EURASIAN NATIONAL UNIVERSITY

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ
ҰЛТТЫҚ УНИВЕРСИТЕТІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РЕСПУБЛИКИ КАЗАХСТАН
ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ Л.Н. ГУМИЛЕВА

MINISTRY OF EDUCATION AND SCIENCE
OF THE REPUBLIC OF KAZAKHSTAN
L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY

16-18 маусым
Нұр-Сұлтан, 2022

# «TURKLANG 2022»

«Түркі тілдерін компьютерлік өңдеу»
атты X халықаралық конференция
ЕҢБЕКТЕРІ

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»

# «TURKLANG 2022»

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages

# «TURKLANG 2022»

**ҚАЗАҚСТАН РЕСПУБЛИКАСЫ**
**БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ**
**Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ**

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ**
**РЕСПУБЛИКИ КАЗАХСТАН**
**ЕВРАЗИЙСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ**
**ИМЕНИ Л.Н. ГУМИЛЕВА**

**MINISTRY OF EDUCATION AND SCIENCE OF**
**THE REPUBLIC OF KAZAKHSTAN**
**L.N. GUMILYOV EURASIAN NATIONAL UNIVERSITY**

# «TURKLANG 2022»
# «Түркі тілдерін компьютерлік өңдеу»
# атты X халықаралық конференция
# ЕҢБЕКТЕРІ
16-18 маусым 2022 ж.

# ТРУДЫ
# X Международной конференции
# «Компьютерная обработка тюркских языков»
# «TURKLANG 2022»
16-18 июня 2022 г.

# PROCEEDINGS
# of the X International Conference
# on Computer processing of Turkic Languages
# «TURKLANG 2022»
16-18 June 2022

Нұр-Сұлтан, 2022

**УДК 80/81:004**
**ББК 81.2:32-973**
**Т 90**

**Техникалық редакция:**
Ергеш Б.Ж.
Елибаева Г.К.
Турсынова Н.А.

**Т 90** ТҮРКІ ТІЛДЕРІН КОМПЬЮТЕРЛІК ӨҢДЕУ. X халықаралық конференция: Еңбектері = КОМПЬЮТЕРНАЯ ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ. X международная конференция: Труды. / - Нұр-Сұлтан: «Булатов А.Ж.» ЖК, 2022.= Нур-Султан: ИП «Булатов А.Ж.»

**ISBN 978-601-326-645-9**

Жинақта «Түркі тілдерін компьютерлік өңдеу» атты X халықаралық конференция қатысушыларының баяндамалары енген.

Компьютерлік лингвистика бағыты бойынша оқитын студенттерге, магистранттарға, докоранттарға және мамандарға арналған.

Жинақ «BR11765535» Қазақ тілі мәдениетін арттыру және функцияларды кеңейту бойынша ғылыми-лингвистикалық негіздер мен IT-ресурстарды әзірлеу» бағдарламасы есебінен жарияланды.

В сборнике представлены доклады участников X международной конференции «Компьютерная обработка тюркских языков».

Предназначен для студентов, магистрантов, докторантов и специалистов специализирующихся в областях компьютерной лингвистика.

Сборник издан за счет средств программы BR11765535 «Разработка научно-лингвистических основ и IT-ресурсов по расширению функций и повышению культуры казахского языка».

**УДК 80/81:004**
**ББК 81.2:32-973**

**ISBN 978-601-326-645-9**

**Ұйымдастырушылар:**

**Қазақстан Республикасы Білім және ғылым министрлігі**
**Л.Н. Гумилев атындағы Еуразия ұлттық университеті**
Ақпараттық технологиялар факультеті, «Жасанды интеллект» ҒЗИ

**Қазақстан Жасанды интеллект қауымдастығы**

**Татарстан Республикасының Ғылым академиясы**
«Қолданбалы семиотика» ҒЗИ

**Ресей Жасанды интеллект қауымдастығы**

**Стамбул техникалық университеті**

---

**Организаторы:**

**Министерство образования и науки Республики Казахстан**
**Евразийский национальный университет имени Л.Н. Гумилева**
Факультет информационных технологий, НИИ «Искусственный интеллект»

**Казахстанская ассоциация искусственного интеллекта**

**Академия наук Республики Татарстан**
**НИИ «Прикладная семиотика»**

**Российская ассоциация Искусственного интеллекта**

---

**Organizers:**

**Ministry of Education and Science of the Republic of Kazakhstan**
**L.N. Gumilyov Eurasian National University**
Faculty of Information Technologies, “Artificial Intelligence”
Scientific Research Institute

