ӘОЖ 004.8+004.912:61:575
ҒТАХР 28.23.21

https://doi.org/10.55452/1998-6688-2026-23-2-171-186

[1,2]Самбетбаева М.А.,
PhD, қауымдастырылған профессор, ORCID ID: 0000-0001-9358-1614,
e-mail: sambetbayeva_ma_1@enu.kz,
[1,2]*Серикбаева С.К.,
PhD, доцент м.а., ORCID ID: 0000-0002-3627-3321,
*e-mail: inf_8585@mail.ru
[1,3]Сұлтанғазиева А.Н.,
магистр, доктарант, ORCID ID: 0009-0009-9038-5234,
e-mail: anara77777@mail.ru
[1,4]Мукажанов Н.К.,
PhD, қауымдыстырылған профессор, ORCID ID: 0000-0003-4835-5751,
e-mail: nurzhan.mukazhanov@narxoz.kz
[1,2]Абдығалым Б.Х.
т.ғ.м., докторант, ORCID ID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com

[1]«Q» University, Алматы қ., Қазақстан
[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан
[3]Астана халықаралық университеті, Астана қ., Қазақстан
[4]Нархоз университеті, Алматы қ., Қазақстан

ГЕНЕТИКАЛЫҚ АУРУЛАР ТУРАЛЫ АҚПАРАТТЫ АЛУ ҮШІН АННОТАЦИЯЛАНҒАН МЕДИЦИНАЛЫҚ МӘТІНДЕР КОРПУСЫН ҚҰРУ

Аңдатпа

Мақалада экзомдық секвенирлеу нәтижелері бойынша алынған орыс тіліндегі клиникалық мәтіндерден құралған аннотацияланған корпус ұсынылады. Аталған корпус гендерге, мутацияларға, тұқым қуалайтын ауруларға, фенотиптік белгілерге және олардың клиникалық маңыздылығына қатысты атаулы нысандар мен семантикалық байланыстарды автоматты түрде анықтау міндеттерін қолдау мақсатында әзірленген. Корпусты қалыптастыру барысында нақты клиникалық экзомдық секвенирлеу есептері пайдаланылып, деректер алдын ала анонимдендіру және мәтінді нормалау кезеңдерінен өткізілді. Таңбалау үдерісінде HGVS, OMIM, ClinVar және HPO сияқты халықаралық стандарттар мен білім базалары басшылыққа алынып, биомедициналық ақпараттың бірізділігі мен дәлдігі қамтамасыз етілді. Корпуста 25 000-нан астам биомедициналық нысан және 6000-нан астам семантикалық байланыс қамтылып, бұл оны көлемі мен мазмұны жағынан клиникалық генетика саласындағы маңызды ресурсқа айналдырады. Аннотациялау бірнеше сарапшының қатысуымен қолмен жүргізіліп, нәтижелер кросс-тексеру арқылы салыстырылды, сондай-ақ аннотациялаушылар арасындағы келісім деңгейі арнайы метрикалар арқылы бағаланды. Алынған нәтижелер корпустың жоғары сапасын және сенімділігін көрсетеді. Дайын корпус медициналық генетика саласындағы табиғи тілді өңдеу модельдерін оқыту мен бағалауға, клиникалық шешім қабылдауды қолдау жүйелерін дамытуға және генетикалық деректерді құрылымдауға арналған қолданбалы зерттеулерде тиімді пайдалануға мүмкіндік береді.

Түйін сөздер: генетикалық аурулар, медициналық мәтінді өңдеу, аннотацияланған корпус, экзомдық секвенирлеу, клиникалық мәтіндер, атаулы нысандарды автоматты тану (NER), семантикалық қатынастарды алу (RE).

Кіріспе

Генетикалық аурулар – әлемдік денсаулық сақтау саласындағы маңызды мәселелердің бірі. Бүкіл әлем бойынша 300 миллионнан астам адам сирек кездесетін ауруларға шалдыққаны белгілі [1]. Orphanet дерекқорына сәйкес 6172 сирек ауру анықталған, олардың 71,9%-ының генетикалық шығу тегі бар [2]. Жалпы халықтың 3,5–5,9%-ы сирек ауруға шалдықса, бұл 263–446 миллион адамға тең болады [2]. Сирек кездесетін аурулардың көпшілігі өмірге қауіп төндіретіндей күрт (жедел) түрде дамымайды, яғни баяу өрбиді немесе симптомдары жеңіл болуы мүмкін. Дегенмен, генетикалық деректерді клиникалық тұрғыдан талдау үшін дәрігерлердің экзомдық секвенирлеу нәтижелеріне негізделген клиникалық қорытындылары басты ақпарат көзі болып табылады.

Соңғы жылдары клиникалық генетика саласы күрделі молекулалық деректерді талдаудың жаңа деңгейіне көтеріліп, экзомдық (WES) және геномдық (WGS) секвенирлеудің клиникалық практикаға енуі жеделдей түсті. Бұл технологиялар тұқым қуалайтын ауруларды дәл диагностикалауға мүмкіндік берсе де, олардың нәтижелері көбіне құрылымсыз мәтіндік есептер түрінде ұсынылады, сондықтан автоматты талдауды қиындатады [3]. Клиникалық есептердегі генетикалық нұсқалар, фенотиптік сипаттамалар, отбасылық анамнез және интерпретациялық түсіндірмелер біркелкі емес лингвистикалық құрылымдар арқылы беріледі. Соның салдарынан ақпаратты автоматты түрде құрылымдау өзекті мәселелердің біріне айналды.

Клиникалық мәтіндерден ақпарат алу мәселесіне арналған зерттеулер көрсеткендей, клиникалық есептер ғылыми мәтіндерге қарағанда анағұрлым күрделі, контекстке тәуелді синтаксиске ие және қысқартулар мен терминдердің ауыспалы нұсқалары жиі кездеседі [3]. Мұндай ерекшеліктер дәстүрлі NLP модельдерінің тиімділігін төмендетіп, клиникалық тілге бейімделген арнайы корпус пен үлгілеу әдістерінің қажеттілігін арттырады. Сол себепті, медициналық ақпараттық жүйелерде, оның ішінде клиникалық шешім қолдау құралдарында қолданылатын NLP әдістерін дамыту үшін жоғары сапалы, аннотацияланған деректердің болуы аса маңызды [4].

Генетикалық есептердің ішінде ерекше қиындық туғызатын міндеттердің бірі – гендік нұсқаларды (variants) тану және оларды HGVS стандартына сәйкестендіру. Зерттеулер көрсеткендей, генетикалық мутацияларды автоматты түрде анықтау үшін арнайы машықтанған NER модельдері қажет және мұндай модельдер тек тиісті домендік корпус болған жағдайда ғана жоғары дәлдікке қол жеткізеді [5]. Бұл клиникалық генетика мәтіндеріне арналған сапалы корпустардың жетіспеушілігін айқын көрсетеді.

Клиникалық NLP саласының 2020 жылғы жағдайын талдаған халықаралық зерттеулер де дәл осы мәселеге назар аударады: ағылшын тілінде биомедициналық мәтіндерге арналған бірнеше корпус болғанымен, нақты клиникалық жазбаларға бағытталған домендік ресурстардың тапшылығы байқалады, ал көптілді корпустар тіпті аз [6]. Орыс тілінде клиникалық есептерге арналған корпус жоқтың қасы, бұл генетикалық есептерді автоматты талдау бағытында зерттеулердің баяу дамуына әкеледі.

