. – Алматы: Ғылым, 1992. – 446 б.
11. Омарұлы Е. Үш томдық шығармалар жинағы: зерттеулер, мақалалар. – Алматы: Алашорда, 2018. – 400 б.
12. Жұбанов Қ. Қазақ тілі жөніндегі зерттеулер. – Алматы: Атамұра, 2010. – 608 б.
13. Жұбанов А.Қ. Қазақ тілтанымындағы статистикалық әдістің орны // Кіт.: Жизнь языка и язык жизни. – Алматы: Қазақ университеті, 2004. – Б. 115–129.
14. Аманжолов С. Қазақ тілі теориясының негіздері. – Алматы: Ғылым, 2002. – 366 б.
15. Аманжолов С. Қазақ әдеби тілінің қысқаша курсы. – Павлодар: Павлодар мемлекеттік университеті, 2008. – 110 б.
16. Арғынов Х. Жай сөйлем синтаксисінің методикасы. – Алматы: Мектеп, 1967. – 112 б.
17. Әмір Р., Әмірова Ж. Жай сөйлем синтаксисі. – Алматы: Қазақ университеті, 2003. – 199 б.
18. Садуақасұлы Ж. Қазақ тіліндегі бір құрамды сөйлемдердің құрылымдық типтері: филол. ғыл. док. ... дис. – Алматы, 1997. – 305 б.
19. Қайыров А.Б. Қазіргі қазақ тіліндегі екі құрамды хабарлы жай сөйлемдердің құрылымдық-мағыналық модельдері: филол. ғыл. канд. ... дис. – Алматы, 1999. – 121 б.
20. Мұсаева Г.Ә. Жай сөйлемдерді меңгеру динамикасы. – Алматы: Әл-Фараби атындағы Қазақ ұлттық университеті, 2013. – 137 б.
21. Жұбанов А.Қ. Қолданбалы лингвистика: формалды модельдер. – Алматы: Қазақ университеті, 2006. – 280 б.
22. Жубанов А.К. Основные принципы формализации содержания казахского текста. – Алматы: Каз НИИ НТИ, 2006. – 617 с.
23. Nivre J. et al. Universal Dependencies v1: A Multilingual Treebank Collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). – Portorož, Slovenia, 2016. – P. 1659–1666.
24. Gerdes K., Guillaume B., Kahane S., Perrier G. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD // Proceedings of the Second Workshop on Universal Dependencies, 2018. – P. 66–74.

REFERENCES

1. Zhubanov A.Q., Zhanabekova A.A. Korpustыq lingvistika [Corpus linguistics]. – Almaty: Qazaq tili, 2017. – 336 b. [In Kazakh]
2. Slovar po iazykoznaniu. Til bilimi sozdigi [Dictionary of Linguistics. Dictionary of Linguistics] / pod obsh. red. E.D. Suleimenovoi. – Almaty: Gylym, 1998. – 540 s. [In Russian and Kazakh]
3. Salqynbai A., Abaqan E. Lingvistikalyq tusindirme sozdik [Explanatory dictionary of linguistics]. – Almaty: Sozdik-slovar, 1998. – 330 b. [In Kazakh]
4. Chomsky N. Syntactic Structures. – The Hague: Mouton, 2002. – 117 b.
5. Homski N. Aspekty teorii sintaksisa [Aspects of the Theory of Syntax]. – M.: Izd. MGU, 1972. – 117 s. [In Russian]
6. Tesnière L. Éléments de syntaxe structurale. – Paris: Librairie C. Klincksieck, 1959. – 690 p.
7. Mel'čuk I.A. Dependency Syntax: Theory and Practice. – Albany: State University of New York Press, 1988. – 428 p.
8. Universal Dependencies [Electronic resource]. URL: https://universaldependencies.org/treebanks/tr_imst/index.html (date of access 01.05.2025)
9. KazDT. kazdet: NLA-NU Kazakh Dependency Treebank [Electronic resource]. URL: <https://github.com/nlacslab/kazdet> (karalan kыni 10.05.2025)
10. Baitursynov A. Til taglymy [The study of language]. – Almaty: Gylym, 1992. – 446 b. [In Kazakh]
11. Omaruly E. Ush tomdyq shygarmalar zhinagy: zertteuler, maqalalar [Collected works in three volumes: studies, articles]. – Almaty: Alashorda, 2018. – 400 b. [In Kazakh]
12. Zhubanov Q. Qazaq tili zhonindegi zertteuler [Studies on the Kazakh language]. – Almaty: Atamura, 2010. – 608 b. [In Kazakh]
13. Zhubanov A.Q. Qazaq tiltanymyndagy statistikalyq adistin orny [The role of the statistical method in Kazakh linguistics] // Kit.: Zhizn iazyka i iazyk zhizni. – Almaty: Qazaq universiteti, 2004. – B. 115–129. [In Kazakh]
14. Amanzholov S. Qazaq tili teoriasynyn negizderi [Fundamentals of the theory of the Kazakh language]. – Almaty: Gylym, 2002. – 366 b. [In Kazakh]
15. Amanzholov S. Qazaq adebi tilinin qysqasha kursy [Short course of the Kazakh literary language]. – Pavlodar: Pavlodar memlekettik universiteti, 2008. – 110 b. [In Kazakh]
16. Argynov H. Zhai soilem sintaksisinin metodikasy [Methods of simple sentence syntax]. – Almaty: Mektep, 1967. – 112 b. [In Kazakh]
17. Amir R., Amirova Zh. Zhai soilem sintaksisi [Simple sentence syntax]. – Almaty: Qazaq universiteti, 2003. – 199 b. [In Kazakh]
18. Saduqasuly Zh. Qazaq tilindegi bir quramdy soilemderdi qurylymdyq tipteri: filol. gyl. dok. ... dis. [Structural types of one-member sentences in Kazakh: dis.]. – Almaty, 1997. – 305 b. [In Kazakh]
19. Qaiyrov A.B. Qazirgi qazaq tilindegi eki quramdy habarly zhai soilemderdin qurylymdyq-magynalyq modelderi: filol. gyl. kand. ... dis. [Structural-semantic models of two-member declarative simple sentences in modern Kazakh: dis.]. – Almaty, 1999. – 121 b. [In Kazakh]
20. Musaeva G.A. Zhai soilemderdi mengeru dinamikasy [Dynamics of mastering simple sentences]. – Almaty: Al-Farabi atyndagy Qazaq ulttyq universiteti, 2013. – 137 b. [In Kazakh]
21. Zhubanov A.Q. Qoldanbaly lingvistika: formaldy modelder [Applied linguistics: formal models]. – Almaty: Qazaq universiteti, 2006. – 280 b. [In Kazakh]
22. Zhubanov A.K. Osnovnye principy formalizatsii soderzhania kazahskogo teksta [Basic principles of formalization of Kazakh text content]. – Almaty: Kaz NIINTI, 2006. – 617 s. [In Russian]
23. Nivre J. et al. Universal Dependencies v1: A Multilingual Treebank Collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). – Portorož, Slovenia, 2016. – P. 1659–1666.
24. Gerdes K., Guillaume B., Kahane S., Perrier G. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD // Proceedings of the Second Workshop on Universal Dependencies, 2018. – P. 66–74.