



UDK:811.512

Nozima UMAROVA,  
ToshDO‘TAU magistranti  
E-mail: nozimaumarova333@gmail.com

ToshDO‘TAU prof. v.b., DSc Sh.Hamroyeva taqrizi asosida

### THEORETICAL AND PRACTICAL FOUNDATIONS OF CREATING A LEGAL QUESTION-AND-ANSWER SYSTEM IN THE UZBEK LANGUAGE BASED ON THE EXPERIENCE OF TURKIC LANGUAGES

Annotation

The article analyzes NLP resources (BERTurk, KazNERD, UD Treebanks) formed in the Turkish and Kazakh languages and proposes a theoretical and practical model for creating a legal question-and-answer (Legal QA, LQA) system in the Uzbek language based on them. The architecture is based on a hybrid principle: IR Retrieval (BM25/dense), LLM Reasoning (transformers) and Rule-based Constraint Filtering (normative checking). The agglutinative morphology of the Uzbek language, SOV order, formulaic constructions of official-legal discourse and the integration of annotation requirements based on UD into LQA are covered. As a result, an integrated architecture, corpus and annotation requirements roadmap for Uzbek LQA was developed.

**Keywords:** natural language processing, legal question-and-answer systems, legal corpus, hybrid approach, large language model, artificial intelligence, semantic role, lemmatization, morphological tagging, tokenization.

### ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ ОСНОВЫ СОЗДАНИЯ ПРАВОВОЙ СИСТЕМЫ ВОПРОСОВ И ОТВЕТОВ НА УЗБЕКСКОМ ЯЗЫКЕ НА ОСНОВЕ ОПЫТА ТЮРКСКИХ ЯЗЫКОВ

Аннотация

В статье анализируются ресурсы обработки естественного языка (BERTurk, KazNERD, UD Treebanks), сформированные на турецком и казахском языках, и предлагается теоретическая и практическая модель создания на их основе системы юридических вопросов и ответов (Legal QA, LQA) на узбекском языке. Архитектура основана на гибридном принципе: поиск информации (BM25/dense), логическое рассуждение (трансформаторы) и фильтрация ограничений на основе правил (нормативная проверка). Рассматриваются агглютинативная морфология узбекского языка, порядок слов SOV, формульные конструкции официально-юридического дискурса и интеграция требований к аннотированию на основе UD в систему LQA. В результате разработана интегрированная архитектура, корпус и дорожная карта требований к аннотированию для узбекской системы LQA.

**Ключевые слова:** обработка естественного языка, системы юридических вопросов и ответов, юридический корпус, гибридный подход, большая языковая модель, искусственный интеллект, семантическая роль, лемматизация, морфологическая разметка, токенизация.

### TURKIY TILLAR TAJRIBASI ASOSIDA O‘ZBEK TILIDA HUQUQIY SAVOL-JAVOB TIZIMINI YARATISHNING NAZARIY VA AMALIY ASOSLARI

Annotatsiya

Maqolada turk va qozoq tillarida shakllangan NLP resurslari (BERTurk, KazNERD, UD Treebanklar) tahlil qilinadi va ular asosida o‘zbek tilida huquqiy savol-javob (Legal QA, LQA) tizimini yaratishning nazariy hamda amaliy modeli taklif etiladi. Arxitektura gibrid tamoyilga tayangan: IR Retrieval (BM25/dense), LLM Reasoning (transformerlar) va Rule-based Constraint Filtering (normativ tekshiruv). O‘zbek tilining agglutinativ morfologiyasi, SOV tartibi, rasmiy-huquqiy diskursning formulaik konstruksiyalari va UD asosida annotatsiyalash talablarining LQAga integratsiyasi yoritiladi. Natijada o‘zbek LQA uchun integratsiyalashgan arxitektura, korpus va annotatsiya talablarining yo‘l xaritasi ishlab chiqildi.

**Kalit so‘zlar:** tabiiy tilni qayta ishlash, huquqiy savol-javob tizimlari, huquqiy korpus, gibrid yondashuv, katta til modeli, sun‘iy intellekt, semantik rol, lemmatizatsiya, morfologik teglash, tokenizatsiya.