**Kazakhstan Academy of Artificial Intelligence**

**Tatarstan Academy of Sciences**
Institute of Applied Semiotics

**Russian Association of Artificial Intelligence**

**Technical University of Istanbul (ITU)**

**Программалық комитет / Программный комитет / Program Committee:**

1. Шәріпбай Алтынбек Әміруұлы (Нұр-Сұлтан, Қазақстан) – тең төраға
2. Сулейманов Джавдет Шевкетович (Қазан, Татарстан Республикасы, РФ) - тең төраға
3. Ешреф Адалы (Ыстанбұл, Түркия) – тең төраға
4. Абдурахмонова Нилуфар (Ташкент, Өзбекстан)
5. Алтынбек Гулила (Үрімші, Қытай)
6. Бекманова Гульмира Тылеубердиевна (Нұр-Сұлтан, Қазақстан)
7. Гатиатуллин Айрат Рафизович (Қазан, Татарстан Республикасы, РФ)
8. Дыбо Анна Владимировна (Мәскеу, РФ)
9. Ергеш Бану Жантуганкызы (Нұр-Сұлтан, Қазақстан)
10. Желтов Валериан Павлович (Чебоксары, Чувашия Республикасы, РФ)
11. Исраилова Нелла Амантаевна (Бішкек, Қырғызстан)
12. Кубединова Ленара Шакировна (Симферополь, Қырым Республикасы, РФ)
13. Мамедова Масума Гусейновна (Баку, Әзірбайжан)
14. Мамырбаев Оркен Жумажанович (Алматы, Қазақстан)
15. Муканова Асель Сериковна (Нұр-Сұлтан, Қазақстан)
16. Офлазер Кемаль (Доха, Катар)
17. Разахова Бибигул Шамшановна (Нұр-Сұлтан, Қазақстан)
18. Садыков Ташполот (Бішкек, Қырғызстан)
19. Салчак Аэлита Яковлевна (Қызыл, Тува Республикасы, РФ)
20. Сиразитдинов Зиннур Амирович (Уфа, Башқұртстан Республикасы, РФ)
21. Сулайманов Мухаммад-али (Қырым Республикасы, РФ)
22. Татевосов Сергей Георгиевич (Мәскеу, РФ)
23. Торотоев Гаврил Григорьевич (Якутск, Саха Республикасы (Якутия), РФ)
24. Тукеев Уалишер Ануарбекович (Алматы, Қазақстан)
25. Рахимова Диана Рамазановна (Алматы, Қазақстан)
26. Чумакаев Алексей Эдуардович (Горно-Алтайск, Алтай, РФ)

# МАЗМҰНЫ / СОДЕРЖАНИЕ / CONTENT

## ПЛЕНАРЛЫҚ МӘЖІЛІС БАЯНДАМАЛАРЫ
## ДОКЛАДЫ ПЛЕНАРНОГО ЗАСЕДАНИЯ
## PLENARY MEETING


| | | |
|---|---|---|
| 1 | ***Дихан Қамзабекұлы***<br>*Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*<br>**ҚАНЫШ – АЛАШ ҒАЛЫМЫ** | 14 |
| 2 | ***Шарипбай Алтынбек Амирович***<br>*Евразийский национальный университет им. Л.Н. Гумилева, Нур-Султан, Казахстан*<br>**ПРОБЛЕМЫ СОЗДАНИЯ АЛФАВИТА КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ЛАТИНСКОЙ ГРАФИКИ** | 20 |
| 3 | ***Сулейманов Джавдет Шевкетович***<br>*Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Татарстан, Россия*<br>**ИНФОКОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ И ТАТАРСКИЙ ЯЗЫК** | 41 |
| 4 | ***Тукеев Уалшер Ануарбекович***<br>*Казахский Национальный Университет им. Аль-Фараби, Алматы, Казахстан*<br>**ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ ПО ВЫЧИСЛИТЕЛЬНОЙ МОДЕЛИ МОРФОЛОГИИ НА ОСНОВЕ ПОЛНОГО НАБОРА ОКОНЧАНИЙ** | 58 |

**ТҮРКІ ТІЛДЕРІ – ЖАҢА ИНТЕЛЛЕКТУАЛДЫ ТЕХНОЛОГИЯЛАР МЕН БІЛІМДЕРДІ ӨҢДЕУ ЖҮЙЕЛЕРІН ҚҰРУДЫҢ НЕГІЗІ РЕТІНДЕ**

**ТЮРКСКИЕ ЯЗЫКИ КАК ОСНОВА ДЛЯ СОЗДАНИЯ НОВЫХ ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЙ И СИСТЕМ ОБРАБОТКИ ЗНАНИЙ**

**TURKIC LANGUAGES AS THE BASIS FOR THE CREATION OF NEW INTELLIGENCE TECHNOLOGIES AND KNOWLEDGE PROCESSING SYSTEMS**


| | | |
|---|---|---|
| 1 | ***Исмаилов Исмаил Ариф оглы***<br>*Азербайджанский архитектурно-строительный университет, Баку, Азербайджан*<br>**ПРИМЕНЕНИЕ СТРУКТУРНОЙ ЭКСПЕРТНОЙ СИСТЕМЫ В ЭТИМОЛОГИЧЕСКИХИЗЫСКАНИЯХ** | 69 |
| 2 | ***Muratbekova Sh.***<br>*Tashkent State University of Uzbek language and literature named after Alisher Navoiy, Tashkent, Uzbekistan*<br>**HOW TO CREATE TEXT VALIDATION SOFTWARE FOR A PLUGIN** | 79 |
| 3 | ***Балсаидов А.Ш.***<br>*Алматинский Технологический Университет, Алматы, Казахстан*<br>**ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ИЗОБРАЖЕНИЙ ДЛЯ НАИЛУЧШЕГО РАСПОЗНАВАНИЯ ТЕКСТА** | 84 |
| 4 | ***Сулейманов Д.Ш., Гильмуллин Р.А., Мухаметзянов И.Р.***<br>*Институт прикладной семиотики Академии Наук Республики Татарстан, Казань, Татарстан, Россия*<br>**О ПОТЕНЦИАЛЕ ГРАММАТИКИ ТАТАРСКОГО ЯЗЫКА ДЛЯ РАЗРАБОТКИ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ** | 92 |
| 5 | ***Сыздықова Г. О.***<br>*Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*<br>**А.БАЙТҰРСЫНҰЛЫ ТЫНЫС БЕЛГІЛЕРІНІҢ ТҮРЛЕРІ МЕН ҚОЛДАНЫСЫ ТУРАЛЫ** | 104 |
| 6 | ***Амангелді Н, Кудубаева С. А., Турсынова Н. А., Баймаханова А., Ерболатова А., Абдиева С.***<br>*Л. Н. Гумилев атындағы Еуразия ұлттық университеті Нұр-Сұлтан, Қазақстан*<br>**ҚАЗАҚ ЫМ ТІЛІНДЕГІ СӨЗДЕРДІҢ КӨРСЕТІЛУ ПІШІНДЕРІН ӨЗГЕ ЫМ ТІЛДЕРІМЕН САЛЫСТЫРМАЛЫ АНАЛИЗІ** | 113 |

