КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

ӘОЖ 004.912
ҒТАХА 28.23.37
https://doi.org/10.53364/24138614_2026_41_2_20

Л.М. Байтенова[1], **С.А. Тусупова**[1], **Г.С. Мухамеджанова**[2*], **Г.Н. Мунайтбас**[3], **Д.Н. Нұртаза**[1]
[1]Тұран университеті, Алматы, Қазақстан
[2*]Нархоз университеті, Алматы, Қазақстан
[3] «Хоум Кредит Банк» АҚ, Алматы, Қазақстан

[*]*E-mail: gulnar.mukhamedzhanova@narxoz.kz*

МОРФОЛОГИЯЛЫҚ ҚАТЕЛЕР ТАКСОНОМИЯСЫ ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН ГИБРИДТІ NLP-ЖҮЙЕЛЕРДЕГІ МАРШРУТТАУ МЕХАНИЗМІ РЕТІНДЕ

Аңдатпа. *Қазақ тілінің агглютинативтік сипаты және кірме сөздердің қарқынды ағымы бірнеше ауытқу кластарын құрайды, олардың NLP-модельдеріне әсер ету механизмі түбегейлі ерекшеленеді. Бұл өз кезегінде автоматтандырылған жүйелер үшін қиындықтар туғызады. Жағдайды одан әрі орыс және ағылшын тілдерінен қосылатын үздіксіз сөздер ағымы қиындатады. Олар тілдің фонологиялық логикасынан ішінара ауытқып кетеді. Нәтижесінде морфологиялық анализатор бір ғана емес, бірнеше түбегейлі қиын формалар түрлеріне тап болады. Қолданыстағы жүйелер NLP модельдерінің өнімділігін талдауға әсер ету механизміне негізделген аномалиялардың түрлерін ажыратпай, біріктірілген түрде бағалайды. Бұл өңдеуді мақсатты басқаруға мүмкіндік бермейді. Осы зерттеудің нысаны болып морфологиялық қателердің түрлері мен қазақ тіліне арналған трансформерлік NLP-модельдердің өнімділік көрсеткіштері арасындағы байланыс табылады.*

Жұмыстың мақсаты – морфологиялық аномалиялардың формалды таксономиясын әзірлеу және оны аналитикалық компоненттерді бағыттау механизмі ретінде гибритті NLP архитектурасына біріктіру. Зерттеуде формалды-лингвистикалық талдау, Universal Dependencies Kazakh-KTB (1 047 сөйлемдер) материалындағы корпустық әдіс, есептік модельдеу және абляциялық талдау қолданылады. KazMorphCorpus-2026 гибридті архитектурасы rule-based FST-талдауды, CRF-дизамбигуацияны, KazRoBERTa трансформерлік модулін және MFRN белгілердің морфологиялық үйлесімділігін тексеру модулін біріктіреді. Зерттеу нәтижелері бойынша жүйеге маршруттаудың басқару механизмі ретінде біріктірілген кірме сөздер (BOR), аффиксалды күрделілік (AFC), сегменттеу бұзылыстары (SEG), грамматикалық белгілердің қақтығыстары (AGR) және бейтарап класс (NONE) атты морфологиялық аномалиялардың бес класты таксономиясы ұсынылады. Сынақ үлгісінде жүйе Accuracy = 87,4% және Macro-F1 = 0,86 деңгейіне жетті, сапаның ең үлкен өсімі AGR (ΔF1 = +0,14) и AFC (ΔF1 = +0,12) кластарында тіркелді. Жүргізілген эксперимент морфологиялық аномалиялардың әртүрлі типтері трансформерлік модельдің жұмысына әртүрлі әсер ететінін растады және бұл айырмашылықтың практикалық маңызы бар. Талдау алдында аномалия түрін диагностикалайтын және токенді тиісті құрамдас бөлікке бағыттайтын жүйелер мақсатты түрде оңай түсіндіруге және жақсартуға болатын нәтиже береді.

Түйін сөздер: *қателердің таксономиясы, морфологиялық маршруттау, гибридті NLP архитектурасы, KazRoBERTa, FST, CRF, MFRN, агглютинативті тіл.*

Кіріспе. Қазақ мәтінін өңдеуде трансформерлік NLP-модельдерінің сапасының төмендеуі морфологиялық стандартты емес сөз формаларында көбірек байқалады. Дегенмен бұл деградацияның ішкі құрылымы тек жартылай сипатталған күйінде қалады [1, 3]. Қолданыстағы зерттеулер әдетте күрделі формалардың, соның ішінде кірме бірліктердің, ұзын аффикстік тізбектердің және морфемалық шекаралары анық емес жағдайларда нәтиженің нашарлау фактісін тіркейді. Сонымен қатар, бақылаулардың өзі, әдетте, мәлімдеме деңгейінде қалады. Модель күрделі формалармен нашар жұмыс істейді. Мұның астарында қандай қиындық бар деген сұрақ әдетте қойылмайды. Қазақ тілі үшін мұнда маңызды мәселе бар. Агглютинативті морфология тек ұзын сөздермен ғана шектелмейді. Бұл сөз формасының заңдылығы морфемалардың түйіскен жеріндегі морфотактикамен де, фонологиялық шектеулермен бір мезгілде анықталатын жүйе [4].

Бұл жүйеге кірме негіздері енген кезде және олардың саны көбеген сайын, ішкі заңдылықтардың бір бөліктері бұзылады. Формалды түрде қазақша, бірақ құрылымы өзгеше орналастырылған формалар пайда болады. Нәтижесінде, әдетте «морфологиялық тұрғыдан қиын сөздер» деген атаулар іс жүзінде бірнеше мүлдем басқа жағдайларға бөлінеді: кірме сөздерді пайдалану, шамадан тыс жүктелген аффиксация, морфемалық шекаралардың бұзылуы, грамматикалық белгілердің сәйкес келмеуі [5]. Олардың әрқайсысы модельге әр түрлі әсер етеді. Стандартты көрсеткіштер бұл айырмашылықты көрсетпейді. Accuracy және Macro-F1 бір нәрсенің дұрыс емес екенін көрсетеді, бірақ нақты не және қайда екенін түсіндірмейді.

Модельдерді сәйкестендіру үшін ең көп таралған көрсеткіштер, ең алдымен Accuracy және Macro-F1 қажет, бірақ олар сапалық жағынан әртүрлі қателіктерді бірыңғай сандық көрсеткішке дейін азайтады [6]. Сонымен қатар, жиынтық көрсеткіштің бірдей төмендеуі әр түрлі шығу көздеріне байланысты болуы мүмкін: сөздің морфемалық құрылымына сөзбе-сөз токенизацияның сәйкес келмеуі, алынған кірме негізді өңдеудегі тұрақсыздық, морфемалық шекаралардың қате қалпына келуі немесе белгілердің морфосинтактикалық үйлесімділігінің бұзылуы. Нәтижелерді ұсынудың осы әдісімен модельдің түпкілікті өнімділігі бақыланады, бірақ оның детерминанттарының айқындығы жеткіліксіз.