Соңғы жылдары трансформер негізіндегі алдын ала оқытылған модельдер (BioBERT, ClinicalBERT, PubMedBERT) биомедициналық мәтіндер үшін жоғары нәтижелер көрсетіп келеді. Дегенмен, бұл модельдер негізінен ағылшын тіліндегі ғылыми деректерде оқытылғандықтан, клиникалық жазбаларды, оның ішінде генетикалық есептерді өңдеу кезінде қосымша бейімдеуді талап етеді [7]. Осыған байланысты нақты клиникалық контентке негізделген аннотацияланған корпустардың жасалуы – көптілді медициналық NLP жүйелерін дамытудағы шешуші қадам.

Бұл жұмыс орыс тіліндегі клиникалық генетикалық есептерден ақпарат алуға арналған, халықаралық стандарттармен үйлескен аннотацияланған корпустардың болмауымен өзекті болып отыр. Сондықтан бұл зерттеу сол олқылықты толтыруды мақсат етеді және генетикалық есептерден автоматты ақпарат алу жүйелерін құруға негіз болатын жаңа корпус әзірлеуді ұсынады.

Жұмыстарға шолу

Биомедициналық табиғи тілді өңдеу (BioNLP) саласындағы соңғы жылдардағы жетістіктер ағылшын тіліндегі мәтіндерге арналған көптеген жоғары сапалы аннотацияланған корпустардың пайда болуына әкелді. Осы ресурстардың ішінде BC5CDR [8], CRAFT [9] және BioRED [10] сияқты корпустар уақыт өте келе эталондық болып қалды. BC5CDR корпусы химиялық заттар мен аурулар арасындағы қатынастарды шығаруға бағытталса [8], CRAFT корпусы онтологиялармен тураланған түсініктердің терең аннотациясын ұсынады [9]. Сонымен қатар, BioRED корпусы гендер, аурулар және химиялық заттар сияқты әртүрлі нысандар арасындағы қатынастарды қамтиды, бұл оны күрделі биомедициналық білімді шығару үшін құнды ресурсқа айналдырады [10].

Атап айтқанда, гендік нұсқалар мен мутацияларды тануға арналған зерттеулер елеулі прогреске қол жеткізді. Bangalore және әріптестері (2021) BioBERT тәрізді алдын ала оқытылған трансформерлік модельдерді қолдана отырып, еркін мәтіндерден мутацияларды анықтау мен олардың нормалану процесіндегі өзекті мәселелерді талдаған [5]. Олардың жұмысы домендік білімі бар модельдердің маңыздылығын көрсетті, бірақ сонымен бірге мұндай модельдерді оқыту үшін жеткілікті аннотацияланған деректердің болуының шектеулілігін де атап өтті [5].

Дегенмен, жоғарыда аталған корпустар негізінен PubMed және PMC сияқты көздерден алынған ғылыми әдебиеттерге негізделген. Нақты клиникалық есеп берулерден, мысалы, экзомдық секвенирлеу нәтижелері бойынша клиникалық қорытындылардан құралған корпустар әлі де сирек кездеседі. 2020 жылы Wang және оның зерттеу тобы клиникалық мәтіндерді өңдеу кезінде жиі кездесетін күрделі синтаксистік құрылымдар, көпмағыналылық және қысқартулардың молдығы сияқты факторларды қарастырып, олардың дәстүрлі NLP әдістерінің қолданылуын қиындататынын көрсетті [3]. Ал сол жылы Neveol бастаған зерттеушілер клиникалық NLP саласына арналған кең көлемді шолу жүргізіп, нақты клиникалық деректерге сүйенген көптілді ресурстарды дамытудың өзектілігін атап өткен [6].

Орыс тіліндегі биомедициналық NLP контекстінде бірнеше бастамалар пайда болды. NEREL-BIO [11] және RuUMLS-Nested [12] сияқты жобалар орыс тілдік биомедициналық мәтіндерге арналған аннотацияланған ресурстарды құруға бағытталған. Алайда, бұл ресурстар негізінен ғылыми абстракттерге негізделген және клиникалық генетикалық есептердің нақты лингвистикалық және семантикалық ерекшеліктерін, мысалы, Human Genome Variation Society (HGVS) номенклатурасы бойынша күрделі гендік нұсқалардың көріністерін толық қамтымайды [13].

Клиникалық генетикалық контекстте ақпаратты шығаруға арналған корпустардың болмауы ерекше проблемалар тудырады. Cohen және оның зерттеу тобы (2020) ғылыми әдебиеттерден гендік варианттар жөніндегі мәліметтерді алуда туындайтын қиындықтарды талдағанымен, клиникалық есептерді өңдеу аспектісіне арнайы тоқталмаған [4]. Клиникалық есептер жиі кірістірілген нысандарды (мысалы, ген ішіндегі HGVS белгілеуі) және клиникалық тұрғыдан маңызды қатынастарды (мысалы, нұсқаның патогенділігі) қамтиды, оларды дұрыс аннотациялау үшін арнайы схемаларды қажет етеді.

Трансформер архитектурасына негізделген алдын ала оқытылған тіл модельдерінің пайда болуы (BioBERT [14], PubMedBERT [15]) биомедициналық NLP саласында тұтас революция жасады. Yoon және әріптестері (2020) бұл модельдердің биомедициналық мәтіндерді өңдеудегі тиімділігін дәлелдегенімен, олардың нақты клиникалық домендерге, әсіресе ағылшын тілінен өзге тілдердегі салаларға бейімделуін қамтамасыз ету үшін қосымша зерттеулердің қажет екенін атап өткен [7]. Осыған байланысты, нақты клиникалық деректер негізінде оқытылған және көптілді контекстке бейімделген модельдерді дамыту өзекті мәселе болып қала береді.

Халықаралық BioNLP сериктестік жобалары мен n2c2 сияқты клиникалық NLP сынақтары клиникалық есептерден ақпарат шығару мәселелерін белсенді түрде зерттеді [3, 6]. Дегенмен, бұл жобалардың басым көпшілігі ағылшын тіліндегі деректерге аударған. Осылайша, орыс тілдік клиникалық генетикалық есептерге арналған жоғары сапалы, қолмен аннотацияланған корпус әлі де жоқ, бұл осы маңызды доменде зерттеулер мен технологиялық дамуды шектейді.

Биомедициналық NLP саласындағы негізгі корпустарды салыстыру олардың күшті және әлсіз жақтарын анық көрсетеді. Төмендегі 1-кестеде корпустардың салыстырмалы талдауы ұсынылған.

Кесте 1 – Биомедициналық NLP бойынша негізгі аннотацияланған корпустарды салыстыру

Корпус атауы	Тілі	Дереккөзі	Негізгі нысандар	Ерекшеліктері	Шектеулері
BC5CDR [8]	Ағылшын	Ғылыми мақалалар	Химия, ауру	Химия-ауру қатынастары	Тек ғылыми әдебиет
CRAFT [9]	Ағылшын	Ғылыми мақалалар	Ген, ауру, белок	Онтологиялармен терең туралау	Түпнұсқа мәтін қиын, клиникалық емес
BioRED [10]	Ағылшын	Ғылыми мақалалар	Ген, ауру, химия	Бірнеше нысан түрлері арасындағы қатынастар	Тек ғылыми әдебиет, HGVS жоқ
NCBI Disease [16]	Ағылшын	Ғылыми мақалалар	Ауру	MeSH-ке нормалау	Тек ауру атауларына бағытталған
NEREL-BIO [11]	Орыс/ Ағылшын	PubMed абстракттері	Ген, Ауру, Белок	Көптілділік, UMLS-ке сілтемелер	Клиникалық емес, HGVS қамтылмаған
RuUMLS-Nested [12]	Орыс	Биомедициналық мәтіндер	Ген, Ауру, Дәрі	Ішкі (nested) нысандар, UMLS	Негізінен ғылыми мәтіндер

Кестеден көрініп тұрғандай, барлық негізгі корпустар ағылшын тіліндегі ғылыми әдебиеттерге негізделген. Орыс тіліндегі соңғы жобалар (NEREL-BIO, RuUMLS-Nested) маңызды болғанымен, олар да ғылыми контентке шектеледі және клиникалық генетиканың нақты қажеттіліктерін, мысалы, HGVS номенклатурасын немесе нұсқаның патогенділігі сияқты клиникалық интерпретациялық нысандарды толық қамтымайды.