## МӘТІНДЕРДІ МОРФОЛОГИЯЛЫҚ, СИНТАКСИСТІК ЖӘНЕ СЕМАНТИКАЛЫҚ ӨҢДЕУ ТЕХНОЛОГИЯЛАРЫ

## ТЕХНОЛОГИИ МОРФОЛОГИЧЕСКОЙ, СИНТАКСИЧЕСКОЙ И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

## TECHNOLOGIES FOR MORPHOLOGICAL, SYNTACTIC AND SEMANTIC TEXT PROCESSING


1 ***Сайранбекова А. Д., Бекманова Г.Т.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*
**МАШИНАЛЫҚ АУДАРМА НЕГІЗІНДЕ МӘТІНДЕГІ ЖАҒЫМСЫЗ СЕНТИМЕНТТІ АНЫҚТАУ** 125

2 ***Кудубаева С.А., Жусупова Б.Т.***
*Евразийский национальный университет им. Л.Н. Гумилева, Нур-Султан, Казахстан*
*Костанайский региональный университет имени А. Байтурсынова, Костанай, Казахстан*
**О ВОЗМОЖНОСТИ УЧЕТА СЕМАНТИЧЕСКОЙ СОСТАВЛЯЮЩЕЙ В СИСТЕМЕ СУРДОПЕРЕВОДА С КАЗАХСКОГО ЯЗЫКА НА КАЗАХСКИЙ ЯЗЫК ЖЕСТОВ** 135

3 ***Азат Абдысадыр уулу***
*Канцелярия Администрации Президента Кыргызской Республики, Бишкек, Кыргызстан*
**МЕТОД БИНАРНЫХ СВЯЗЕЙ ДЛЯ ВИЗУАЛИЗАЦИИ СМЫСЛА ПРЕДЛОЖЕНИЯ** 145

4 ***Khamroeva Shahlo Mirdjanovna***
*Tashkent State university of Uzbek language and literature named Alisher Navai, Tashkent, Uzbekistan*
**FINITE STATE MACHINE MODEL OF NOUNS FOR UZBEK LANGUAGE MORPHOLOGICAL ANALYZER** 153

5 ***Nasrullayeva A., Mukanova A.***
*Astana International University, Nur-Sultan, Kazakhstan*
**RESEARCH AND ANALYSIS OF MECHANISMS FOR DETECTING PROHIBITED CONTENT ON THE INTERNET** 165

6 ***Оралбекова І.Т., Ергеш Б.Ж.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*
**ҚАЗАҚ ТІЛІНДЕГІ ҚОНАҚҮЙЛЕР ТУРАЛЫ ПІКІРЛЕРДІҢ РЕҢКІН АСПЕКТІЛЕРГЕ ТАЛДАУ** 172

## ТҮРКІ ТІЛДЕРІНІҢ ЭЛЕКТРОНДЫ МӘТІНДІК ЖӘНЕ АУДИО КОРПУСТАРЫ

## ЭЛЕКТРОННЫЕ ТЕКСТОВЫЕ И АУДИО КОРПУСЫ ТЮРСКИХ ЯЗЫКОВ

## ELECTRONIC TEXT AND AUDIO CORPUS OF THE TURKIC LANGUAGES


| | | |
|---|---|---|
| 1 | ***Мухамедшин Д.Р.***<br>*Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Татарстан, Россия*<br>**НЕКОТОРЫЕ НОВЫЕ ПОИСКОВЫЕ И ИССЛЕДОВАТЕЛЬСКИЕ ВОЗМОЖНОСТИ СИСТЕМЫ УПРАВЛЕНИЯ КОРПУСНЫМИ ДАННЫМИ** | 180 |
| 2 | ***Мадиева Г. Б., Мансурова М. Е.***<br>*Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан*<br>**РАЗРАБОТКА УЧЕБНОГО МАТЕРИАЛА ПО КОРПУСНОЙ ЛИНГВИСТИКЕ: К ВОПРОСУ О ЯЗЫКОВОМ РЕСУРСЕ** | 188 |
| 3 | ***Сакенова Ж. Ж., Маткаримов Б.Т.***<br>*Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*<br>**ТҮРКІ ТІЛДЕРІНДЕГІ ЖӘНЕ ТҮРКІ МУЗЫКАСЫНДАҒЫ ӘН ЖАЗБАЛАРЫНЫҢ МУЗЫКАЛЫҚ-КОРПУСЫ** | 196 |
| 4 | ***Abjalova Manzura Abdurashetovna***<br>*Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, Tashkent, Uzbekistan*<br>**THE AUTHOR'S CORPUS OF ALISHER NAVOI AND IT'S IMPORTANCE** | 208 |
| 5 | ***Хакимов Б. Э., Шаехов М.Р.***<br>*Институт прикладной семиотики Академии Наук Республики Татарстан, Казань, Татарстан, Россия*<br>*Казанский федеральный университет, Казань, Татарстан, Россия*<br>**СРАВНЕНИЕ КАЧЕСТВА МАШИННЫХ ПЕРЕВОДЧИКОВ В РУССКО-ТАТАРСКОЙ ПАРЕ С ПОМОЩЬЮ ТЕСТОВОГО ПАРАЛЛЕЛЬНОГО КОРПУСА** | 212 |