Жағдайды 1047 сөйлемнен тұратын Universal Dependencies Kazakh-KTB стандартты корпусының шектеулі көлемі одан әрі күрделендіреді [1], [2], [3], [7]. Көлемді белгіленген ресурстары бар тілдер үшін жеке құрылымдық аномалиялар статистикалық тұрғыдан перифериялық болып қалуы мүмкін. Қазақ корпусында тіпті салыстырмалы түрде сирек кездесетін ауытқулар да талдау сапасына айтарлықтай әсер ете бастайды, себебі олар тұрақты мысалдардың көп мөлшерінде жоғалып кетпейді. Бұл құрылымдық қателерді диагностикалау міндетін көмекші ғана емес, сонымен қатар негізгі міндетке айналдырады. Қазақ тілін морфологиялық талдаудың гибридті тәсілдері моноархитектуралық шешімдерге қарағанда тұрақты артықшылықты көрсетеді: FST, CRF және трансформерлік модульді біріктіру әрбір компонентке қатысты жинақталған қатені жеке-жеке азайтады [5]. Бұл нәтижені UD Kazakh-KTB материалында жаңғыртуға болады. Алайда, қолданыстағы зерттеулерде мәселе басқаша сұрақ туғызады: қандай нақты ауытқулардың түрлері оған ықпал ететіні емес, архитектураның жалпы пайдасын қалай бағалауға болатыны. Бұл айтарлықтай шектеу. Грамматикалық белгілердің кірмелілігі мен қайшылықтары түбегейлі әртүрлі құбылыстар болып табылады, егер де олар модельдің мінез-құлқына әртүрлі әсер етсе, онда олардың үлесін талдауды бірыңғай қорытынды көрсеткішке дейін қысқартуға болмайды. Бұл зерттеу дәл осы олқылықты толтыруға тырысады.

Бұл мақалада морфологиялық қателер қазақ тілі үшін трансформерлік NLP-модельдердің өнімділігінің анықтаушысы ретінде қарастырылады. Зерттеудің мақсаты морфологиялық ауытқулардың типологиясын ресімдеуге, олардың Universal Dependencies

Kazakh-KTB корпусындағы сапаны төмендетуге қосқан үлесін бағалауға және қате түрінің айқын диагностикасы басқарудың таксономиялық деңгейін пайдаланбайтын жүйемен салыстырғанда гибридті архитектураның дәлдігі мен интерпретациясын қамтамасыз ететінін тексеруге бағытталған.

NLP жүйелеріндегі қателерді зерттеу қарапайым жиынтық көрсеткіштерді салыстырудан модель сапасының төмендеуімен жүйелі түрде байланысты факторларды талдауға ауысты. Бұл логикада қате енді өңдеудің қалдық әсері ретінде емес, тілдік материалдың қасиеттері, ұсыну түрі және модельдің архитектуралық шектеулері арасындағы өзара әрекеттесудің нәтижесі ретінде қарастырылады. Морфологиялық тұрғыдан бай тілдер үшін бұл тәсіл әсіресе нәтижелі, өйткені бірыңғай сыртқы нәтиже, мысалы, дұрыс емес POS-тегі немесе қате лемма әртүрлі көздерден туындауы мүмкін.

Тіларалық контексте Боллманн мен Согаардты зерттеуі іргелі маңызға ие [8], мұнда морфологиялық белгілер бірнеше NLP тапсырмаларында қателердің тұрақты предикторлары ретінде әрекет ете алатыны көрсетілген. Авторлар маңызды ескерту жасайды: морфологиялық категориялардың түсіндіруші күші тілдің типологиялық құрылымына байланысты өзгереді. Басқаша айтқанда, флекциялық тілдерге тән қателіктер үлгілері агглютинативті тілдерге механикалық түрде ауыса алмайды, яғни қазақ тіліне өзіндік аналитикалық негіз қажет. Бұл тұжырым екі себеп бойынша нақты зерттеу үшін маңызды. Біріншіден, ол морфологияны модель сапасының түсіндірмелі айнымалысы ретінде қарастыруға болатындығын растайды. Екіншіден, бұл морфологиялық тұрғыдан қанықкан тілдер грамматикалық белгілерді санаудан гөрі қате көздерін нақтырақ ажыратуды қажет ететіндігін көрсетеді [8].

Әдебиеттерге шолу. Морфологиялық қателердің типологиясы деңгейінде CoNLL-SIGMORPHON Shared Task материалында орындалған Gorman және басқалардың еңбектері маңызды бағдар береді [9]. Авторлар лексикалық қасиеттерге, күтпеген инфлекциялық паттерн мен эталондық деректердің ақауларына байланысты қателіктерді ажыратады. Мұнда тек бөлудің өзі ғана емес, сонымен бірге оның әдіснамалық салдары да маңызды: сыртқы көрінісі бойынша бірдей модельдік қателіктің әртүрлі шығу көздері болуы мүмкін, сондықтан оны автоматты түрде біртекті құбылыс ретінде түсіндіруге болмайды. Морфологиялық талдау тапсырмалары үшін бұл тезис ерекше маңызды, өйткені ол зерттеуді қорытынды дәлсіздік деңгейінен сәтсіздіктің құрылымдық себептері деңгейіне дейін көтереді [9].

Қазақ тілін морфологиялық өңдеу саласындағы зерттеулер екі дербес бағытты қалыптастырды. Олардың бірінші бағыты соңғы автоматтар негізінде анализаторлар мен ережеге бағдарланған жүйелерді әзірлеуді қамтыды. Apertium-kaz [10], KazMorph 1.0 [11] және басқа да FST-тәсілдері морфотактиканы формализациялауды қамтамасыз етті және тұрақты нативті лексиканы сенімді талдаудың негізін қалады [10,12,13]. Сонымен бірге, дәл осындай жүйелердің сөздік қоры мен грамматикалық қатаңдығы күтілетін құрылымнан ауытқитын формаларды өңдеудегі олардың осалдығын көрсетті. Осы зерттеу үшін маңыздысы, бұл ауытқулар қиындықтың бір түрімен шектелмейтіні, сондықтан оларды бір ғана қалдық қателіктер класы ретінде қанағаттанарлықтай сипаттау мүмкін емес.

Екінші бағыт қателерді талдау және гибридті архитектураларды құрумен байланысты. Л.М. Байтенова және т.б. [5] жартылай құрылымдалған қазақ мәтіндерінде қателік көздері біркелкі таралмағандығын және материалдың сипатына, соның ішінде орыс орфографиялық кедергісіне және пайдаланушы формаларының өзгергіштігіне айтарлықтай тәуелді екендігін көрсетеді. Л.М. Байтенова және т.б. жұмыстарында [5] FST + CRF + KazRoBERTa архитектурасы морфологиялық тұрғыдан күрделі жағдайларда, соның ішінде кірме сөздерде, омонимдік формаларда және ұзын аффикстік тізбектерде біріктірілген қатені азайтатынын көрсеткен. Сонымен қатар, аталған нәтижелердің практикалық маңыздылығы айтарлықтай олқылықты жоймайды. Әдебиеттерде әлі күнге дейін диагностикаланатын морфологиялық аномалия түрін оның модельдің жұмысына әсерін

түсіндірумен және оны өңдеу үшін аналитикалық компонентті таңдау механизмімен бір уақытта байланыстыратын формальды схема ұсынылмаған.

Зерттелген дереккөздер біркелкі көрініс құрайды, одан үш негізгі бақылаулар туындайды. Біріншіден: морфологиялық мәтін ерекшеліктері NLP жүйелерінің қателерімен статистикалық тұрғыдан корреляцияланады [9] – бұл тіларалық материалдарда расталған. Екіншіден: агглютинативті тілдер үшін мұндай корреляция әмбебап сұлбаны емес, тілге тән типологияны қажет етеді [8, 14]. Үшіншіден: қазақ тілінің гибридті архитектурасы сапаның өлшенетін өсуін қамтамасыз етеді [5].

Сонымен бірге, морфологиялық ауытқулардың қандай нақты кластары трансформерлік өңдеу өнімділігінің негізгі детерминанттары болып табылатыны және олардың нақты диагностикасы жүйенің өнімділігін қалай жақсарта алатыны туралы әдебиеттерде шешілмеген. Бұл зерттеу осы олқылықтың орнын толтыруға бағытталған.