Әзірленетін GENEXOM корпусы осы олқылықтарды ескере отырып, келесі қағидалар бойынша құрылатын болады:

- Дереккөз: нақты клиникалық экзомдық секвенирлеу есептері пайдаланылады;
- Тіл: орыс тіліне негізделеді;
- Аннотацияның тереңдігі: халықаралық стандарттармен (HGVS, OMIM, ClinVar, HPO) сәйкестендірілген көпқабатты нысандар мен семантикалық қатынастар қарастырылады;
- Домендік маңыздылығы: генетикалық диагностикаға тікелей қатысты нысандарды қамту жоспарланған.

Осылайша, қазіргі заманғы BioNLP ландшафты құрылымдық тұрғыдан бай болса да, ол біздің зерттеуіміздің негізгі мақсаты болып табылатын нақты клиникалық генетикалық есептерге арналған көптілді, доменге арналған ресурстар бойынша айтарлықтай олқылықты сақтайды. GENEXOM корпусы осы олқылықты толтыруға, орыс тіліндегі клиникалық генетика үшін бірегей, халықаралық стандарттармен туралanған және көпқабатты аннотацияланған ресурсты ұсынуға бағытталған.

Материалдар мен әдістер

1) Дереккөздер және алдын ала өңдеу

GENEXOM корпусының негізін 2018–2023 жж. аралығында жиналған 200 нақты экзомдық секвенирлеу (Whole Exome Sequencing, WES) бойынша клиникалық есептер құрайды.

Деректер жинағы Қазақстанның түрлі медициналық генетикалық орталықтарынан анонимді түрде алынды. Корпустың клиникалық сәйкестігі мен деректердің алуан түрлілігін қамтамасыз ету үшін есептерді іріктеу келесі критерийлер бойынша жүргізілді:

- Экзомдық секвенирлеу нәтижелері бойынша толық клиникалық қорытындының болуы;
- Гендік нұсқалардың HGVS стандарты бойынша сипаттамасы;
- Фенотипік сипаттамалардың егжей-тегжейлі сипаттамасы;
- Аурудың OMIM идентификаторымен немесе атауымен анықталуы;
- Нұсқаның ClinVar дерекқорына сілтемесінің болуы.

Деректерді жинау және алдын ала өңдеу конвейері төмендегі кезеңдерден тұрды (1-сурет):

Сурет 1 – GENEXOM корпусын құрудың деректерді өңдеу конвейері

1. Деректерді жинау және анонимдеу. Бастапқы кезеңде клиникалық-диагностикалық мекемелерден алынған медициналық есептердің PDF және DOCX форматтары жинақталып, қауіпсіздік және конфиденциалдылық талаптарына сәйкес, жеке басын анықтауға болатын барлық ақпараттар (пациенттің аты-жөні, туған күні, жеке куәлік нөмірі, мекен-жайы) деректерді өңдеудің бастапқы кезеңінде жойылды. Анонимдеу процедурасы автоматты түрде жүргізіліп, артын қалдырмай тексерілді.

2. Мәтінді парсингтеу және шығару. Клиникалық есептердің құрылымсыз мәтіндерін өңдеу үшін PDFPlumber [17] және python-docx [18] кітапханалары қолданылды. Бұл кітапханалар медициналық құжаттардың күрделі форматтарымен (кестелер, баспа беттері, суреттер) тиімді жұмыс істей алады. Әсіресе, PDFPlumber кітапханасы кестелерді және күрделі құрылымды анықтауда жоғары дәлдік көрсеткен [17]. Парсингтеу нәтижесінде алынған мәтіндер біріктіріліп, бір файл ретінде сақталды.

3. Мәтінді нормалау және бірыңғайландыру. Клиникалық генетикалық есептердегі терминологиялық әртүрлілікті жою үшін келесі нормалау операциялары жүргізілді:

- Медициналық аббревиатураларды анықтау және оларды толық атауларымен ауыстыру [8]. Мысалы, «AD» – «autosomal dominant» (аутосомдно-доминантты), «AR» – «autosomal recessive» (аутосомдно-рецессивті), «VUS» – «variant of uncertain significance» (белгісіз маңыздылығы бар нұсқа).
- HGVS номенклатурасына сәйкес гендік нұсқаларды тану және нормалау [13]. Мысалы, «c.5266dupC», «p.Glu6Val» т.б. Бұл процесте HGVS стандартына сәйкес емес жазулар (мысалы, «дупликация цитозина») формальды белгілерге аударылды.

◆ Терминдерді стандарттау және орфографиялық қателерді түзету. Мысалы, «brca1» – «BRCA1», «хромосома 13» – «хромосома 13». Ген атауларының дұрыс жазылуын тексеру үшін HGNC [19] дерекқоры пайдаланылды.

4. Деректердің сапасын тексеру. Өңделген мәтіндердің дәлдігін тексеру үшін арнайы әзірленген Python скрипттері қолданылды. Скрипттер төмендегідей тексерулерді жүргізді:

◆ HGVS форматына сәйкестікті тексеру;

◆ Ген атауларының HGNC стандартына сәйкестігін растау;

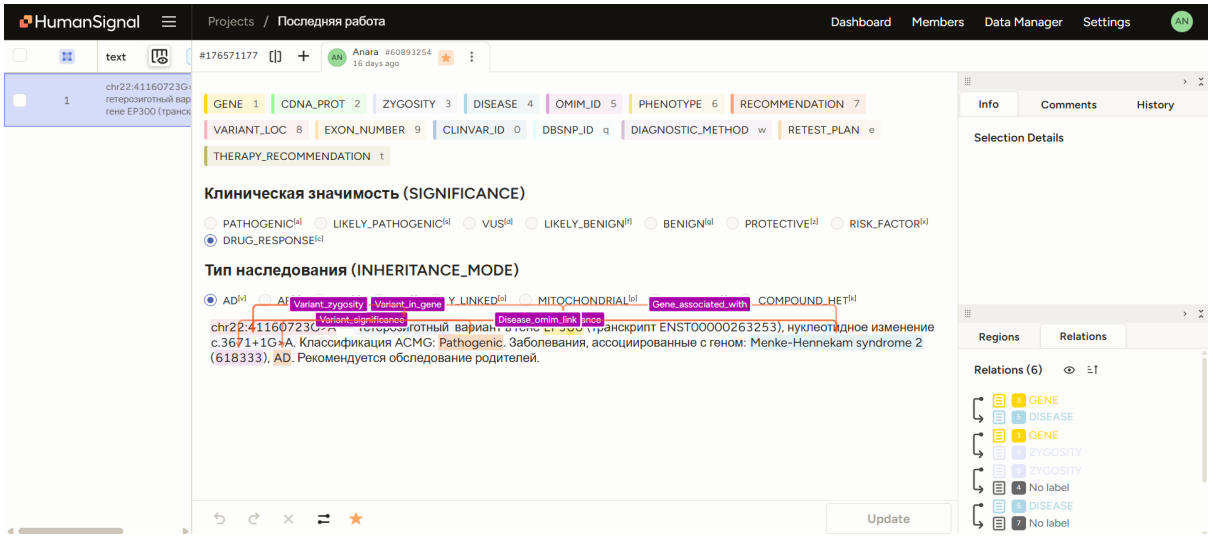
◆ OMIM және ClinVar идентификаторларының форматын тексеру.