| | | |
|---|---|---|
| 6 | ***Леспекова А.А., Муканова А.С., Елибаева Г.К.***<br>*Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Қазақстан, Нұр-Сұлтан*<br>*Астана Халықаралық университеті, Қазақстан, Нұр-Сұлтан*<br>**ТЫЙЫМ САЛЫНҒАН КОНТЕНТТІ АНЫҚТАУ ҮШІН МӘТІНДІК КОРПУС ҚҰРУ** | 223 |
| **7** | ***Abdurakhmonova N., Tuliyev U., Ismailov A., Abduvahobo G.***<br>*National university of Uzbekistan, Tashkent, Uzbekistan*<br>*Andijan Machine Building institute, Andijan, Uzbekistan*<br>*Fergana state university, Fergana, Uzbekistan*<br>**UZBEK ELECTRONIC CORPUS AS A TOOL FOR LINGUISTIC ANALYSIS** | 231 |
| 8 | ***Кутдусова Е. П., Прокопьев Н. А.***<br>*Казанский Федеральный Университет, Казань, Татарстан, Россия*<br>**СИНТЕЗ ТАТАРСКОЙ РЕЧИ ПРИ ПОМОЩИ ГЛУБОКОГО ОБУЧЕНИЯ НА ОСНОВЕ МОДЕЛИ VITS** | 241 |

## ЖАСАНДЫ ИНТЕЛЛЕКТ ЖҮЙЕЛЕРІНДЕ АҚПАРАТТЫ ҰСЫНУ МЕН ӨҢДЕУДІҢ СЕМИОТИКАЛЫҚ МОДЕЛДЕРІ

## СЕМИОТИЧЕСКИЕ МОДЕЛИ ПРЕДСТАВЛЕНИЯ И ОБРАБОТКИ ИНФОРМАЦИИ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

## SEMIOTIC MODELS OF INFORMATION REPRESENTATION AND PROCESSING IN ARTIFICIAL INTELLIGENCE SYSTEMS

1 ***Маңмұрын М. М., Шәріпбай А.Ә***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*
**ШАБЛОН ҚҰЖАТ ҮЛГІЛЕРІНІҢ ФРЕЙМДІК МОДЕЛІ** 250

2 ***Aktaeva A., Kubigenova A., Esmagambetova G.***
*Kokshetau University named after Sh.Ualikhanov, Kokshetau, Kazakhstan*
*Kazakh Agrotechnical University named after S. Seifullin, Nur-Sultan, Kazakhstan*
*Mongolian University of Science and Technology, Ulaanbaatar, Mongolia*
**SEMANTIC ASPECTS OF SECURITY IN BLOCKCHAIN TECHNOLOGIES: CRYPTOSEMANTICS** 259

## ТҮРКІТІЛДЕС ИНТЕРНЕТ-РЕСУРСТАР, ОНТОЛОГИЯЛАР, ТЕЗАУРУСТАР ЖӘНЕ СӨЗДІКТЕР

## ТЮРКОЯЗЫЧНЫЕ ИНТЕРНЕТ-РЕСУРСЫ, ОНТОЛОГИИ, ТЕЗАУРУСЫ И СЛОВАРИ

## TURKIC INTERNET RESOURCES, ONTOLOGIES, THESAURI AND DICTIONARIES


| | | |
|---|---|---|
| 1 | ***Бурнашев Р. А., Галимов М.Р.***<br>*Институт прикладной семиотики Академии Наук Республики Татарстан, Казань, Татарстан, Россия*<br>**ПОСТРОЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ ЯЗЫКОВЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ ГЕОИНФОРМАЦИОННЫХ СИСТЕМ** | 268 |
| 2 | ***Nizomova Zuhra Komil qizi***<br>*Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, Tashkent, Uzbekistan*<br>**THE IMPORTANCE OF THE THESAURUS OF CHEMICAL TERMS** | 274 |
| 3 | ***Шарипбай А.А., Омарбекова А.С.***<br>*Евразийский национальный университет им. Л. Н. Гумилева, Нур-Султан, Казахстан*<br>**СТРУКТУРА БАЗЫ ДАННЫХ ТЕРМИНОВ ШКОЛЬНЫХ ПРЕДМЕТОВ И ВИДЫ ЗАПРОСОВ К НИМ** | 280 |
| 4 | ***Кадирхан А. К.***<br>*Ш. Шаяхметов атындағы «Тіл-Қазына» ұлттық ғылыми-практикалық орталығы, Нұр-Сұлтан, Қазақстан*<br>**«ТӘУЕЛСІЗ» СӨЗІНІҢ ОРФОГРАММАСЫ ЖӘНЕ ҚОЛДАНЫС ЕРЕКШЕЛІГІ** | 286 |
| 5 | ***Жұмаш Б., Муканова А. С.***<br>*Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*<br>*Астана Халықаралық университеті, Нұр-Сұлтан, Қазақстан*<br>**ОНТОЛОГИЯЛЫҚ МОДЕЛЬ НЕГІЗІНДЕ КӨЛІК САЛЫҒЫ БОЙЫНША АҚПАРАТТЫҚ ЖҮЙЕНІ ӘЗІРЛЕУ** | 294 |
| 6 | ***Орынбай Л. О., Сайранбекова А. Д., Елибаева Г. К., Бекманова Г. Т***<br>*Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*<br>**ҚАЗАҚ ЕСІМДЕРІНІҢ СЕМАНТИКАЛЫҚ БАЗАСЫНЫҢ ҚҰРЫЛЫМЫН АНЫҚТАУ ЖОЛДАРЫ** | 302 |