Қазақ тілінің морфологиялық қателері құрылымдық жағынан гетерогенді болып табылады және трансформерлік NLP-модельдерінің өнімділігінің сараланған детерминанттары ретінде әрекет етеді. Іс жүзінде қазақ тіліндегі морфологиялық ауытқулар басқаша әрекет етеді және бұл күрделі модельдеусіз корпус деңгейінде айқын көрінеді. Кірме негізі бар сөз модельді қатарынан жеті жалғанған қосымшасы бар сөзге қарағанда басқаша «бұзады». Сегменттеудің тұрақсыздығы септік пен уақыт белгілерінің қайшылығынан гөрі қателіктің басқа түрін тудырады. Осының барлығын «қиын сөз формаларының» бір класына біріктіру, бұл деректердегі бұрыннан бар ақпараттан әдейі бас тартуды білдіреді.

Сондықтан бұл мақалада морфологиялық ауытқулар құрылымдық жағынан ерекшеленетін әртүрлі құбылыстар тобы ретінде қарастырылады. Олардың әрқайсысы модельге өзінің механизмі арқылы әрекет етеді және бұл механизмді диагностикалауға, сипаттауға және аналитикалық компонентті таңдауда басқару параметрі ретінде пайдалануға болады.

Зерттеу материалдары мен әдістері.

Қазақ тіліндегі морфологиялық қателер табиғаты бойынша гетерогенді болып келеді. Шеттен алу, аффикстік шамадан тыс жүктелу, сегменттік тұрақсыздық және грамматикалық белгілердің қақтығыстары бұл дегеніміз әр түрлі құбылыстар және әрқайсысы өзіндік механизмі арқылы морфологиялық талдау сапасын төмендетеді. Дәл осы айырмашылық осы зерттеудің негізін құрайды. «Қиын формалардың» бір класының орнына мұнда құрылымдық тұрғыдан сараланған аномалиялар тобы анықталады. Олардың әрқайсысы бөлек диагностикалауды және бөлек аналитикалық шешімді қажет етеді.
Әрбір x_i токені үшін белгілердің векторы есептеледі:

$$F_i = (r_i, a_i, n_i, l_i, v_i, z_i) \quad (1)$$

Онда
r_i – сөздің негізі,
a_i – аффикстердің тізбегі,
n_i – аффикстер саны,
l_i – негіздің ұзындығы,
v_i – фонологиялық ауытқудың көрсеткіші,
z_i – кірме сөздердің индикаторы.
Осы параметрлер негізінде аномалия түрі анықталады:

$$e_i = \Phi(F_i). \quad (2)$$

Мұнда e_i BOR, AFC, SEG, AGR немесе NONE мәндерінің бірін қабылдайды:
BOR: шеттен алу (заимствование)

Бұл класқа сыртқы шығу тегінің немесе графикалық-фонологиялық интерференция белгілерін көрсететін токендер кіреді. Оларды бағалау үшін келесі көрсеткіш қолданылады:

$$B_i = f_B(z_i, v_i, g_i). \quad (3)$$

Онда g_i қосымша графикалық маркерлерді көрсетеді.

B_i - мәні, неғұрлым жоғары болса, токеннің кірме сөздер класына тиесілі ықтималдылығы соғұрлым жоғары болады.

АФК: аффиксалды күрделілік

Бұл класс терең немесе типтік емес құрылымдық аффиксі бар сөз формаларын біріктіреді.

$$A_i = f_A(n_i, m_i, v_i) \text{ көрсеткіші} \quad (4)$$

Онда, m_i – токеннің класқа қосылу ықтималдығын сипаттайтын аффикстік аймақтың ұзындығы.

n_i және m_i мәндерінің жоғарылауы құрылымдық күрделіліктің артқанын көрсетеді.

SEG: сегментациялық бұзылу

Бұл класс морфемалардың шекараларын тұрақсыз қалпына келтіру жағдайларын көрсетеді. Диагностикалық көрсеткіш ретінде анықталады:

$$S_i = f_S(l_i, d_i). \quad (5)$$

Мұндағы d_i қалпына келтірілген және бақыланатын сөз формалары арасындағы сәйкессіздікті білдіреді.

AGR: грамматикалық қайшылық

AGR класы грамматикалық белгілердің үйлесімсіздігімен байланысты. Сәйкестік коэффициенті формула бойынша есептеледі:

$$G_i = 1 - c_i, c_i \in [0,1]. \quad (6)$$

c_i мәні неғұрлым төмен болса, белгілер қақтығысының ықтималдығы соғұрлым жоғары болады.

NONE: бейтарап класс

Егер диагностикалық көрсеткіштердің ешқайсысы жіктеу шегінен аспаса, токен *NONE* мәнін алады:

$$e_i = \text{NONE} \iff \max(B_i, A_i, S_i, G_i) < \theta. \quad (7)$$

Мұнда, θ – жіктеу шегі.

Класты таңдау ережесі

Токен аномалиялардың бірнеше класының белгілерін бірден анықтаған кезде, диагностикалық көрсеткіші ең жоғары болып табылатын белгілерге басымдық береді:

$$e_i = \arg\max(B_i, A_i, S_i, G_i). \quad (8)$$

Осылайша, бұл типология морфологиялық қателерді сипаттау құралы ғана емес, сонымен қатар KazMorphCorpus-2026 жүйесі ішіндегі кейінгі талдауды бағыттаудың жұмыс механизмі ретінде қызмет етеді.

Зерттеудің әдіснамалық логикасы лингвистикалық, есептеу және салыстырмалы тәсілдердің үйлесіміне негізделген. Бастапқы міндет тек қазақ мәтінінің морфологиялық талдауының дәлдігін арттыру ғана емес, сонымен қатар трансформерлік NLP-модельдердің өнімділігін төмендетуге морфологиялық ауытқулардың қандай түрлері үлкен үлес қосатынын анықтау болып табылады. Қойылған міндеттерді шешу үшін өзара байланысты әдістер кешені құрылды, олардың әрқайсысы зерттеудің жалпы логикасында тәуелсіз функцияны орындайды (кесте 1).

Кесте 1 – Зерттеу әдістері және олардың зерттеу функциясы

Әдіс	Қолданылуы	Зерттеу функциясы
Формальды-лингвистикалық талдау	Морфологиялық ауытқулардың диагностикалық белгілерін оқшаулау	Қателер типологиясын қалыптастыру
Корпустық талдау	UD Kazakh-KTB стандартты белгіленген корпус ретінде пайдалану	Репродуктивтілік пен салыстырмалылықты қамтамасыз ету
Есептік модельдеу	FST + CRF + KazRoBERTa + MFRN гибридті архитектурасын қолдану	Қате түрлерінің бірдей емес әсері туралы гипотезаны тексеру
Нысандырылған маршруттау	Диагностикаланған аномалия түрі бойынша аналитикалық компонентті таңдау	Таксономияны талдауды басқару механизміне айналдыру
Сандық бағалау	Accuracy, Macro-Precision, Macro-Recall және Macro-F1 көрсеткіштері бойынша модельдің сапасын сандық бағалау	Сапаны өлшеу, конфигурацияларды салыстыру
Абляциялық талдау	Конфигурацияларды модульдердің әртүрлі құрамымен салыстыру	Таксономия мен маршруттау үлесін бағалау

Бірінші деңгей ресми лингвистикалық талдау әдістерінен тұрады. Бұл кезеңде қазақ мәтінінде кездесетін морфологиялық ауытқулардың тұрақты түрлері анықталып, олардың диагностикалық белгілері белгіленеді. Бұл тәсіл қателерді жиынтық қарастырудан олардың құрылымдық типологиясына көшуге мүмкіндік береді. Зерттеу шеңберінде морфологиялық аномалия беткі сөз формасы мен токеннің күтілетін морфологиялық ұйымы арасындағы байқалатын алшақтық ретінде түсіндіріледі. Оны сипаттау үшін түбірлік құрылымның белгілері, аффикс тізбегінің тереңдігі, сегменттеу сипаты, алынған элементтердің болуы және грамматикалық сипаттамалардың үйлесімділігі қолданылады.