5. Аннотациялауға дайындық. Алдын ала өңделген мәтіндер аннотациялау процесіне дайындалды. Мәтіндер Label Studio платформасына жүктеліп, аннотациялау схемасына сәйкес бапталды [20]. Әрбір құжатқа арнайы идентификатор тағайындалып, аннотациялау тарихын бақылау жүйесі орнатылды.

Деректерді өңдеудің әрбір кезеңінде стандартты процедуралар қолданылып, нәтижелердің тұрақтылығы қамтамасыз етілді. Бұл тәсіл корпус құру процесінің қайталанатындығын және өңделген деректердің жоғары сапасын қамтамасыз етті.

2) Аннотациялау схемасы және процесі

Аннотациялау процесі корпус құрудың негізгі кезеңі болып табылады. Аннотациялау үшін HumanSignal компаниясының Label Studio ашық бастапқы кодты платформасы таңдалды [20]. Бұл таңдау платформаның келесі артықшылықтарына негізделді: ішкі (nested) және қиылысатын (overlapping) нысандарды қолдау, семантикалық қатынастарды белгілеу мүмкіндігі, JSON сияқты машиналық өңдеуге ыңғайлы форматтарда деректерді экспорттау, сондай-ақ бірнеше аннотаторлармен бірлескен жұмыс істеуге арналған интуитивті интерфейс. Платформаның аннотациялау интерфейсінің көрінісі 2-суретте көрсетілген.



Сурет 2 – Label Studio (HumanSignal) платформасындағы аннотациялау интерфейсінің көрінісі

2-суретте көрсетілгендей, аннотациялау интерфейсі үш негізгі аймақтан тұрды:

◆ Орталық панель: Аннотация процесі үшін бастапқы мәтін орналасты. Аннотаторлар тікелей осы аймақта мәтін бөліктерін тұтас ретінде бөліп (highlight), оларға сәйкес тегтерді тағайындады;

◆ Тегтер панелі: Алдын ала анықталған нысан түрлерінің (GENE, CDNA_PROT, т.б.) тізімі орналасты. Аннотатор бөліп алған мәтін аралығына бір немесе бірнеше тегтерді тағайындай алатын;

◆ Басқару панелі: Тапсырмалар арасында навигациялау, аяқталған аннотацияларды сақтау және қателерді түзету үшін басқару элементтері берілген.

Интерфейстің ерекшелігі – іріктеу-тағайындау әдісінің қарапайымдылығы және бір мәтін ішінде әртүрлі түстік кодтау арқылы бірнеше нысан түрлерін бір мезгілде көру мүмкіндігі. Мысалы, 2-суретте «EP300» тіркесі GENE ретінде сары түспен, ал «Pathogenic» сөзі SIGNIFICANCE ретінде басқа түспен белгіленген. Сонымен қатар, платформа ішкі құрылымдарды (мысалы, «BRCA1 c.5266dupC» тіркесінде «BRCA1» GENE, ал «c.5266dupC» CDNA_PROT ретінде) дәл шекаралармен белгілеуге мүмкіндік берді.

Аннотациялау схемасы халықаралық түрде танылған биомедициналық стандарттар мен білім базалары негізінде әзірленді. Негізгі нысандар (entity) түрлері және олардың сәйкес келетін стандарттары 2-кестеде ұсынылған.

Кесте 2 – Аннотацияланған негізгі нысандар және олардың сипаттамалары

Нысан атауы	Сипаттама	Қолданылған стандарт
GENE	Геннің атауы (HGNC сәйкестендірілген)	HGNC [19]
CDNA_PROT	Мутацияның HGVS-форматтағы түрленуі	HGVS [13]
ZYGOSITY	Зиготалық күйі (гетерозиготалық немесе гомозиготалық)	Клиникалық генетика [21]
DISEASE	Ауру немесе синдром	OMIM [22]
OMIM_ID	OMIM дерекқорындағы идентификатор	OMIM [22]
PHENOTYPE	Науқас белгілері мен симптомдары	HPO [23]
RECOMMENDATION	Клиникалық ұсыныстар немесе тұжырымдар	Клиникалық нұсқаулар [21]
VARIANT_LOC	Варианттың геномдық/хромосомалық орны	Genomic координаттар [13]
EXON_NUMBER	Экзон нөмірі	Транскрипт аннотациясы [13]
CLINVAR_ID	ClinVar вариант идентификаторы	ClinVar [24]
DBSNP_ID	dbSNP дерекқорындағы вариант идентификаторы	dbSNP [25]
DIAGNOSTIC_METHOD	Қолданылған диагностикалық әдіс	Клиникалық диагностика [21]
RETEST_PLAN	Қайта тексеру жоспары	Клиникалық бақылау [21]
THERAPY_RECOMMENDATION	Емдеу бойынша ұсыныстар	Терапия нұсқаулары [21]

Label Studio платформасының икемділігі күрделі аннотациялау схемаларын жүзеге асыруға мүмкіндік берді [20]. Платформа арнайы медициналық мәтіндерді өңдеуге бейімделген және көпдеңгейлі аннотациялауды қолдайды.

Аннотациялау барысында клиникалық маңыздылық (SIGNIFICANCE) ACMG/AMP нұсқаулығының жаңартылған нұсқасына сәйкес белгіленді [21]. Негізгі категориялар 3-кестеде келтірілген.

Кесте 3 – Клиникалық маңыздылық категориялары

Категория	Ағылшынша атауы	Сипаттама
Патогенді	Pathogenic	Дәлелденген патогенді нұсқалар
Ықтимал патогенді	Likely_pathogenic	Ықтималды патогенді нұсқалар
Белгісіз маңыздылығы	VUS (Variant of Uncertain Significance)	Клиникалық маңыздылығы белгісіз нұсқалар
Ықтимал зақымсыз	Likely_benign	Ықтималды зақымсыз нұсқалар
Зақымсыз	Benign	Дәлелденген зақымсыз нұсқалар
Қорғанышты	Protective	Қорғанышты мутация
Тәуекел факторы	Risk_factor	Тәуекел факторы
Дәрілік реакция	Drug_response	Дәрілік реакция (фармакогенетика)

Тұқым қуалау сұлбасы (INHERITANCE_MODE) генетикалық аурулардың заманауи классификациясы бойынша анықталды [21]. INHERITANCE_MODE нысандары 4-кестеде келтірілген:

Кесте 4 – Тұқым қуалау сұлбалары

Мән	Ағылшынша атауы	Сипаттамасы
AD	Autosomal dominant	Аутосомдно-доминантты
AR	Autosomal recessive	Аутосомдно-рецессивті
XLD	X-linked dominant	X-хромосомаға байланысты доминантты
XLR	X-linked recessive	X-хромосомаға байланысты рецессивті
Y_LINKED	Y-linked	Y-хромосомаға байланысты
MITOCHONDRIAL	Mitochondrial	Митохондриалды
DE_NOVO	De novo mutation	Де novo мутация
COMPOUND_HET	Compound heterozygote	Композитті гетерозиготалы

Аннотациялау барысында нақты клиникалық есептердің мысалдары қолданылды. 2-суретте көрсетілген «chr22-41160723G>A» жазбасында келесі ақпараттар аннотацияланған:

- HGVS форматындағы нұсқа: c.3671+1G>A;
- ACMG классификациясы: Pathogenic [21];
- Байланысты ауру: Menke-Hennekam syndrome 2 (OMIM: 618333) [22];
- Тұқым қуалау сұлбасы: AD (аутосомдно-доминантты);
- Клиникалық ұсыныс: ата-ананы тексеру ұсынылған.