## КОМПЬЮТЕРЛІК ЖҮЙЕЛЕРДІҢ ҰЛТТЫҚ ЛОКАЛИЗАЦИЯСЫ МЕН ТЕРМИНОЛОГИЯ

## НАЦИОНАЛЬНАЯ ЛОКАЛИЗАЦИЯ КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕРМИНОЛОГИЯ

## NATIONAL LOCALIZATION OF COMPUTER SYSTEMS AND TERMINOLOGY


| | | |
|---|---|---|
| 1 | ***Сеилов Ш. Ж., Ахметова Ж. Ж., Зұлпыхар Ж.Е.***<br>*Евразийский национальный университет им. Л.Н. Гумилева, Нур-Султан, Казахстан*<br>**МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ПЕРЕВОДА ТЕРМИНОВ НА КАЗАХСКИЙ ЯЗЫК И ИХ РАЗВИТИЕ В СФЕРЕ ЦИФРОВОЙ ЭКОНОМИКИ** | 310 |
| 2 | ***Илиуф Хаджи-Мурат Шаяхметович***<br>*Государственный университет им. Шакарима, Семей, Казахстан*<br>**О СОВЕРШЕНСТВОВАНИИ ТЕРМИНОЛОГИИ В КАЗАХСКОМ ЯЗЫКЕ** | 317 |
| 3 | ***Қожахмет А. Қ.***<br>*Л.Н.Гумилев атындағы Еуразия ұлттық университеті Нұр-Сұлтан, Қазақстан*<br>**ҒЫЛЫМИ МӘТІНДІ ОҚЫТУДЫҢ ТАНЫМДЫҚ СИПАТЫ** | 336 |

## ТҮРКІ ЖӘНЕ ШЕТ ТІЛДЕРІН ОҚЫТУДЫҢ ИНТЕЛЛЕКТУАЛДЫ ТЕХНОЛОГИЯЛАРЫ

## ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ОБУЧЕНИЯ ТЮРКСКИМ И ИНОСТРАННЫМ ЯЗЫКАМ

## INTELLIGENCE TECHNOLOGIES FOR LEARNING TURKIC AND FOREIGN LANGUAGES


1 ***Абишева Ж.М., Разахова Б.Ш.***
*Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Нұр-Сұлтан, Қазақстан*
**БІЛІМ БЕРУДЕГІ ГЕЙМИФИКАЦИЯ** 342

2 ***Бекбауова А.У., Ғилманова Н.Т.***
*Қ.Жұбанов атындағы Ақтөбе өңірлік университеті, Ақтөбе, Қазақстан*
**CLIL ТЕХНОЛОГИЯСЫН АЛГЕБРА ПӘНІН ОҚЫТУДА ҚОЛДАНУ** 352

3 ***Алханов А.А., Туребаева Р.Д., Маткаримов Б.Т.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті Нұр-Сұлтан, Қазақстан*
**АВТОМАТТАНДЫРЫЛҒАН БІР ДҰРЫС ЖАУАПТЫ ТАҢДАУ СҰРАҚТАРЫН ҚҰРУ ЖҮЙЕСІ ҮШІН БІЛІМ ҚОРЫНЫҢ МОДЕЛІН ЖАСАУ** 359


## COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH» СЕМИНАРЫНЫҢ МАТЕРИАЛДАРЫ

## МАТЕРИАЛЫ СЕМИНАРА «COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH»


**MATERIALS OF THE SEMINAR «COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH** 369

# ПЛЕНАРЛЫҚ МӘЖІЛІС БАЯНДАМАЛАРЫ

# ДОКЛАДЫ ПЛЕНАРНОГО ЗАСЕДАНИЯ

# PLENARY MEETING

***Дихан Қамзабекұлы***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*филология ғылымдарының докторы*
*Басқарма мүшесі – әлеуметтік-мәдени даму жөніндегі проректор*
*ҚР ҰҒА академигі*
*Нұр-Сұлтан, Қазақстан*
*dikhan.kamzabek@enu.kz*

## ҚАНЫШ – АЛАШ ҒАЛЫМЫ

2009 жылы Л.Н. Гумилев атындағы Еуразия ұлттық университетінің баспасынан «Қаныш Сәтбаев. Алгебра. Том қаласы, 1924 жыл» атты толымды ғылыми-мәдени мұра 150 таралыммен ғана басылып шықты.

Кітаптың маңдайына «Алаш қайраткерлерінің «Оян, қазақ!» тұжырымдамасына - 100 жыл, Академик Қ.И.Сәтбаевтың туғанына – 110 жыл, Қазақ Білімпаздарының тұңғыш съезіне – 85 жыл» деп жаздық. Осы маңызды даталардың ішінде бүгінгі оқырманға түсініктісі - «110» жыл ғана. Бірақ тарихты жақсы білетін азаматтар көрсетілген деректердің ұлт тарихына және жаңа замандағы қазақ ғылымы мен мәдениетінің мақтанышы саналатын академик Қаныш Имантайұлы Сәтбаевқа тікелей қатысы барын аңғарады.