Екінші деңгей корпустық әдіспен ұсынылған. Барлық есептеулер мен салыстырулар қазақ тілі үшін стандартты белгіленген корпус ретінде пайдаланылатын Universal Dependencies Kazakh-KTB материалында орындалады [15]. Корпус талдау сапасын салыстырмалы форматта бағалауға мүмкіндік береді және POS санаттарының, леммалардың және морфологиялық белгілердің бірыңғай жиынтығын қамтамасыз етеді. Бұл ресурсты пайдалану өте маңызды, өйткені мақаланың міндеті жеке мысалды көрсету емес, танылған эталонда қайталанатын нәтиже алу болып табылады.

Үшінші деңгей есептік моделдеу әдістерін құрайды. Гипотезаны тексеру үшін KazMorphCorpus-2026 гибридті архитектурасы қолданылады, оның ішінде ережеге бағытталған FST талдауы, CRF дизамбигуациясы, KazRoBERTa трансформерлік модулі және MFRN детерминирленген морфологиялық үйлесімділік модулі. Осы зерттеуде KazMorphCorpus-2026 архитектурасы дәлдікті жақсарту құралы ретінде ғана емес, сонымен қатар бақыланатын эксперименттік орта ретінде де қолданылады. Әртүрлі аномалия кластарындағы жеке компоненттердің мінез-құлқын бақылау арқылы берілген аномалия түрі үшін қайсысы ең тиімді екенін анықтауға болады. Осылайша зерттеудің орталық гипотезасын тексеруге болады.

Төртінші деңгей формальды маршруттау әдісімен байланысты. Диагностикалық ерекшеліктерді есептегеннен кейін әрбір токенге морфологиялық аномалия класы салыстырылады, содан кейін ауытқудың осы түрімен эмпирикалық түрде жақсырақ жұмыс істейтін аналитикалық компонент таңдалады. Бұл қате түрін талдаудан кейінгі қорытынды белгі ретінде емес, бүкіл жүйенің ішіндегі түсіндіру және басқару параметрі ретінде

пайдалануға мүмкіндік береді. Мұндай қадам зерттеу мақсатына жету үшін өте маңызды, себебі ол қателіктер таксономиясын сипаттамалық жазықтықтан операциялық жазықтыққа ауыстырады.

Бесінші деңгей сандық бағалау әдістерінен тұрады. Сапаны өлшеу үшін Accuracy, Macro-Precision, Macro-Recall және Macro-F1 қолданылады. Бұл көрсеткіштер, бір жағынан, ұсынылған жүйені базалық модельдермен салыстыруға, ал екінші жағынан, морфологиялық ауытқулардың жеке кластары бойынша архитектураның мінез-құлқын бағалауға мүмкіндік береді.

Сонымен қатар, абляциялық талдау жүргізіледі. Бұл бізге архитектуралық гибридтіліктің өзіндік үлесі мен маршруттаудың таксономиялық деңгейінің алынған сапа артуына қосқан үлесін ажыратуға мүмкіндік береді [16, 17]. Қолданылатын әдістемелік кешен дәлдікті өлшеу шеңберінен шығатын мәселені шешуге бағытталған. Морфологиялық ауытқулардың қай кластары трансформерлік модельдің өнімділігін неғұрлым анықтайтынын және қате түрінің нақты диагностикасы талдау нәтижесін қаншалықты анықтайтынын анықтау қажет.

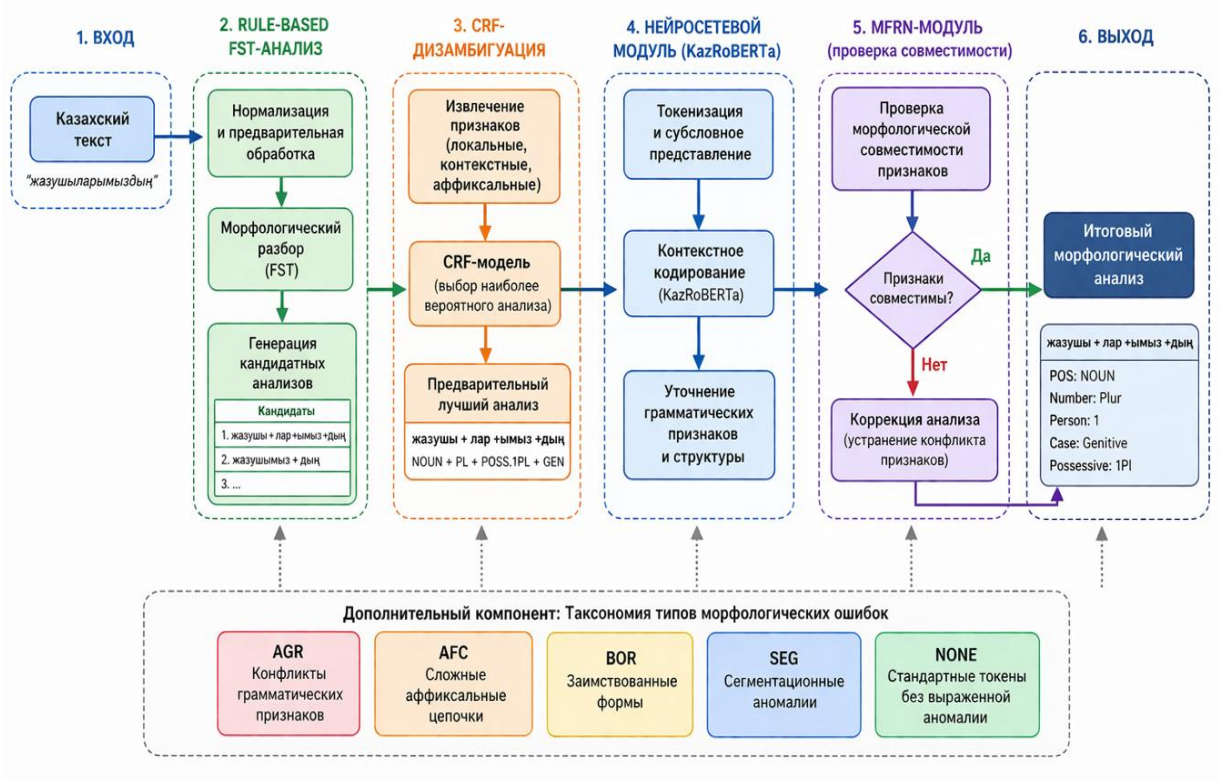
Көрсетілген мақсатқа ауытқулардың құрылымдық диагностикасы, таксономиялық бақыланатын маршруттауды есептік модельдеуі және абляциялық бақылауды қолдана отырып, жүйе конфигурацияларын салыстырмалы бағалау арқылы дәйекті түрде қол жеткізіледі.

Бастапқы қадам морфологиялық қателіктерді сапаның төмендеуінің жалғыз көзі ретіндегі жалпы тұжырымдамадан олардың рәсімделген саралауға көшуді қамтиды. Нәтижесінде, морфологиялық аномалия дұрыс емес талдаудың бір ғана мысалы ретінде емес, байқалатын сипаттамалары бар типтік диагностикалық бірлік ретінде анықталады.