Label Studio платформасы нысандар арасындағы семантикалық қатынастарды белгілеуге мүмкіндік берді [20]. Бұл функционал ген-ауру, ген-фенотип және нұсқа-клиникалық маңыздылық сияқты маңызды байланыстарды анықтауға мүмкіндік берді. Label Studio бағдарламасы бұл үшін арнайы құралдарды ұсынды: аннотатор бірінші нысанды таңдап, содан кейін екінші нысанды таңдап, оларды байланыстыратын қатынастың түрін тізімнен таңдады. Бұл корпусты тек NER үшін ғана емес, сонымен қатар білімді шығару (relation extraction) және клиникалық білім графтарын құру үшін де құнды етеді. Негізгі қатынас түрлері 5-кестеде келтірілген.

Аннотациялау дәлдігін арттыру үшін келесі заманауи әдістер қолданылды:

- Автоматты түрде терминдерді стандартты онтологиялармен тексеру;
- Аннотациялар арасындағы қайшылықтарды анықтау;
- Күрделі жағдайларды шешу үшін эксперттердің келісім сессиялары;
- Клиникалық маңыздылықты белгілеуде ClinVar деректерінің көмегімен растау [24]

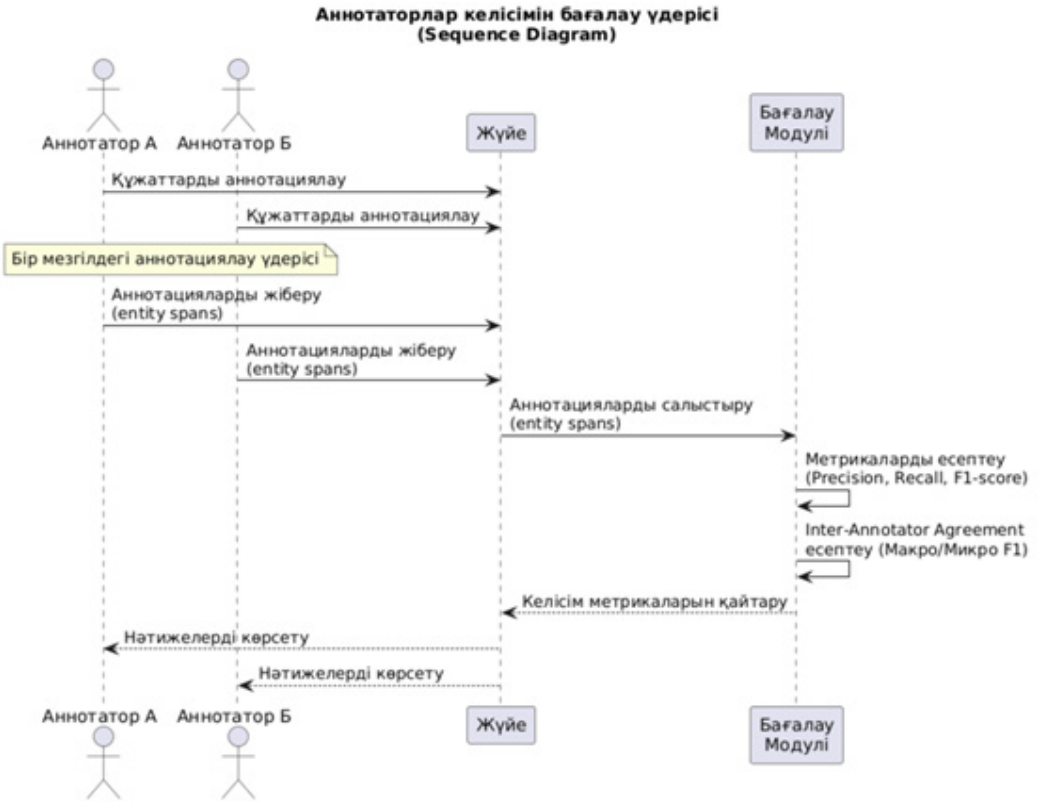
Кесте 5 – Семантикалық қатынастар

Қатынас түрі	Мақсаты	Мысал
variant_in_gene	Мутацияның белгілі бір генге тиесілілігі	c.3671+1G>A – EP300
gene_associated_with	Гендің аурумен байланысы	EP300 – Menke-Hennekam syndrome 2
variant_significance	Нұсқаның патогенділігі	c.3671+1G>A – Pathogenic
disease_inheritance	Аурудың тұқым қуалау түрі	Menke-Hennekam syndrome 2 – AD
phenotype_supports_disease	Фенотиптің ауруды растауы	syndactyly – Menke-Hennekam syndrome
disease_omim_link	Аурудың OMIM идентификаторымен байланысы	Menke-Hennekam syndrome – 618333
variant_zygosity	Нұсқаның зиготальность түрі	c.3671+1G>A – heterozygous

Бұл кешенді тәсіл клиникалық генетикалық мәтіндердің барлық аспектілерін қамтуға мүмкіндік берді және корпустың жоғары сапасын қамтамасыз етті. Аннотацияланған деректер кейінгі NLP модельдерін оқыту үшін дайындалды.

3) Деректер сапасын қамтамасыз ету және кросс-аннотациялау

Аннотацияланған корпустың сапасын және сенімділігін қамтамасыз ету үшін көпсатылы кросс-аннотациялау процедурасы жүргізілді. Әрбір қатысушының рөлі және олардың өзара байланысы 3-суреттегі реттік диаграммасында (Sequence diagram) көрсетілген. Бұл процессте үш тәуелсіз сарапшы-аннотатор қатысты, олардың әрқайсысы биомедициналық лингвистика немесе клиникалық генетика саласында тәжірибеге ие болды.



Сурет 3 – Корпусты аннотациялау процесінің реттік диаграммасы (Inter-Annotator Agreement)

3-суретте көрсетілгендей, процесс келесі негізгі кезеңдерден тұрды:

◆ Тәуелсіз аннотациялау (Annotator A, Annotator B): Әр аннотатор бірдей мәтіндер жиынтығын Label Studio платформасында [20] тәуелсіз түрде белгіледі;

◆ Аннотацияларды салыстыру (Compare annotations): Екі аннотатордың нысан аралықтары (entity spans) және олардың категориялары салыстырылды;

◆ Метрикаларды есептеу (Compute Precision, Recall, F1-score): Салыстыру нәтижелері бойынша дәлдік (Precision), өтіл (Recall) және F1-өлшемі сияқты стандартты метрикалар есептелді [26];

◆ Келісім деңгейін бағалау (Inter-Annotator Agreement): Соңғы кезеңде әрбір нысан категориясы бойынша макро және микро-орталама F1-өлшемі ретінде аннотаторлар арасындағы жалпы келісім деңгейі бағаланды.

Кросс-аннотациялау процесі екі негізгі кезеңнен тұрды:

Бірінші кезең – бастапқы аннотациялау. Бұл кезеңде әр аннотатор бірдей мәтіндер жиынтығын тәуелсіз түрде аннотациялады. Содан соң аннотациялар Label Studio платформасында жүргізілді [20]. Бастапқы аннотациялардың сәйкестігін тексеру үшін F1-өлшемі қолданылды [26]

Екінші кезең – келісім сессиялары және түзетулер. Мұнда айтарлықтай айырмашылықтар кездесетін аннотациялар бірлескен талқылауға ұсынылды. Күрделі клиникалық жағдайлар үшін клиникалық генетик маманының кеңесі пайдаланылды. Кейін аннотациялық нұсқаулық қайта қаралып, нақтыланды.

Келісім деңгейін бағалау үшін NER жүйелерінің бағалауында кеңінен қолданылатын F1-өлшемі таңдалды [26]. Бұл таңдау дәл шекаралар мен нысан категорияларын есепке алатын ең объективті көрсеткіш болғандықтан жасалды.