Ұлтымыздың азаматтық тарихының бір белесі – Алаш қозғалысы. Әрине, оның бастауы XIX ғасы мен XX ғасырдың түйіскен тұсы. Бұл ретте осы қозғалыстың манифесіне баланған Міржақып Дулатұлының «Оян, қазақ!» жинағы – азат Қазақстан үшін аса құнды мұра. Өйткені, аталған еңбекте ел зиялыларының қоғамдық дамуға берген бағасы және тығырықтан шығудың жолы жүйелі түрде көрсетілген. Соның ішінде елдік жауапкершілік пен ғылым-білім бірінші кезекке қойылған. Мектеп жасындағы Қаныш осы рухта өсіп-жетілгені әмбеге аян.

Болашақ академикке баға берген халыққа аса танымал ақын, діндар-ағартушы Мәшһүр Жүсіп Көпейұлы 20-жылдары былай депті: «Осы күнгі жастарда Қаныш Сәтбаев – адамның жорғасы. Тірі болса,

тірі жүрсе, бақты, талайлы болатын жігіт» («Көп томдық шығармалары», 13-том, 22-бет). Ал осы сөзден кейін Мәшһүрді, халық тілімен айтқанда, әулие емес деп көріңіз!

Қазақтың тағдырын шешкен төңкеріс тұсында Қаныш бар болғаны 20-жаста екен. Оның ағасы Әбікей, басқа да сыйлас-пікірлес аға буыны Алаш қозғалысының жуан ортасында жүрді. Ендеше жігіт Қаныш та олардан бөлек кеткен жоқ.

Осы «Алгебра» жазылатын шамада, 1924 жылы 12-17 маусым күндері Орынборда Қазақ Білімпаздарының тұңғыш съезі (орысша «Первый съезд ученых-казахов») өтті. А.Байтұрсынұлы, Ә.Бөкейхан, Х.Досмұхамедұлы т.б. ұйытқы болған осы жиында оқулық жазу мен ғылымды қалыптастыру мәселесі нақты қарастырылды.

1924 жылы 23 мамырда “Еңбекшіл қазақ” газеті осы жиынның қарайтын мәселесін жариялады. Олар: “1. Қазақ емлесін бірөңкейлеу. 2. Қаріп жағдайын қарастыру. 3. Халық әдебиетінің халін һәм оқылатын пәндерді жоспарлау. 5. Оқу һәм білім кітаптарын көбейту шарасын қарастыру. 6. Қазақ пән сөздері бір болу мәселесі”.

Съезд Орынборда ағарту қызметкерлері ұйымының үйінде өтеді. Жиынды Қазақстан халық ағарту комиссары Н.Зәлиұлы ашты. Оның орынбасары М.Жолдыбайұлы съезге келген кісілерді таныстырады. Олар: Мәскеудегі Күншығыс баспасөз тарататын кіндік ұйымнан – Әлихан Бөкейханұлы мен Нәзір Төреқұлұлы; Бұхарадағы қазақтар атынан – Мырза Наурызбайұлы; Түркістаннан – Халел Досмұхамедұлы мен Ишанғали Арабайұлы; Қазақстанның аймақтық партия комитетінінен – Аспандияр Кенжеұлы; Жалпыресейлік кәсіпшілер кеңестер ұйымының қазақ аймағы бөлімінен – Мұхтар Саматаұлы; Орынбордағы қазақ институтынан (КИНО) – Мұхтар Мырзаұлы; Қазақстан халық ағарту комиссариатынан – Ахмет Байтұрсынұлы, Елдес Омарұлы, Нұртаза Ералыұлы, Нұғман Зәлиұлы, Молдағали Жолдыбайұлы; Қостанай гүбернелік оқу бөлімінен – Ерғали Алдоңғарұлы; Семей гүбернелік оқу бөлімінен – Мәннан Тұрғанбайұлы; Орал гүбернелік оқу бөлімінен – Нығмет Шағиұлы; Бөкей гүбернелік оқу бөлімінен – Рүстем Ағыбайұлы; Қазақ аймағын зерттеп, ғылым жиятын қауымнан – Міржақып Дулатұлы; Орынбордағы газет-журнал басқармаларынан – Рақым Сүгіұлы.

Мұнан соң съезд басқармасы сайланды. Олар: Ә.Бөкейханұлы, Н.Зәлиұлы, А.Кенжеұлы, М.Наурызбайұлы, И.Арабайұлы. Басшылар съездің хатшылығына М.Тұрғанбайұлы мен Е.Алдоңғарұлын сайлайды.

Қазақ Білімпаздар съезіі тезге салған күрделі мәселенің ауқымдысы – қазақ тіліндегі пән сөздерінің (терминдердің) жайы болды. Қазіргі таңда қазақ тілі грамматикасында орныққан заңдардың

көбісіне дерлігі білімпаздар тобында сөз етілді. Пән сөздері бойынша баяндама жасаған – Е.Омарұлы. Баяндамада қазақ тіліне тән дыбыстардың үндестік жүйесі, сөзжасамдық қуаты сөз болып, өмірдің сан-саласындағы өзгерістерге сәйкес төл тілімізге еніп жатқан жат сөздерді қалай етене ету, қалай пайдалану хақындағы пайымдар мен ұсыныстар алға тартылды. Ғалым жат сөздердің тігісін жатқызып, сөздік қорымызға қабылдаудың 9 тәсілін көрсетіп берді. Съезд оны қабылдады.Сондағы негізгі шешім мынау: «Пән сөздері ең әуелі қазақ тілінен ізделінеді. Бүтін дүниедегі қауымға бірдей сөдердің қазақшасы табылса, ол пән сөзі әлгі көп ел алған сөзбен жарыстырыла қолданылсын. Қайсысы оңды болады – оны өмір көрсетеді. Қазақ тілінен табылмаған пән сөздері түркі тілінен ізделінсін. Табылса, жатырқаусыз қазақ тілі заңына бағындырылып алынсын. Ал, пән сөзі түркі жұртында да ұшыраспаса, Шығыс немесе Батыс тілдерінен алынып, ана тіліміздің заңына бас идіріп, бұрылып пайдаланылсын».