Әрі қарай, таңдалған типология операциялық формаға ауыстырылады және есептеу схемасы ішінде қолданылады. Әрбір аномалия класы үшін токенді белгілі бір ауытқу түрімен автоматты түрде байланыстыруға мүмкіндік беретін ерекшеліктер орнатылады. Бұл аналитикалық компоненттер арасындағы рөлдердің функционалды таралуын мүмкіндік етеді: ережеге-бағдарланған анализатор морфемалық шекаралар негізгі болып табылатын жерде; CRF-модулі - контекстке тәуелді екіұштылықта, ең алдымен шеттен алынған және омонимдік формаларда; трансформерлік деңгей - ұзын аффиксалды тізбектерді және күрделі морфосинтаксистік конфигурацияларды өңдеу үшін; ал MFRN модулі белгілердің сәйкестігін тексеруден кейін орындауда қолданылады.

Гипотеза корпус материалында бір-бірін толықтыратын екі құрал арқылы тексеріледі: жүйенің әртүрлі конфигурацияларымен салыстырмалы эксперименттер және жекелеген компоненттердің үлестерін абляциялық талдау. Мұндай дизайн гибридті архитектураның жалпы әсерін таксономиялық басқарылатын маршруттау нәтижесінде пайда болатын әсерден ажыратуға мүмкіндік береді. Осының арқасында сапаның түпкілікті артуы жаңа модульдерді қосудың қарапайым салдары ретінде емес, кіріс аномалиясының құрылымы мен оны өңдеу әдісі арасындағы дәлірек сәйкестіктің нәтижесі ретінде түсіндіріледі.

Жүйенің функционалдық ұйымдастырылуы қорытынды морфологиялық талдау бір аналитикалық модульмен емес, диагностикаланатын аномалия түріне байланысты өңдеудің бірнеше деңгейлерінің үйлестірілген өзара әрекеттесуінің нәтижесімен анықталады деп болжайды. Жүйенің функционалдық ұйымдастырылуы және токеннің аналитикалық компоненттер арқылы өту бағыты 1-суретте көрсетілген.



Сурет 1 – KazMorphCorpus-2026 жүйесінің архитектурасы және морфологиялық ауытқулар кластары бойынша маршруттау механизмі

Ұсынылған схема KazMorphCorpus-2026 жұмысының логикасын көрсетеді және диагностикаланатын морфологиялық аномалия түрі аналитикалық маршрутты таңдау процедурасына қалай енгізілетінін көрсетеді. Бұл ұйымдастыру жүйенің құрылымын сипаттаудан оның корпустық деректердегі нақты тиімділігін бағалауға көшуге мүмкіндік береді.

Нәтижелері және оларды талқылау.

Эксперимент нәтижелері морфологиялық қателердің трансформерлік өңдеу өнімділігіне әсері біркелкі таралмағанын көрсетеді. UD Kazakh-KTB сынақ үлгісінде жүйенің толық конфигурациясы $Accuracy = 87{,}4\%$ және $Macro - F1 = 0{,}86$ жетеді. Алайда, нәтижелерді интерпретациялау үшін көрсеткіштің қорытынды мәні емес, қателер кластары арасындағы айырмашылықтар анағұрлым маңызды болып табылады, өйткені олар модельге біркелкі емес аналитикалық жүктемені қалыптастырады.

Қорытынды деректерге сәйкес, AGR және AFC сыныптары үшін сараланған өңдеуден кейін айқын өсім байқалады. AGR жағдайында бұл грамматикалық белгілердің қайшылығы талдаудың ішкі дәйектілігіне тікелей әсер ететіндігімен байланысты болуы мүмкін, сондықтан қорытынды классификацияға көбірек әсер етеді. AFC үшін тұрақсыздықтың ықтимал көзі терең аффиксация болып табылады, онда грамматикалық ақпарат бірнеше морфемалық позицияларға бөлінеді және стандартты шартты көрініс нашар сақталады. Берілген іріктеу шеңберінде дәл осы сыныптар дифференциалды өңдеуге ең жоғары сезімталдықты көрсетеді.

BOR класы орташа айқын, бірақ тұрақты әсерді көрсетеді. Кірме формалары міндетті түрде сапаның максималды жоғалуына әкелмейді. Дегенмен, олар жүйелі түрде аналитикалық тұрақсыздық қаупін арттырады, себебі олар әдеттегі фонологиялық және сөзжасамдық үлгілерден ішінара аутқиды. SEG класы AGR және AFC-ке қарағанда аз зиян келтіреді, бірақ сапаның нашарлауына ықпал етеді. Бұл сегменттеу бұзушылықтары да

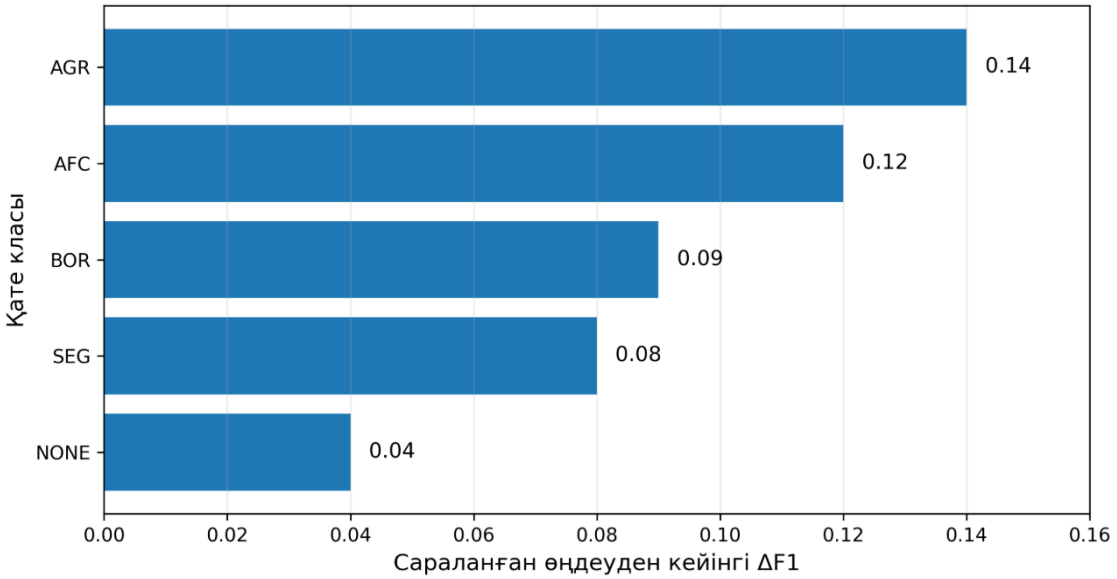
бөлек қарастыруды қажет ететінін және оларды біріктірілген көрсеткіштерге қоспау керектігін көрсетеді.

2-ші кестеде келтірілген мәндерге сәйкес, $\Delta F1$-дің ең аз өсуі $NONE$ класы үшін байқалады. Интерпретациялық тұрғыдан алғанда, бұл дифференциалды өңдеудің әсері негізінен құрылымдық анықталған морфологиялық аномалиялары бар токендерде шоғырланғанын, ал стандартты сөз формалары бастапқыда сенімдірек өңделетінін көрсетеді. Бұл тұрғыда морфологиялық қателік түрі нәтижені сипаттау құралы ретінде ғана емес, сонымен қатар модель өнімділігінің өзгергіштігімен байланысты фактор ретінде де әрекет етеді. Дегенмен, бұл тұжырымды ағымдағы іріктеу және қолданылатын таксономия схемасы аясында қарастырған жөн.