Бастапқы кросс-аннотациялау нәтижелері келесідей болды:

◆ Жалпы F1-өлшемі: 62.6%;

◆ Ең төмен келісім SIGNIFICANCE категориясы үшін: 48%;

◆ Ең жоғары келісім GENE категориясы үшін: 78%.

Аннотациялық нұсқаулықты екі рет қайта қарап, нақтылағаннан кейін және келісім сессияларын өткеннен кейін, соңғы келісім көрсеткіштері айтарлықтай жақсарды және олар 6-кестеде көрсетілген.

Кесте 6 – Аннотаторлар арасындағы соңғы келісім көрсеткіштері

Категория	F1-өлшемі (%)	Түсіндірме
GENE	92	Жоғары стандарттау дәрежесі
CDNA_PROT	87	HGVS номенклатурасының нақтылығы
DISEASE	85	OMIM сәйкестендіруінің арқасы
SIGNIFICANCE	76	Контекстке тәуелділік
ZYGOSITY	74	Терминологиялық әртүрлілік
Орташа	83.5	Жоғары сенімділік деңгейі

Келісім деңгейіне әсер еткен негізгі факторлар:

◆ Лексикалық стандарттау: GENE және CDNA_PROT сияқты формальды нотациялар жоғары келісімге әкелді;

◆ Контекстке тәуелділік: SIGNIFICANCE және ZYGOSITY сияқты категориялар төменірек келісімге ие болды, себебі олар клиникалық контекстке байланысты түрленуі мүмкін;

◆ Күрделі лингвистикалық құрылымдар: анафоралар және ішкі құрылымдар аннотацияны қиындатып, келісімді төмендетті.

Сапаны қамтамасыз етудің қосымша шаралары:

◆ Автоматты түрде UMLS [27] және HPO [23] онтологияларымен сәйкестікті тексеру;

- Аннотациялық қайшылықтарды анықтауға арналған арнайы скрипттерді әзірлеу;
- Мерзімді түрде клиникалық генетиктердің эксперттік бағалауын жүргізу.

Нәтижесінде, 83.5% орташа F1-өлшемімен GENEXOM корпусы биомедициналық корпустар үшін қолданылатын халықаралық стандарттарға сәйкес келетін жоғары сапалы аннотацияланған ресурс болып табылады [26, 28].

Нәтижелер мен талқылау

Бұл бөлімде GENEXOM корпусының негізгі сипаттамалары мен статистикалық көрсеткіштері, сондай-ақ корпустың биомедициналық табиғи тілді өңдеу (nlp) зерттеулеріндегі қолдану мүмкіндіктері талқыланады.

1) Корпус статистикасы

GENEXOM корпусы 200 клиникалық экзомдық секвенирлеу есебін қамтиды. Корпус оқыту (70%), тексеру (10%) және сынақ (20%) жинақтарына бөлінген. Жиынтық статистикалық көрсеткіштер 7-кестеде келтірілген.

Кесте 7 – Корпустағы құжаттар, сөйлемдер және токендер саны

Категория	Оқыту жиыны	Тексеру жиыны	Сынақ жиыны	Барлығы
Құжаттар	140	30	30	200
Сөйлемдер	8574	1836	1840	12250
Токендер	111340	22840	22680	156860

Корпуста барлығы 25096 аннотацияланған нысан және 6428 семантикалық байланыс белгіленген. Нысандардың таралуы 8-кестеде көрсетілген.

Кесте 8 – Корпустағы нысандар мен байланыстар саны

Категория	Оқыту жиыны	Тексеру жиыны	Сынақ жиыны	Барлығы
GENE	870	125	250	1245
CDNA_MUTATION	685	100	197	982
DISEASE	810	123	220	1153
SIGNIFICANCE	925	130	262	1317
ZYGOSITY	582	83	184	849
OMIM_ID	770	150	169	1089

Ең жиі кездесетін нысандардың үлесі:

- GENE – 20.2%;
- CDNA_PROT – 16.8% ;
- DISEASE – 14.8%;
- SIGNIFICANCE – 10.7%;
- OMIM_ID – 7.6%.

Кросс-аннотациялау нәтижелері бойынша аннотаторлар арасындағы келісім (IAA) орташа есеппен 83.5% F1-өлшемін құрады. Бұл көрсеткіш биомедициналық корпустар үшін халықаралық стандарттарға сәйкес келеді [26,28].

Ең жоғары келісім көрсеткіштері:

- GENE – 92% (HGNC стандартының арқасында);
- CDNA_PROT – 87% (HGVS номенклатурасының нақтылығы);
- DISEASE – 85% (OMIM сәйкестендіруі).

Төменгі келісім көрсеткіштері:

- SIGNIFICANCE – 76% (контекстке тәуелділік);
- ZYGOSITY – 74% (терминологиялық әртүрлілік).

GENEXOM корпусының практикалық құндылығын көрсету үшін базалық NER және RE модельдері оқытылды. Нәтижелер 9-кестеде көрсетілген.

Кесте 9 – NER жүйелерінің салыстырмалы нәтижелері

Модель	GENE	DISEASE	CDNA_PROT	SIGNIFICANCE	Орташа F1
CRF	0.79	0.75	0.77	0.71	0.77
RuBERT	0.90	0.88	0.89	0.83	0.88

Қатынастарды анықтау тапсырмасында BERT негізіндегі модель ережелерге негізделген жүйеден айтарлықтай жоғары нәтиже көрсетті (0.74 F1 және 0.61 F1).

GENEXOM корпусы орыс тілдік клиникалық генетикалық мәтіндер үшін бірегей ресурс болып табылады. Біздің зерттеу нәтижелері мынаны көрсетеді:

1. Тілдік ерекшеліктер: орыс тілдік клиникалық есептер ағылшын тіліне қарағанда терминологиялық әртүрлілігі жоғары екенін атап өту керек, бұл NER жүйелерінің жұмыс істеуін қиындатады.
2. Домендік маңыздылық: корпустағы SIGNIFICANCE және ZYGOSITY сияқты арнайы категориялар клиникалық шешім қабылдау үшін өте маңызды болып табылады.
3. Салыстырмалы артықшылықтар: GENEXOM барлық белгілі биомедициналық корпустармен салыстырғанда клиникалық генетикалық есептерді егжей-тегжейлі қамтитын жалғыз ресурс болып табылады.

Корпустың шектеулері:

- Тек экзомдық секвенирлеу деректерін қамтиды;
- Тек орыс тіліндегі мәтіндер;
- Клиникалық есептердің форматтық әртүрлілігі.

Болашақ зерттеулер:

- Корпусты геномдық секвенирлеу есептерімен кеңейту;
- Көптілді модельдерді әзірлеу;
- Клиникалық шешім қолдау жүйелерін енгізу.

Қорытынды

Бұл зерттеуде орыс тіліндегі клиникалық экзомдық секвенирлеу есептері негізінде GENEXOM корпусы жасалды. Бұл корпус клиникалық генетика саласындағы табиғи тілді өңдеуге арналған алғашқы көпқабатты аннотацияланған ресурс болып табылады. Жұмыстың ғылыми және практикалық маңыздылығы оның клиникалық генетика мен жасанды интеллект саласының өзара интеграциясына қосатын үлесінде жатыр.

Зерттеу барысында келесі негізгі нәтижелерге қол жеткізілді.