**Kirish.** Raqamli texnologiyalar jadal rivojlanayotgan hozirgi davrda huquqiy axborotga tezkor kirish, aniq ma’lumot olish va fuqarolarga qulay maslahat berish davlat boshqaruvi va adliya tizimi uchun ustuvor yo‘nalishlardan biri bo‘lib qoldi. Raqamli adliya konsepsiyasi dunyo miqyosida elektron sudlar, avtomatlashtirilgan izlash tizimlari va sun‘iy intellektga asoslangan huquqiy yordam platformalarini shakllantirmoqda. O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev tomonidan ilgari surilgan “Raqamli O‘zbekiston – 2030” strategiyasi va unga muvofiq 2020-yil 5-oktabrda qabul qilingan PQ–6079-sonli qarorda davlat xizmatlarini raqamlashtirish, jamiyatning barcha sohalarida raqamli texnologiyalarni keng joriy etish ustuvor vazifalardan biri etib belgilangan [1]. Shunga mos ravishda, dunyo miqyosida Legal

AI, xususan Legal Question Answering (LQA) tizimlari hujjat izlash, normativ moslikni tekshirish va huquqiy maslahatni avtomatlashtirishda samaradorlik ko‘rsatmoqda. Biroq o‘zbek tilida LQA tizimi mavjud emas, bunga sabab ochiq huquqiy korpus va annotatsiyaning yetishmasligi, o‘zbek uchun domen modellar (LegalBERT va h.k.) yo‘qligi, agglutinativ morfologiya va rasmiy diskursning formulaik, ammo variativ sintaksisidir. Shunga qaramay, turk va qozoq tillaridagi ishonchli Natural Language Processing (NLP) resurslari o‘zbek LQA uchun transfer-learning imkoniyatini ochadi.

Tadqiqot maqsadi – turkiy tillar tajribasiga tayangan holda o‘zbek tilida LQA tizimini yaratishning nazariy asoslari, arxitekturasi va texnik talablarini ishlab chiqish. Tadqiqotning ilmiy yangiligi – o‘zbek LQA uchun integratsiyalashgan

arxitekturani (IR+LLM+rule-based) taklif etish, turkiy tillar tajribalarini moslashtirish konsepsiyasini ko'rsatishdan iboratdir.

Mavzuga oid adabiyotlarning tahlili. Turkiy tillarda buy o'nalishda ko'plab tadqiqotlar olib borilgan va amaliy ishlar qilingan. Xususan, Turk tilida NLP bo'yicha BERTurk (Turkish BERT) turk tilida kontekstual embeddinglar samaradorligini isbotladi [2]; KazNERD qozoq tilida ishonchli NER korpusini taqdim etdi [3]; UD (Universal Dependencies) bosqichma-bosqich turkiy tillar sintaksisini formal standartga olib keldi (Kazakh UD, Uzbek UD) [4][5].

Tadqiqot metodologiyasi. Tadqiqot quyidagi ilmiy yondashuvlarga asoslanadi:

1) Deduktiv metod – xalqaro LQA tajribalarini tahlil qilib, o'zbek tiliga moslashtirilgan arxitektura ishlab chiqildi.

2) Transfer-learning metodologiyasi – Turkic NLP modellariidan o'zbek tili uchun asos sifatida foydalanildi.

3) Arxiv va matn lingvistik tahlili – Lex.uz hujjatlari korpus sifatida o'rganildi.

4) Gibrid AI dizayni – IR Retrieval + LLM Reasoning + Rule-based filtr konsepsiyasi asoslandi.

Tahlil va natijalar. Turkiy tillar so'nggi o'n yil ichida tabiiy tilni qayta ishlash (NLP) bo'yicha sezilarli ilmiy yutuqlarga erishdi. Til tuzilishining agglutinatив xususiyatlari va boy morfologik tizim turkiy tillar uchun maxsus modellar, korpuslar va embeddinglar yaratishni talab qildi. Ayniqsa, turk va qozoq tillarida sifatli, ilmiy nashrlarda tasdiqlangan resurslar paydo bo'lgani o'zbek tilida ham shunga o'xshash tizimlarni yaratish uchun muhim asos yaratadi.