Кесте 2 – Морфологиялық қателік типтерінің модель сапасына әсерінің таралуы

Класы	**Үлесі, %**	**F1**	**ΔF1**	**Әсер ету күшін түсіндіру**
AGR	4,3	0,82	+0,14	Сапаның төмендеуінің ең күшті детерминанты
AFC	9,8	0,85	+0,12	Терең аффиксацияға байланысты жоғары әсер
BOR	11,3	0,87	+0,09	Интерференция арқылы орташа әсер ету
SEG	6,4	0,89	+0,08	Локализацияланған, бірақ тұрақты әсер
NONE	68,2	0,94	+0,04	Минималды пайда; стандартты токендер

Алынған деректерді жинақтап алғанда, қазақ тіліндегі морфологиялық күрделілік трансформерлік өңдеу үшін біртекті қиындық көзі емес деген қорытынды жасауға мүмкіндік береді. Оның әсері ауытқудың нақты түріне байланысты. Белгілердің қақтығысы, терең аффиксация, кірме құрылым және сегменттеу тұрақсыздығы аналитикалық жүктеменің әртүрлі режимдерін тудырады, сондықтан модельдің сапасына әртүрлі әсер етеді. Осы себепті қазақ NLP-жүйелерін бағалау тек Accuracy және Macro-F1 жалпы мәндерімен шектелмеуі тиіс. Сөз формаларын қай түрлері негізгі қауіп аймағын құрайтынын көруге мүмкіндік беретін сыныптық талдау неғұрлым мағыналы көрініс береді (2-сурет).



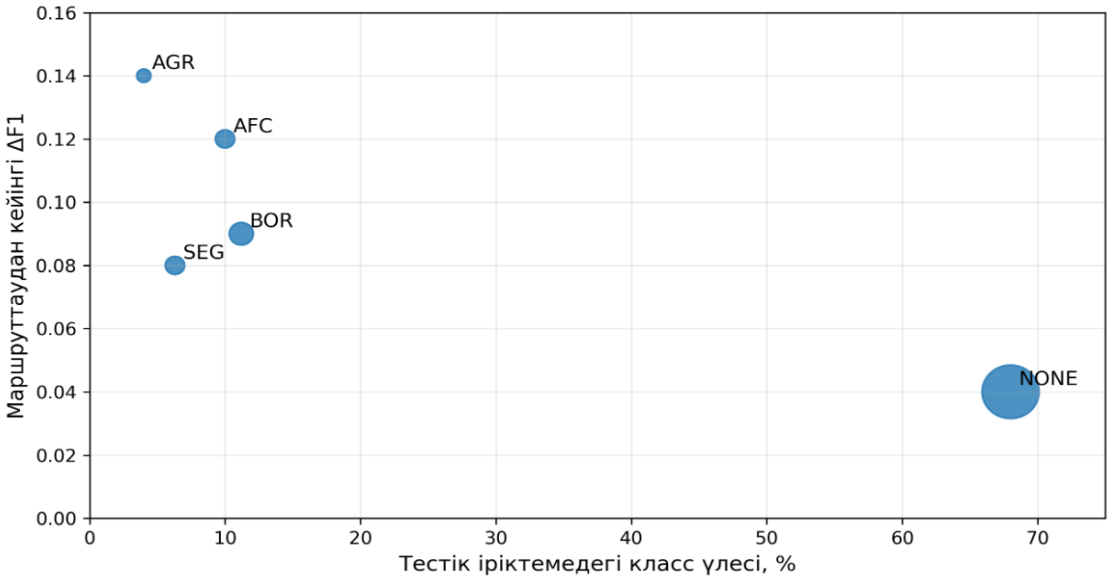
Сурет 2 – Дифференциалданған өңдеуден кейінгі морфологиялық қателер класы бойынша F1 өсуі

Ең үлкен ұтыс AGR және AFC кластары үшін тіркелген. Барлық диагностикалық кластардың ішінде AGR және AFC маршруттауға ең жоғары сезімталдықты көрсетеді. Олар үшін сапаның өсуі тіркелген, бұл олардың трансформерлік модель өнімділігінің нашарлауындағы шешуші рөлін көрсетеді (3-кесте).

Кесте 3 – Өнімділік детерминанттары ретінде қателіктер типтерінің мазмұнды сипаттамасы

Кластар	**Қиындықтың негізгі механизмі**	**Трансформерлік өңдеуге әсері**	**Практикалық қорытынды**
AGR	Морфологиялық белгілердің қайшылығы	Талдаудың ішкі дәйектілігін бұзады	Белгілердің үйлесімділігін тексеруді күшейту
AFC	Ұзын аффиксалды реттілік	Бағыныңқы сөздердің ықшамдылығын көрсетеді	Терең аффиксацияның арнайы диагностикасы қажет
BOR	Графикалық-фонологиялық құрылымның интерференциясы	Контекстік және лексикалық тұрақсыздықты арттырады	Кірме негіздерін бөлек қарастырған жөн
SEG	Морфемалық шекаралардың бұзылуы	Негізді және аффиксалды құрылымды қалпына келтіруді бұрмалайды	Бөлек сегменттеу бақылауы ұсынылады
NONE	Айқын морфологиялық аномалияның болмауы	Айқын морфологиялық аномалияның болмауы	Айқын морфологиялық аномалияның болмауы

Нәтижелерді интерпретациялау ΔF1 абсолютті мәндерімен шектелмейді. Сапа өсімін үлгідегі сынып үлесімен салыстыру – мысалы, 4,3% AGR ΔF1 = +0,14 береді, ал NONE 68,2% тек +0,04 береді – бұл шын мәнінде аномалиялардың қандай түрлері модель өнімділігінің өзгермелілігін анықтайтынын түсіну үшін түбегейлі маңызды диспропорцияны анықтайды. Бұл байланыс 3-ші суретте көрсетілген.



Сурет 3 – Әсер ету картасы (іріктемедегі класс үлесінің ΔF1 өсуіне қатынасы)

Ұсынылған деректер қазақша NLP-модельдерін бағалау кезінде тек Accuracy және Macro-F1 жиынтық көрсеткіштерімен шектелу жеткіліксіз екенін көрсетеді. Сыныптық талдау жинақталған көрсеткіштерге қарағанда мағынааралық көрініс береді. Бұл аналитикалық тәуекелдің негізгі аймағы терең аффиксациясы сөз формалары мен грамматикалық белгілердің қақтығыстарынан құралатынын, ал NONE класының стандартты токендері нәтижелердің өзгермелілігіне іс жүзінде ешқандай әсер етпейтінін көрсетеді [18]. Осыдан практикалық қорытынды шығады: морфологиялық талдау жүйелерін әзірлеу және валидациялау үшін корпустық іріктемені қалыптастыру кезінде AGR, AFC, BOR және SEG кластарының формаларын жеткілікті түрде ұсынуды қамтамасыз еткен жөн. Олар морфологиялық біртекті емес мәтін жағдайында модельдің нақты тұрақтылығын анықтайды.

Қорытынды.

Жүргізілген зерттеу морфологиялық қателіктердің түрлері трансформерлік NLP модельдерінің өнімділігінің тәуелсіз детерминанттары ретінде әрекет ететінін растайды. Агглютинативті тілдегі морфологиялық талдау сапасының төмендеуі құрылымдық тұрғыдан сараланған сипатқа ие екені анықталды. Аномалиялардың әртүрлі кластары – AGR, AFC, BOR, SEG – модельдің өнімділігін әртүрлі механизмдер арқылы және әртүрлі дәрежеде төмендетеді. Нәтижесінде, олардың жиынтық сипаттамасы, бірыңғай жиынтық көрсеткіш арқылы байқалған ауытқулардың нақты сипатын көрсетпейді.