Бұл шешім 1926 жылы Бакуде болған Түрік Білімпаздары съезінде қолдау тапқанын жақсы білеміз. Мұны баспасөзде Қаныш Сәтбаев та жазды («Қазақ тілі» газеті, 1926 жыл, 19 мамыр). Әйтсе де, осы шешімнен «пантүркілік» астар, «түркілік томаға-тұйықтық» іздеушілер болды...

Қазақ Білімпаздары пән сөздерін қадағалайтын комиссия (қазіргіше терминком) құру қажет деп, қаулы қабылдады. Ол бойынша комисияда әр салаға қатысты секциялар болу керек. Әрбір секция өз саласында тапқан сөздерін комиссияның ғылым кеңесі алдына қоюға міндетті. Секциялар пән сөздерімен жұртты мәшһүр етуге, ол жөнінде ғалымдардың пікіріне құлақ қоюға ынталы болуға тиіс.

Пән сөздері туралы шараларды көрсетіп бергеннен кейінгі Білімпаздар тобының қараған ісі, алдын-ала хабарланғандай, “Бастауыш мектептерде оқылатын пәндерді жоспарлау” делінген тақырыптан асып кетіп, жалпы оқу-ағарту жүйесінің құрылымын қамтыды. Бұл туралы Т.Шонанұлы баяндама жасады.

Съезд уәкілдері оқу-ағарту жүйесін тексеретін, біліми-әдістемелік оқулықтарды сараптайтын, жоғары мектеп пен бастауыш оқу орындарының мақсатнамаларын електен өткізетін комиссия құрылсын деп ұсыныс ендірді. Жасақталған комиссия осы отырыста ағартуға мамандандырылған оқу орындарының: институттар мен техникумдердың, негізгі және пысықтауыш курстардың, 7 жылдық пен 4 жылдық мектептердің жүйесі хақында қаулылар шығарады. Онда әрбір ағарту орнының оқу мерзімі, негізгі пән мен қосалқы пән қай уақытта жүргізілуі керектігі, мамандарды қалай бөлу айтылған. Бұл қаулының кейбір баптары әлі де зәрулігін жойған жоқ. Мысалы,

“Барлық техникумдардың прогрaмы бірөңкей болмауы керек”, “Институттар үлкен өлкелерде болады”, “Оқу ана тілінде болады”, “Амалсыз кездерде жоғары бөлімдерде оқуды орыс тілінде жүргізуге бөгет болмасын”, “Отырықшы халықтың мектебі бір жерде тұрады”, “Көшпелі елдің мектебі өздерімен бірге көшіп жүреді”, “Көшпелі елдің, иа қыстауы бытыраңқы, иа там-үйі жоқ елдерде оқу мезгілі жергілікті жағдайға қарай болады”.

Іргелі ел болуға ниеттенген жұрт әліппесін, бастауыш мектебін түзейтінін ұққан топқа келген уәкілдер бастауыш мектеп оқуын жоспарлауға жіті көңіл аударады. “Өтілетін барлық пәндердің: шама тілінің (математика), ана тілінің, тұрмыс жүйесі ғылымының (тарих пен қоғамтану), табиғаттанудың, жертанудың, сүгірет салудың жоспары келеге салынып, “мазмұны тұрмысқа жуық, баларға ұғымды, ауыр емес етіп” бекітілді. Жоспар талқыға түскенде, ана тілі бағдарламасы қаз-қалпында қалдырылып, қалғандарына түзетулер ендірілді.

Сиездің оқу-білім кітаптарын көбейту үшін түрлі шараларды белгілеуге арналған отырысы да Білімпаздар тобының маңызын арттырды. Мәжілісте баяндама жасаған – М.Мырзаұлы. Қазақ халыққа білім беру институтының (КИНО) басшысы болған баяндамашы ана тіліндегі білім кітаптарының тілі, мазмұны, құрылымы жағынан әлі де болса кемшін екеніне тоқталып, алаш балаларының рухани жетілуіне жәрдемдесетін оқулықтар мен нұсқаулықтарды, ғылымға негізделген жинақтарды даярлау жөнінде төмендегі шараларды үкіметке ұсынды:

“а) Әдебиет жұмысы (әңгіме оқу кітаптары жөнінде болып отыр – Д.Қ.) төтенше жұмыс ретінде табылсын.