Turk tilida zamonaviy modellar va korpuslar keng miqyosda yaratilgan. Eng muhim ilmiy asoslangan resurslardan biri – BERTurk modeli bo'lib, u Şahin va Eryigit (2020) tomonidan katta turkcha korpus asosida yaratilib, turk tilidagi sintaktik va semantik vazifalarda BERT asosidagi eng kuchli modellar qatoriga kiradi [2]. Ushbu modelning muvaffaqiyati, birinchidan, turk tilining boy morfologik tizimiga moslashtirilgan tokenizatsiya, ikkinchidan, optimallashtirilgan transformer arxitekturasi bilan izohlanadi. Bundan tashqari, Çöltekin (2020) tomonidan turk tilidagi morfologiya va uning NLP uchun muammolari haqida keng qamrovli tahliliy tadqiqot amalga oshirilgan bo'lib, unda turk tilining agglutinatив xususiyati, paradigma variativligi va affiksatsiya chuqurligining mashinaviy modellar samaradorligiga ta'siri chuqur asoslab berilgan [6].

Qozoq tili ham turkiy tillar o'rtasida NLP bo'yicha faol rivojlanayotgan yo'nalishlardan biridir. Rasmiy ilmiy nashrlar orasida eng muhimlaridan biri KazNERD Mukhamadiyev (2021) tomonidan yaratilgan qozoq tilining Named Entity Recognition (NER) korpusi bo'lib, 300 000 dan ortiq tokenlarni qamrab oladi va qozoq tilida sintaktik-semantik tahlilni chuqurlashtirishga xizmat qiladi [3]. Sud hujjatlari bo'yicha qozoq tilida katta ko'lamli korpuslar hozircha cheklangan, biroq Qozoq Milliy Universiteti huzurida yuritilgan "Kazakh Legal Language Processing" loyihasi doirasida normativ-huquqiy matnlarning struktural tahlili yuzasidan ishlar olib borilmoqda[7].

Turkiy tillar tizimi umumiy genealogik ildizga ega bo'lib, ular orasida fonetik, morfologik va sintaktik o'xshashliklar yuqori darajada. Ushbu o'xshashliklar NLP jarayonlarida, xususan transfer-learning uchun asos bo'lib xizmat qiladi. Turkiy tillarga xos bo'lgan asosiy xususiyat – affikslarning ketma-ket qo'shilishi orqali so'z shakllarining cheksiz variantlarda hosil bo'lishidir. Bu holat morfologik tahlilchilar, lemmatizatorlar va NER modellar uchun katta ahamiyatga ega. Çöltekin (2020) turk tilida morfologik kombinatsiyalar soni millionlab bo'lishi mumkinligini ilmiy asoslagan [6]. Affikslar shaxs, egalik, kelishik, modal, zamon,

nisbat va negatsiya kabi grammatik munosabatlarni kodlaydi. Bu esa semantik vazifalarni murakkablashtiradi. Turkiy tillar, jumladan o'zbek tili, SOV tartibini ustuvor deb biladi, ammo pragmatik omillarga ko'ra so'zlar o'rnini erkin o'zgartirish mumkin. Bu huquqiy matnlar sintaksisini qayta ishlashda muhimdir. Ega-to'ldiruvchi-hol munosabatlari ko'pincha affikslarda kodlanadi. Shu bois SRL (semantic role labeling) modellarini turkiy tillar uchun maxsus moslashtirish talab etiladi.

Turkiy tillar bo'yicha ilmiy tajribalarga tayanadigan bo'lsak, o'zbek tili uchun texnik moslashuv imkoniyati nihoyatda keng va BERTurk, KazNERD, UD Treebank kabi ilmiy resurslarning mavjudligi transfer-learning asosida o'zbek LQA tizimini yaratishda juda qulay. Morfologik o'xshashlik tufayli embeddinglarni o'zbek tiliga moslashtirish osonlashadi. Turk va qozoq modellari o'zbek tilidagi affiksatsiya va sintaktik strukturaga yaqinligi tufayli turkiy tillarda huquqiy matnlar kam o'rganilgan bo'lsa-da, mavjud tajribalar o'zbek tili uchun uslubiy asos yaratadi.