Корпус құру сапасы:

- 200 клиникалық экзомдық секвенирлеу есебі негізінде жоғары сапалы корпус құрылды
- 25 000-нан астам нысан және 6 400-ден астам семантикалық байланыс дәл белгіленді
- Аннотаторлар арасындағы келісім 83.5% F1-өлшемімен жоғары сенімділік деңгейін көрсетті

Әдіснамалық тұрғыда, зерттеу барысында халықаралық стандарттар – HGVS, OMIM, ClinVar, HPO және ACMG/AMP – негізінде арнайы аннотациялау схемасы әзірленді. Сонымен қатар, Label Studio платформасының мүмкіндіктері тиімді пайдаланылып, көпқабатты аннотациялау процесі іске асырылды. Нәтижесінде, орыс тіліндегі клиникалық мәтіндерге арналған алғашқы арнайы аннотацияланған корпус жасалып, ол биомедициналық

NLP зерттеулері үшін сенімді және сапалы дереккөзге айналды. Практикалық жағынан, RuBERT моделі NER тапсырмасында 88% F1 көрсеткішіне жетіп, қатынастарды анықтау тапсырмасында BERT негізіндегі модель 0.74 F1 деңгейінде жоғары нәтиже көрсетті. Бұл көрсеткіштер GENEXOM корпусын клиникалық мәтіндерден ақпарат алу және генетикалық есептерді автоматты талдау үшін сенімді негіз ретінде пайдалануға мүмкіндік береді.

GENEXOM корпусының дамуы ғылыми және практикалық тұрғыдан бірқатар маңызды үлестерді қосады. Ғылыми тұрғыда, ол орыс тіліндегі биомедициналық NLP зерттеулеріне арналған бірегей дереккөз болып табылады, бұл клиникалық генетикалық мәтіндерді өңдеудің жаңа әдістемелерін әзірлеуге мүмкіндік береді. Сонымен қатар, корпус көптілді биомедициналық NLP зерттеулерінің дамуына да негіз бола алады, өйткені орыс тіліндегі клиникалық контентке бейімделген деректер базасы халықаралық зерттеулер үшін қажетті деректерді қамтамасыз етеді.

Практикалық жағынан, GENEXOM корпусы клиникалық шешім қолдау жүйелерін әзірлеу және генетикалық есептерді автоматты түрде өңдеу мен талдау үшін тікелей қолдануға жарамды. Бұл дәрігерлердің клиникалық шешім қабылдау процестерін жеңілдетіп, диагностикалық дәлдікті арттыруға ықпал етеді. Клиникалық генетика саласына әсері де маңызды: сирек кездесетін генетикалық ауруларды диагностикалауды жеделдету, генетикалық нұсқаларды интерпретациялауды автоматтандыру және клиникалық генетикалық деректерді стандарттау арқылы медициналық тәжірибені жетілдіру мүмкіндігін береді.

Зерттеу нәтижелері GENEXOM корпусын әрі қарай дамыту үшін бірнеше перспективалық бағыттарды айқындайды. Біріншіден, корпусты кеңейту бойынша жұмыстар жүргізу жоспарланып отыр: геномдық секвенирлеу деректерін қосу, басқа тілдердегі клиникалық есептерді қамту және фенотиптік ақпаратты толықтыру. Екіншіден, технологиялық жетілдіру мәселелері маңызды болып отыр, бұл бағытта көптілді NLP модельдерін әзірлеу, клиникалық шешім қолдау жүйелерін енгізу және нақты уақыт режимінде талдау жүргізу мүмкіндіктерін құру көзделуде. Үшіншіден, интеграциялық жобаларға мән беріледі: электрондық денсаулық карталар жүйелерімен біріктіру, халықаралық генетикалық дереккорлармен интеграция және клиникалық тәжірибеге тікелей енгізу, бұл корпустың практикалық құндылығын арттырады.

GENEXOM корпусының дамуы қоғамдық денсаулық сақтау саласына бірнеше маңызды әсерлерін тигізеді. Диагностикалық сапаны арттыру арқылы генетикалық ауруларды уақытында анықтауға, диагностикалық қателерді азайтуға және дәл диагностика негізінде жекелендірілген медицинаны дамытуға мүмкіндік береді. Экономикалық тұрғыда, корпустың қолданысы генетикалық тестілеу шығындарын азайтып, дәрігерлердің уақытын үнемдеуге және клиникалық зерттеулерді жеделдетуге септігін тигізеді. Сонымен қатар, GENEXOM корпусы білім беру және кадрлар даярлау саласында да маңызды рөл атқарады: генетик мамандарды оқытуға, клиникалық генетика бойынша оқу ресурстарын жасауға және жас мамандарды кәсіби тұрғыдан даярлауда қолдануға мүмкіндік береді. Осылайша, GENEXOM корпусы ғылыми зерттеулер, клиникалық практика және қоғамдық денсаулық сақтау саласында кешенді ықпал ету әлеуетіне ие.

GENEXOM корпусы орыс тілдік клиникалық генетика саласындағы табиғи тілді өңдеу (nlp) зерттеулеріне негіз болып, жекелендірілген медицинаның дамуына үлес қосады. Корпустың ашық қол жетімділігі (https://github.com/Anara-Sultangaziyeva/GENEXOM) оны кеңінен қолдануға мүмкіндік береді.

Жұмыстың негізгі үлесі:

- Орыс тілдік клиникалық генетика үшін алғашқы арнайы корпус;
- Халықаралық стандарттарға сәйкес келетін жоғары сапалы аннотация;
- Практикалық қолданылымдылығы дәлелденген шешім;
- Болашақ зерттеулерге кен перспективалар ашатын ресурс.

GENEXOM корпусы тек ғылыми зерттеулер үшін ғана емес, сонымен қатар клиникалық тәжірибеде де нақты пайда әкелетін күрделі және көпқырлы ресурс болып табылады. Оның

дамуы клиникалық генетика мен жасанды интеллект саласының интеграциясында маңызды қадам болып саналады.

Келесі кезеңде корпусты кеңейту және оны клиникалық тәжірибеге енгізу болады. Бұл жолда генетикалық аурулармен күресте жаңа мүмкіндіктер ашылуы мүмкін, әсіресе сирек кездесетін ауруларды диагностикалау және емдеу саласында.

Қаржыландыру туралы ақпарат. «Бұл зерттеуді Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландырады (Грант № AP26105113)».

ӘДЕБИЕТТЕР

1 Rare Diseases International. Rare Diseases International (online). URL: https://rarediseasesinternational.org (accessed: 13.11.2025).

2 Nguengang Wakap, S., Lambert, D.M., Olry, A., et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics, 28, 165–173 (2020).

3 Wang, Y., Wang, L., Rastegar-Mojarad, M., et al. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 103, 103466 (2020).

4 Cohen, T., Whitfield, G., and Xu, H. Extracting genetic variant information from the literature. Database (Oxford), 2020, baz158 (2020).

5 Bangalore, S., Hegde, H., Bharadwaj, A., et al. Extraction and normalization of mutations in free text using biomedical named entity recognition models. Bioinformatics, 37 (7), 949–957 (2021).

6 Névéol, A., Zweigenbaum, P., and Roberts, A. Clinical natural language processing in 2020: State of the art and future challenges. Journal of Biomedical Informatics, 110, 103604 (2020).

7 Yoon, W., Lee, J., Kim, S., et al. Pre-trained language models for biomedical natural language processing. Briefings in Bioinformatics, 22 (6), 1686–1702 (2020).

8 Li, M., Lu, Y., Shen, H., et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database, 2016, baw068 (2016).

9 Bada, M., Eckert, M., Evans, D., et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics, 13 (1), 161 (2012).

10 Zhao, W., Wu, X., Ma, Y., et al. BioRED: A biomedical relation extraction dataset for complex relation types. Bioinformatics, 38 (8), 2210–2216 (2022).

11 Gurulingappa, H., Mateen, A., and Dima, A. NEREL-BIO: A new resource for biomedical named entity recognition and classification in Russian. Proceedings of the LREC Conference (2024).