Сапаның төмендеуіне ең үлкен әсер ететін грамматикалық белгілердің қайшылықтары (AGR) және күрделі аффикстік тізбектер (AFC) болып табылады – дәл осы кластар сараланған өңдеу кезінде сапаның ең үлкен өсуін көрсетеді (ΔF1 = сәйкесінше +0,14 және +0,12). Кірме нысандар (BOR) және сегменттеудің бұзылуы (SEG) бөлек есепке алуды қажет ететін аналитикалық тұрақсыздықтың азырақ бұзылатын, бірақ тұрақты аймағын құрайды. NONE класы стандартты таңбалауыштардың айқын ауытқуларсыз архитектураның негізгі компоненттерімен жоғары тұрақтылықпен өңделетіндігін растайды.

Алынған нәтижелердің практикалық маңыздылығы толық гибридті маршруттау жүйелерімен шектелмейді. Аномалиялардың құрылымдық типтерін айқын ажырату жобалау және валидация кезеңі жиынтық көрсеткіштер шеше алмайтын мәселені шешуге мүмкіндік береді: тұрақсыздықтың белгілі бір көзін локализациялау – бұл грамматикалық ерекшеліктердің қайшылығы, терең аффиксация немесе сегменттеудің бұзылуы – және

мақсатты түрде модельдің мінез-құлқын түзету енгізу формаларының осы класына қолданылады.

Зерттеулердің даму келешегі ұсынылған таксономияны басқа жанрлық сипаттағы корпустарда тексерумен, біріктірілген жағдайларда кластар арасындағы шекараларды нақтылаумен және тәсілді басқа түркі тілдеріне ауыстырумен байланысты. Ең алдымен, морфологиялық қиындықтардың анықталған таралуы бір корпус пен бір модель конфигурациясының шегінен тыс қаншалықты тұрақты түрде қайталанатынын анықтау.

Алғыс: *Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті (грант № ЖТН AP23487753) «Қазақ тіліндегі мәтіндерді автоматтандырылған түзетуге арналған инновациялық технологиялар: машиналық оқыту және морфологиялық талдау» жоба аясында қаржыландырды.*

Әдебиеттер тізімі

1. Tyers, F.M. & Washington, J.N. (2015). Towards a Free/Open-source Universal-dependency Treebank for Kazakh // Proc. of the 3rd International Conf. on Turkic Languages Processing (TurkLang 2015). Kazan: Kazan Federal University, P. 276–289. URL: https://www.antat.ru/ru/ips/science/editions/Turklang_2015.pdf .

2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, P. 4171–4186. DOI: 10.18653/v1/N19-1423.

3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Yu, L., Donnec, J. & Lample, G. (2020). Unsupervised Cross-lingual Representation Learning at Scale // Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, P. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

4. Johanson, L. & Csató, É.Á. (Eds.). (2021). The Turkic Languages. 2nd ed. London; New York: Routledge, 980 p. DOI: 10.4324/9781003243809.

5. Baitenova, L., Tussupova, S., Mambetov, S., Munaitbas, G. & Mukhamejanova, G. (2025). Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language // Frontiers in Artificial Intelligence.Vol. 8. Art. 1708566. DOI: 10.3389/frai.2025.1708566.

6. Nivre, J., Agić, Š., Ahrenberg, L., Antonsen, M., Aranzabe, A., Asahara, M., Ataman, D., Badmaev, B., Ballesteros, M., Banerjee, E. & Bank R. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proc. of the 12th International Conf. on Language Resources and Evaluation (LREC 2020). Marseille: European Language Resources Association, P. 4034–4043. ISBN 979-10-95546-34-4.

7. Makazhanov, A., Sultangazina, A., Makhambetov, O. & Yessenbayev, Zh. (2015). Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines // Proc. of the 3rd International Conf. on Turkic Languages Processing (TurkLang 2015). Kazan: Kazan Federal University, P. 338–350.

8. Bollmann, M. & Søgaard, A. (2021). Error Analysis and the Role of Morphology // Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, P. 1887–1900. DOI: 10.18653/v1/2021.eacl-main.162.

9. Gorman, K., McCarthy, A.D., Cotterell, R., Hulden, M., Kementchedjhieva, Y., Liu, Y., Muriuki, G. & Wisniewski, G. (2019). Weird Inflects but OK: Making Sense of Morphological Generation Errors // Proc. of the 23rd Conf. on Computational Natural Language Learning (CoNLL 2019). Hong Kong: Association for Computational Linguistics, P. 140–151. DOI: 10.18653/v1/K19-1014.

10. Washington, J.N., Salimzyanov, I. & Tyers, F.M. (2014). Finite-state morphological transducers for three Kypchak languages // Proc. of the Ninth International Conf. on Language Resources and Evaluation (LREC'14). Reykjavik: European Language Resources Association (ELRA), P. 3378–3385. URL: https://aclanthology.org/L14-1143/.

11. Makhambetov, O., Makazhanov, A., Yessenbayev, Zh. & Sultangazina, A.A. (2013). Knowledge-Based Approach to Kazakh Part-of-Speech Tagging // Intelligence and Security Informatics: 8th International Workshop, ISI 2013. Seattle: Springer, P. 139–145. DOI: 10.1007/978-3-642-39693-7_15.

12. Beesley, K.R. & Karttunen, L. (2003). Finite State Morphology. Stanford: CSLI Publications, 528 p. ISBN 978-1-57586-434-9.

13. Tyers, F.M., Washington, J.N. & Salimzyanov, I. (2020). Finite-State Morphological Transducer for Kazakh // Language Resources and Evaluation. Vol. 54, No. 2. P. 523–548. DOI: 10.1007/s10579-019-09474-2.

14. Uz, H. & Eryiğit, G. (2023). Towards Automatic Grammatical Error Type Classification for Turkish // Proc. of the 17th Conf. of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. Dubrovnik: Association for Computational Linguistics, P. 134–142. DOI: 10.18653/v1/2023.eacl-srw.14.

15. UD Kazakh-KTB [Электрондық ресурс] // Universal Dependencies. URL: https://universaldependencies.org/treebanks/kk_ktb/index.html.

16. Pan, Y., Li, X., Yang,Y. & Dong, R. (2020). Morphological Word Segmentation on Agglutinative Languages for Neural Machine Translation // arXiv. arXiv:2001.01589 [cs.CL]. DOI: 10.48550/arXiv.2001.01589. URL: https://arxiv.org/abs/2001.01589.

17. Ratinov, L. & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition // Proc. of the Thirteenth Conf. on Computational Natural Language Learning (CoNLL-2009). Boulder, Colorado: Association for Computational Linguistics, P. 147–155.

18. Haslett, D.A. (2025). Tokenization Changes Meaning in Large Language Models: Evidence from Chinese // Computational Linguistics. Vol. 51, No. 3. P. 785–814. DOI: 10.1162/coli_a_00557.