O'zbek tilining morfologik, sintaktik va pragmatik xususiyatlari huquqiy javol-javob tizimini yaratishda markaziy o'rin tutadi. Tilning agglutinatив tuzilishi, murakkab affiksatsiya tizimi, rasmiy huquqiy diskursning o'ziga xos terminologiyasi va sintaktik konstruksiyalari LQA arxitekturasini ishlab chiqishda maxsus yondashuvni talab etadi.

O'zbek tili agglutinatив til bo'lib, grammatik ma'no qo'shimchalar orqali ifodalanadi. Har bir qo'shimcha o'zining aniq semantik yuklamasiga ega va so'zning grammatik rolini belgilaydi. Umuman olganda o'zbek tili affiksatsiyaning tizimli tabiati va grammatik kategoriyalarning ketma-ket affikslar orqali qatlamlanishi jihatdan o'ziga xos tabiatga ega [8].

Huquqiy matnlarda fe'llar asosiy semantik yukni ko'taradi. Fe'l paradigmasi zamon, nisbat, modallik, inkor-tasdiq, shaxs-son kabi ma'nolarni o'z ichiga oladi. Bu ko'rinishlar huquqiy iboralarda semantik o'zgarishlar keltirib chiqaradi. Masalan, "ta'qiqlanadi", "ruxsat etiladi", "majburdir", "amalga oshirilishi shart" kabi konstruksiyalar normativ kuchga ega bo'lib, LQA tizimi uchun modal-huquqiy tasniflashni talab qiladi. Shuningdek, egalik va kelishik qo'shimchalari ham alohida ahamoyatha ega. O'zbek tilida 6 ta kelishik va 3 darajadagi egalik kategoriyasi mavjud. Bu grammatika elementlari huquqiy subyektni, huquqiy obyektini, mulkchilik munosabatlarini, vakolat va javobgarlikni aniqlashda markaziy funksiyani bajaradi. Normativ matnlarda "fuqarolarning huquqlari", "tadbirkorlik subyektlarining majburiyatlari", "ma'muriy javobgarlik to'g'risidagi kodeks" kabi birikmalar aynan egalik va kelishik kategoriyalari orqali aniqlashtiriladi.

Morfologik boylik sababli lemmatizatsiya, morfologik teglash, Affiks-aware tokenization kabi jarayonlar LQA uchun zarur bazaviy komponentlardir. O'zbek tilida ishlatiluvchi UD (Universal Dependencies) morfologik tavsifi modellashtirish uchun asos sifatida qabul qilinishi mumkin. O'zbek tili SOV so'z tartibiga ega bo'lsa-da, rasmiy-huquqiy matnlarda sintaktik strukturalarning qat'iy va normativ tabiatga egaligi ko'zga tashlanadi. Huquqiy hujjatlarning normativ tabiatidan kelib chiqib, quyidagi shakllar juda faol: "amalga oshiriladi", "tasdiqlanadi", "qabul qilinadi", "ko'rib chiqiladi", "belgilanadi". Bu konstruksiyalar ko'pincha majhul nisbatda bo'lib, bajaruvchini ko'rsatmaydi; ular me'yorga yo'naltirilgan. LQA tizimi bunday konstruksiyalardan normativ harakatning yo'nalishini aniqlashi kerak. Huquqiy matnlarda "agar ... bo'lsa, ...", "qonunda boshqacha belgilangan bo'lsa...", "zarur hollarda...", "shuningdek..." kabi konstruksiyalar ko'p uchraydi. Shart mayli orqali huquqiy javobgarlikning shartlari aniqlanadi. LQA modelining logical

reasoning modulida aynan shart bog'lovchilar bilan ishlash muhim ahamiyatga ega.

Huquqiy diskursning asosiy belgilaridan yana biri formulaik konstruksiyalardir. O'zbek matnlarida ham bu norma kuzatiladi. Masalan, "ushbu Qonunga muvofiq...", "tegishli davlat organi...", "mazkur Kodeksda nazarda tutilgan tartibda..." kabi. LQA tizimi ushbu formulaik birliklarni pattern-level annotatsiya orqali tanib olishi kerak. O'zbek tilida huquqiy savol-javob tizimini ishlab chiqish uchun quyidagi annotatsiya resurslari zarur:

1. Named Entity Recognition (NER). Shaxs, tashkilot, huquqiy institut, hudud, hujjat nomlari bu tarkibga kiradi. Turk va qozoq tillaridagi NER tajribasi (KazNERD) buni qo'llab-quvvatlaydi.