12 Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., and Panchenko, A. Named entity recognition in Russian: Data, analysis, and model. Proceedings of the EMNLP Conference, 5284–5295 (2020).

13 den Dunnen, J.T., and Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Human Mutation, 15 (1), 7–12 (2000).

14 Lee, J., Yoon, W., Kim, S., et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36 (4), 1234–1240 (2020).

15 Gu, Y., Tinn, R., Cheng, H., et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3 (1), Article 2 (2021).

16 Islamaj, R., Kim, S., Kwon, D., et al. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of Biomedical Informatics, 92, 103132 (2019).

17 Chen, J. PDFPlumber: PDF text extraction library (online). URL: https://github.com/jsvine/pdfplumber

18 C2 Team. (n.d.). python-docx: Python library for creating and updating Microsoft Word (.docx) files. GitHub. URL: https://github.com/python-openxml/python-docx (accessed: 13.11.2025).

19 C2 Team. python-docx: Python library for creating and updating Microsoft Word (.docx) files (online). URL: https://github.com/python-openxml/python-docx

20 Tkachenko, M., Malyuk, M., Shevchenko, N., & Holubiev, B. (2023). Label Studio: Data labeling software. GitHub. URL: https://github.com/heartexlabs/label-studi (accessed: 13.11.2025).

21 Seal, R.L., et al. Genenames.org: the HGNC resources in 2023. Nucleic Acids Research, 51, D1003–D1009 (2023).

22 Tkachenko, M., Malyuk, M., Shevchenko, N., and Holubiev, B. Label Studio: Data labeling software (online). URL: https://github.com/heartexlabs/label-studio (accessed: 1.01.2023).

23 Richards, S., et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the ACMG and AMP. Genetics in Medicine, 17 (5), 405–424 (2015).

24 Amberger, J.S., Bocchini, C.A., Scott, A.F., and Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Research, 47, D1038–D1043 (2019).

25 Köhler, S., et al. The Human Phenotype Ontology in 2021. Nucleic Acids Research, 49, D1207–D1217 (2021).

26 Landrum, M.J., et al. ClinVar: improvements to accessing data. Nucleic Acids Research, 48, D835–D844 (2020).

27 Sherry, S.T., et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research, 29, 308–311 (2001).

28 Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., and Gómez-Berbís, J.M. Named entity recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces, 35 (5), 482–489 (2013).

29 Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32, 267–270 (2004).

30 Yang, X., Saha, S., Venkatesan, A., Tirunagari, S., Vartak, V., and McEntyre, J. Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms. Scientific Data, 10, Article 722 (2023).

[1,2]Самбетбаева М.А.,
Phd, ассоциированный профессор, ORCID ID: 0000-0001-9358-1614,
e-mail: sambetbayeva_ma_1@enu.kz
[1,2]*Серикбаева С.К.,
Phd, и.о.доцента, ORCID ID: 0000-0002-3627-3321,
*e-mail: inf_8585@mail.ru
[1,3]Сұлтанғазиева А.Н.,
магистр, докторант, ORCID ID: 0009-0009-9038-5234,
e-mail: anara77777@mail.ru
[1,4]Мукажанов Н.К.,
PhD, ассоциированный профессор, ORCID ID: 0000-0003-4835-5751,
e-mail: nurzhan.mukazhanov@narxoz.kz
[1,2]Абдығалым Б.Х.,
магистр, докторант, ORCID ID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com

[1]«Q» University, Алматы қ., Қазақстан
[2]Евразийский национальный университет имени Л. Н. Гумилева, Казахстан, г. Астана
[3]Международный университет Астаны, Казахстан, г. Астана,
[4]Университет Нархоз, Казахстан, г. Алматы,

СОЗДАНИЕ КОРПУСА АННОТИРОВАННЫХ МЕДИЦИНСКИХ ТЕКСТОВ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ О ГЕНЕТИЧЕСКИХ ЗАБОЛЕВАНИЯХ

Аннотация

В статье представлен аннотированный корпус, состоящий из клинических текстов на русском языке, полученных по результатам экзомного секвенирования. Данный корпус был разработан в поддержку задач автоматического определения именованных объектов и семантических связей в отношении генов, мутаций, наследственных заболеваний, фенотипических признаков и их клинической значимости. В ходе формирования корпуса были использованы отчеты фактического клинического экзомного секвенирования, данные прошли этапы предварительной анонимизации и нормирования текста. В процессе маркировки использовались международные стандарты и базы знаний, такие как HGVS, OMIM, ClinVar и HPO, а также обеспечивалась согласованность и точность биомедицинской информации. Корпус содержит более 25 000 биомедицинских объектов и более 6000 семантических связей, что делает его важным ресурсом в области клинической генетики с точки зрения объема и содержания. Аннотация проводилась вручную с участием нескольких

экспертов, и результаты сравнивались путем перекрестной проверки, а уровень согласия между аннотаторами оценивался с помощью специальных показателей. Полученные результаты свидетельствуют о высоком качестве и надежности корпуса. Готовый корпус позволяет эффективно использовать модели обработки естественного языка в области медицинской генетики для обучения и оценки, разработки систем поддержки принятия клинических решений и прикладных исследований для структурирования генетических данных.

Ключевые слова: генетические заболевания, обработка медицинских текстов, аннотированный корпус, экзомное секвенирование, клинические тексты, автоматическое распознавание именованных сущностей (NER), извлечение семантических отношений (RE).

[1,2]Sambetbayeva M.,
PhD, Associate Professor, ORCID ID: 0000-0001-9358-1614,
e-mail: sambetbayeva_ma_1@enu.kz
[1,2]*Serikbaeyva S.,
PhD, Acting Associate Professor, ORCID ID: 0000-0002-3627-3321,
*e-mail: inf_8585@mail.ru
[1,3]Sultangaziyeva A.,
Master, PhD student, ORCID ID: 0009-0009-9038-5234,
e-mail: anara77777@mail.ru
[1,4]Mukazhanov N.,
PhD, Associate Professor, ORCID ID: 0000-0003-4835-5751,
e-mail: nurzhan.mukazhanov@narxoz.kz
[1,2]Abdy-galym B.,
Master, PhD student, ORCID ID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com

[1]Q university, Almaty, Kazakhstan
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[3]Astana International University, Astana, Kazakhstan
[4]Narxoz University, Almaty, Kazakhstan

CONSTRUCTION OF AN ANNOTATED MEDICAL TEXT CORPUS FOR INFORMATION EXTRACTION ON GENETIC DISEASES

Abstract

The article presents an annotated corpus consisting of clinical texts in Russian, obtained by exomic sequencing. This corpus was developed to support the tasks of automatically identifying named objects and semantic relationships in relation to genes, mutations, hereditary diseases, phenotypic traits and their clinical significance. During the formation of the corpus, reports of actual clinical exomic sequencing were used, the data went through the stages of preliminary anonymization and text normalization. The labeling process used international standards and knowledge bases such as HGVS, OMIM, ClinVar, and HPO, and ensured consistency and accuracy of biomedical information. The corpus contains more than 25,000 biomedical objects and more than 6,000 semantic links, making it an important resource in the field of clinical genetics in terms of volume and content. The annotation was carried out manually with the participation of several experts, and the results were compared by cross-checking, and the level of agreement between the annotators was assessed using special indicators. The results obtained indicate the high quality and reliability of the case. The finished corpus makes it possible to effectively use natural language processing models in the field of medical genetics for teaching and evaluation, development of clinical decision support systems, and applied research for structuring genetic data.

Keywords: genetic diseases, medical text processing, annotated corpus, exome sequencing, clinical texts, named entity recognition (NER), relation extraction (RE).

Received December 13, 2025; revised April 24, 2026; accepted May 8, 2026.