ТАКСОНОМИЯ МОРФОЛОГИЧЕСКИХ ОШИБОК КАК МЕХАНИЗМ МАРШРУТИЗАЦИИ В ГИБРИДНЫХ NLP-СИСТЕМАХ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Аннотация. *Казахский язык устроен так, что грамматическая информация распределяется по цепочке аффиксов, а не концентрируется в отдельных словах. Это само по себе создаёт трудности для автоматических систем. Ситуацию дополнительно осложняет постоянный приток русских и английских заимствований, которые частично выбиваются из фонологической логики языка — и в итоге морфологический анализатор сталкивается не с одним типом трудных форм, а с несколькими принципиально разными. Существующие системы оценивают производительность NLP-моделей агрегированно, не разграничивая типы аномалий по механизму их воздействия на анализ, что не позволяет целенаправленно управлять обработкой. Предметом настоящего исследования является связь между типами морфологических ошибок и показателями производительности трансформерных NLP-моделей для казахского языка.*

Цель работы – разработать формализованную таксономию морфологических аномалий и интегрировать её в гибридную NLP-архитектуру в качестве механизма маршрутизации аналитических компонентов. В исследовании применяются формально-лингвистический анализ, корпусный метод на материале Universal Dependencies Kazakh-KTB (1 047 предложений), вычислительное моделирование и абляционный анализ. Гибридная архитектура KazMorphCorpus-2026 объединяет rule-based FST-анализ, CRF-дизамбигуацию, трансформерный модуль KazRoBERTa и модуль проверки

морфологической совместимости признаков MFRN. По результатам исследования предложена пятиклассовая таксономия морфологических аномалий – заимствования (BOR), аффиксальная сложность (AFC), сегментационные нарушения (SEG), конфликты грамматических признаков (AGR) и нейтральный класс (NONE), – интегрированная в систему в качестве управляющего механизма маршрутизации. На тестовой выборке система достигает Accuracy = 87,4% и Macro-F1 = 0,86; наибольший прирост качества зафиксирован для классов AGR (ΔF1 = +0,14) и AFC (ΔF1 = +0,12). Проведённый эксперимент подтвердил: разные типы морфологических аномалий по-разному сказываются на работе трансформерной модели, и это различие имеет практическое значение. Системы, которые диагностируют тип аномалии до анализа и направляют токен к подходящему компоненту, дают результат, который проще интерпретировать и легче улучшить целенаправленно.

Ключевые слова: *таксономия морфологических ошибок, морфологическая маршрутизация, гибридная NLP-архитектура, трансформерные модели, KazRoBERTa, FST, CRF, MFRN, агглютинативный язык.*

MORPHOLOGICAL ERROR TAXONOMY AS A ROUTING MECHANISM IN HYBRID NLP SYSTEMS FOR THE KAZAKH LANGUAGE

Abstract. *The Kazakh language is designed so that grammatical information is distributed along a chain of affixes, rather than concentrated in individual words. This in itself creates difficulties for automated systems. The situation is further complicated by the constant influx of Russian and English borrowings, which are partially dislodged from the phonological logic of the language, and as a result, the morphological analyzer is faced with not one type of difficult forms, but with several fundamentally different ones. Existing systems evaluate the performance of NLP models in an aggregated manner, without distinguishing the types of anomalies by the mechanism of their impact on the analysis, which does not allow targeted management of processing. The subject of this study is the relationship between the types of morphological errors and performance indicators of transformative NLP models for the Kazakh language.*

The aim of the work is to develop a formalized taxonomy of morphological anomalies and integrate it into a hybrid NLP architecture as a routing mechanism for analytical components. The study uses formal linguistic analysis, the corpus method based on the Universal Dependencies Kazakh-KTB (1,047 sentences), computational modeling and ablative analysis. The KazMorphCorpus-2026 hybrid architecture combines rule-based FST analysis, CRF disambiguation, the KazRoBERTa transformer module and the MFRN feature morphological compatibility verification module. Based on the results of the study, a five–class taxonomy of morphological anomalies was proposed – borrowings (BOR), affixal complexity (AFC), segmentation disorders (SEG), conflicts of grammatical features (AGR) and neutral class (NONE), integrated into the system as a control routing mechanism. In the test sample, the system reaches Accuracy = 87.4% and Macro-F1 = 0.86; the largest increase in quality was recorded for the AGR (ΔF1 = +0.14) and AFC (ΔF1 = +0.12) classes. The experiment confirmed that different types of morphological anomalies have different effects on the operation of the transformer model, and this difference is of practical importance. Systems that diagnose the type of anomaly prior to analysis and direct the token to a suitable component produce a result that is easier to interpret and easier to improve purposefully.

Keywords: *morphological error taxonomy, morphological routing, hybrid NLP architecture, transformer models, KazRoBERTa, FST, CRF, MFRN, agglutinative language.*

Сведение об авторах

Байтенова Лаура Маратовна	д.э.наук, профессор Высшей школы информационных технологий, Университет Туран, г. Алматы, Казахстан, E-mail: l.baitenova@turan-edu.kz
Тусупова Сауле Амангелдиевна	д.т.наук, профессор Высшей школы информационных технологий, Университет Туран, г. Алматы, Казахстан, E-mail: s.tussupova@turan-edu.kz
Мухамеджанова Гульнар Сайлаубаевна	Магистр информационных систем, ст.преподаватель школы Цифровых технологий, Университет Нархоз, г. Алматы, Казахстан, E-mail: gulnar.mukhamedzhanova@narxoz.kz
Мунайтбас Гаухар Нуртазакызы	Магистр технических наук, Начальник Отдела стандартизации и защиты информации Департамента информационной безопасности АО «Home Credit Bank», г. Алматы, Казахстан, E-mail: gmunaitbas@gmail.com
Нұртаза Дарын Нұртазаұлы	Магистрант 1 курса ОП «Информационные системы» (научно-педагогическое направление), Университет «Туран», г. Алматы, Казахстан, E-mail: 25262929@turan-edu.kz

Авторлар туралы мәлімет

Байтенова Лаура Маратовна	э. ғ. д., Ақпараттық технологиялар жоғары мектебінің профессоры, Тұран университеті, Алматы қ., Қазақстан, E-mail: l.baitenova@turan-edu.kz
Тусупова Сауле Амангелдиевна	т. ғ. д., Ақпараттық технологиялар жоғары мектебінің профессоры, Тұран университеті, Алматы қ., Қазақстан, E-mail: s.tussupova@turan-edu.kz
Мухамеджанова Гульнар Сайлаубаевна	Ақпараттық жүйелер магистрі, Цифрлық технологиялар мектебінің аға оқытушысы, Нархоз университеті, Алматы қ., Қазақстан E-mail: gulnar.mukhamedzhanova@narxoz.kz
Мунайтбас Гаухар Нуртазақызы	Техника ғылымдарының магистрі, «Хоум Кредит Банк» АҚ Ақпараттық қауіпсіздік департаментінің стандарттау және ақпаратты қорғау бөлімінің бастығы, Алматы қ., Қазақстан, E-mail: gmunaitbas@gmail.com
Нұртаза Дарын Нұртазаұлы	«Ақпараттық жүйелер» білім беру бағдарламасының 1 курс магистранты (ғылыми-педагогикалық бағыт), Тұран университеті, Алматы қ., Қазақстан, E-mail: 25262929@turan-edu.kz

Information about the authors

Baitenova Laura	Doctor of Economics, Professor, Higher School of Information Technologies, Turan University, Almaty, Kazakhstan, E-mail: l.baitenova@turan-edu.kz
Tussupova Saule	Doctor of Technical Sciences, Professor, Higher School of Information Technologies, Turan University, Almaty, Kazakhstan, E-mail: s.tussupova@turan-edu.kz
Mukhamedzhanova Gulnar	Master of Information Systems, Senior Lecturer, School of Digital Technologies, Narxoz University, Almaty, Kazakhstan, E-mail: gulnar.mukhamedzhanova@narxoz.kz
Munaitbas Gaukhar	Master of Technical Sciences, Head of Standardization and Information Protection Departament of Information Security Department «Home Credit Bank» JSC, Almaty, Kazakhstan, E-mail: gmunaitbas@gmail.com
Nurtaza Daryn	1st-year Master's student in Information Systems (scientific and pedagogical direction), Turan University, Almaty, Kazakhstan, E-mail: 25262929@turan-edu.kz