2. Legal Entity Recognition (LER). Huquqiy subyektlar "davlat organi", "fuqaro", "xususiy tadbirkor", "ijro hokimiyati" kabi terminlar.

3. Semantic Role Labeling (SRL) Huquqiy voqea ishtirokchilarini (Agent, Action, Object) aniqlash.

4. Relation Extraction. "Subyekt-modda", "qoidaning tatbiq doirasi", "ruxsat-ta'qiqlov bog'lanishi".

Huquqiy savol-javob tizimi uchun asosiy poydevor — puxta tuzilgan huquqiy korpusdir, LQA samaradorligi bevosita korpus sifati va hujjatlar semantik belgilanilishiga bog'liq. Bu jarayonda Lex.uz – O'zbekiston Respublikasi qonunlari va qarorlari platformasidan foydalanish mumkin.[9] Ma'lumotni tozalashda ortiqcha texnik belgilarni olib tashlash; band, modda, paragraf raqamlarini standartlashtirish; formulaik iboralarni normalizatsiya qilish kabi ishlar bajarilishi kerak.

Keyingi bosqichda esa huquqiy korpus yaratish asosiy vazifa sanaladi. O'zbek tilida hozirda maxsus huquqiy korpus yo'qligi sababli korpusni strukturali tarzda yaratish zarur. Huquqiy korpus normativ matnlar, sud qarorlari (agar mavjud bo'lsa), izoh va sharhlar, savol-javob juftliklari (annotatsiya qilingan) kabi segmentlardan iborat bo'lishi kerak. Bu segmentlar o'zaro bog'langan holda RAG (retrieval-augmented generation)ni amalga oshiradi.

Dunyo bo'yicha ilmiy nashrlardagi LQA tizimlari asosan gibril yondashuvga tayangan. Huquqiy domenning murakkabligi sabab, faqat transformerlarga tayanish noto'g'ri bo'ladi. LQA uchun eng samarali arxitektura aynan quyidagi uch bosqichdan iborat:

1. IR Retrieval (BM25). Huquqiy savol-javobning birinchi bosqichi tegishli hujjatlarni izlab topishdir. BM25 algoritmi shu kungacha eng ishonchli IR modellardan biri bo'lib, LexGLUE va COLIEE musobaqalarida bazaviy yechim sifatida qo'llanilgan [1][2]. IR bosqichi savolni tokenizatsiya qilish, matnlardan BM25 bo'yicha relevans ballini hisoblash, eng mos modda yoki bandni tanlab olish kabi vazifalarni bajaradi.

2. LLM Reasoning (Semantik tahlil va xulosa chiqarish). Tanlangan matn asosida savolning mazmunini

semantik tahlil qilish – LQA yuragi hisoblanadi. Bu jarayonda huquqiy matnlar uzoq, murakkab va ko'p qatlamli bo'lgani uchun transformer arxitekturasi eng mos keladi. O'zbek tili uchun maxsus LegalBERT mavjud emas. Shuning uchun transfer-learning asosida TurkBERT, multilingual BERT (mBERT) kabi modellarni moslashtirish ilmiy jihatdan asosli yechim bo'ladi. Bu yondashuv turkiy tillar tipologik o'xshashlikka ega ekanini ko'rsatgan Çöltekin (2020) tadqiqoti bilan ham o'xshash [6].

3. Rule-based Constraint Filtering (Normativ moslikni tekshirish). LLMlar ko'pincha semantik jihatdan to'g'ri bo'lsa ham, huquqiy normaga mos javob bera olmaydi. Shu sababli yakuniy bosqichda huquqiy moddalarga zid bo'lmaganlikni tekshirish, ta'qiqlov va ruxsat konstruksiyalarini aniqlash, shartli gaplar (agar..., unda...) ning to'g'ri talqini tekshirish uchun rule-based filtr zarur. Huquqiy matnlarda formulaik struktura qat'iyligi aynan rule-based qayta tekshiruvni osonlashtiradi.

O'zbek tilida LQA tizimi quyidagi asosiy modullardan tashkil topadi:

1. Savolni tahlil qilish moduli (Question Understanding). Bunda savol turi va modal ma'no aniqlanadi.

2. Matnni semantik indekslash (Semantic Indexing). Lemmatizatsiya, morfologik teglash, UD sintaktik daraxtlari.

3. Huquqiy hujjatlarni qidirish moduli (Retrieval Engine) BM25 baseline asosida amalga oshiriladi.

4. Entailment/ Yes-No tasnifi. Xalqaro LQA tizimlarida aynan entailment modeli (matnda savolga ijobiy yoki salbiy dalil mavjudligini aniqlash) eng muhim bosqichdir.

5. Javob generatsiyasi. LLM asosida huquqiy shakllantiriladi va bunda javob normativ hujjatga tayangan bo'lishi, kontekstga mos bo'lishi, o'zboshimchalik qilmasligi (gollyutsinatsiya xavfini kamaytirish uchun IR + rule-based uyg'un ishlatiladi) kerak.

**Xulosa.** Tadqiqot natijalari asosida birinchi bor o'zbek tilida huquqiy savol-javob tizimi yaratish konsepsiyasi ishlab chiqildi. Bu konsepsiya asosida adliya organlari uchun tezkor izlash mexanizmlarini joriy qilish mumkin hamda fuqarolar uchun avtomatik huquqiy yordam botlari ishlab chiqish imkoniyati paydo bo'ladi. Bu raqamli adliya konsepsiyasini rivojlantirishda muhim qadam hisoblanadi.

Kelgusidagi tadqiqotlar uchun takliflar sifatida o'zbek LegalBERT modelini yaratish (Chalkidis, 2020 metodologiyasiga asoslanib); keng ko'lamli o'zbek huquqiy korpusini shakllantirish va ochiq e'lon qilish; annotatsiya standartlarini ishlab chiqish (NER, LER, SRL, RE); explainable Legal AI yo'nalishida tadqiqotlarni kengaytirishni keltirish mumkin. Ushbu yo'nalish bo'yicha izlanishlar O'zbekistonda huquqiy texnologiyalarni zamonaviy bosqichga olib chiqishga xizmat qiladi.

#### ADABIYOTLAR

1. O'zbekiston Respublikasi Vazirlar Mahkamasi. Raqamli iqtisodiyotni rivojlantirish bo'yicha qo'shimcha chora-tadbirlar to'g'risida: PQ-6079-son qaror, 2020-yil 5-oktabr. – Elektron ma'lumot. – URL: <https://lex.uz/docs/5030048> (murojaat qilingan sana: 11.12.2025).
2. GitHub - turkish-bert by stefan-it. – Elektron resurs. – URL: <https://github.com/stefan-it/turkish-bert> (murojaat qilingan: 11.12.2025).
3. GitHub - KazNERD by IS2AI. – Elektron resurs. – URL: <https://github.com/IS2AI/KazNERD> (murojaat qilingan: 11.12.2025).
4. Universal Dependencies - Kazakh. – Elektron resurs. – URL: <https://universaldependencies.org/kk/> (murojaat qilingan: 11.12.2025).
5. Universal Dependencies - Uzbek. – Elektron resurs. – URL: <https://universaldependencies.org/uz/index.html> (murojaat qilingan: 11.12.2025).
6. Çöltekin, Ç. A Corpus of Turkish Offensive Language on Social Media // Proceedings of the Twelfth Language Resources and Evaluation Conference. – Marseille, France, 2020. – P. 6174–6184. – European Language Resources Association.
7. Rakhimova, D.; Turarbek, A.; Karyukin, V.; Sarsenbayeva, A.; Alieyev, R. Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation // Computers. – 2025. – Vol. 14. – Art. 354. – DOI: <https://doi.org/10.3390/computers14090354>.
8. kurs ishi. - Elektron manba. - URL: <https://www.scribd.com/document/920374070/Grammatik-Kategoriyalar-Va-Ularning-Uslubiy-Xususiyatlari> (murojaat qilingan sana: 11.12.2025).
9. Lex.uz: rasmiy huquqiy ma'lumotlar portali. - Elektron manba. - URL: <https://lex.uz/> (murojaat qilingan sana: 11.12.2025).