ә) Қазақстан орталығында жігерлі, қолдарынан іс келетін азаматтардан Кеңес сайлансын. Ол Кеңестің міндеттері мыналар болсын: 1) Әдебиет жұмысының ретін, бағытын тексеріп, жазушыларға басшылық қылып, жөн көрсету; 2) Баспасөз мекемелерінің жұмыстарын ынтымақтастырып, кем істерін жоюға шаралар қолдану; 3) Жазушылардың басын қосып отыру және соларды қамсыздандыру шаралары; 4) Кеңестің осы айтқан міндеттерін орындай аларлық арнаулы қаржысы болуға тиіс. Жетпеген қаржы өкіметтен сұралсын; 5) Кітаптардың кезектік ретімен, алдымен мектеп құралдары түгенделсін деген нұсқау берілсін. Содан кейін негізгі пәндерден мағлұмат берерлік жәрдемші кітаптар аударылсын; 6) Жалпы ел үшін денсаулық, тазалық, шаруа, тұрмыс жайларынан, қазақ тарихынан кітаптар аударылатын болсын. Орысшадан тек қазаққа үйлесетін реттеріндегі өзгерістермен ғана аудартылсын; 7) Жалғыз бұл шаралармен ғана тынбай, жәрдемші ретінде білім әдебиетін көбейту туралы түрлі қауымдар, серіктер

ашылып отыратын болсын. Бұларға үкіметтен, партиядан түрлі жеңілдіктер, жәрдемдер берілсін".

Әрине, біз бүгін тарихымыз бен тағдырымыз үшін Қазақ Білімпаздарының тұңғыш съезінің орнын дұрыс бағамдауымыз қажет. Мұнда дәл қазір де көкейкесті рухани-мәдени, ғылыми-біліми жайттардың шешілу жолы айтылды. Ал, енді кейін оның дұрыс және жүйелі орындалмауының себебі – ұлт зиялыларының репрессиялануы байланысты.

Қазақстанда ұлттық ғылым тілі қазір қалыптасу кезеңін бастап кешіріп отыр. Ал, оның тууы XX ғасырдың 10-30 жылдар аралығын қамтиды. Бізде ғылым тілі алғашқы оқулықтардың шығуымен, әр сала бойынша танымдық, әліппе іспеттес еңбектердің жариялануымен қабат пайда болды.

Әдеби тіл мен ғылым тілі бір емес. Соңғысы әбден сарапталған, нақтылықты, дәлдікті қажет ететін, тұжырымдалған аталым-терминдерге иек артатын тіл. Сондықтан бұл тілдің туу процесі оңайлыққа түспегені айқын.

20-жылдардың басында А.Байтұрсынұлы жетекшілік ететін Халық ағарту комиссариаты сол шақтың оқығандарына ғылымның әр саласынан оқулық жазу мен орыс тіліндегі ғылыми әдебиеттерді аудару жөнінде тапсырыс берді. Нәтижесінде Х.Досмұхамедұлы – анатомия, зоология, М.Дулатұлы, С.Қожанұлы, Ә.Ермекұлы, Қ.Сәтбайұлы – математика, алгебра, Т.Шонанұлы – география, тарих, М.Жұмабайұлы – педагогика, Ж.Аймауытұлы – психология, Қ.Кемеңгерұлы – химия, тіл, тарих, Ж.Күдеріұлы – тіл, әдебиет, мәдениет, Ә.Бөкейханұлы – астрономия, география, М.Әуезұлы – әдебиет, С.Садуақасұлы – театртану, А.Байтұрсынұлы, Е.Омарұлы, Т.Шонанұлы, Н.Төреқұлұлы – тіл саласынан еңбектер жазды (ол кезде фамилия соңына «ұлы» қосымшасы жалғанған және ол жоғарыда аталған съезде бекітілген). Бұл тізім, әрине, толық емес. 30-жылдардың соңында жоғарыда аты аталған авторлардың бәрі дерлік репрессияланды. "Халық жауы" атанған олардың араб харпіндегі еңбектері қайта жарық көрмеді. Бірақ олар қалыптастырған, пайдаланған терминдер ғылыми тілімізде орныға бастады. Сонымен бірге сол зиялылар шешуге ұмтылған жетпей кеткен проблемалар осы күнге дейін күн тәртібінде тұр. Өйткені қазақ ғылыми тілі проблемасын әр саланың мамандарын қатыстыра отырып қарастырған басқосу 1924 жылдан кейін өткізілген жоқ. Кеңес тұсында ғылыми тілдің көп мәселесі бір жақты шешілді. Алаш тағылымына жан-жақты тыйым салынғанда, М.Әуезов, Қ.Сәтбаев, Ә.Марғұлан үшін аса күрделі кезең басталғаны рас. Ғылым академиясының Тіл білімі институты ғылым тілінің практикалық қолданысы проблемасымен

шұғылдана алмады. Сондықтан қазір "бізде ғылым тілін жүйелендірдік" деп айту әлі ерте. Дұрысы, біз сол үшін күресіп жатырмыз. Сондай-ақ Қаныш Сәтбаевтың «Алгебрасы» сынды мұралар – осы жолда бізге үлкен таяныш.

Ізденуші-әдіскер жас Қаныш бұл оқулықты Том қаласында студент болып жүрген кезде жазды (Байқайсыздар ма, ол бүгінгілер секілді «Томск» демейді, «Том» дейді. Айтқандайын, қазір кейбір білгіштердің Омбыға ұқсатып, «Томбы» деп жүргені де ұшырасады. Бірақ қазақ тілінің заңы мен тарихы «Том»-ды қолдайды).

Жұмыста кездесетін әлпет (выражение), қоспа (слагаемое), өрнек (формула), шама (величина), қосынды (сумма), үдеме дәреже (восходящая степень), азба дәреже (нисходящая степень), пропорция (тең ара), түйін (теорема), берен (функция), белдік (ось), жік (грань) дәуірлеу (прогрессия), өріс (радиус), тек (элемент), шоғыр (комплекс), өре (диаметер), нағыз (натуральный), жасалым (устройство) т.б. аталымдар А.Байтұрсынұлы қалыптастырған ұлттық ғылыми терминдердің өресінен табылады. Халықтың ғасырлар бойы тірнектеп жиып, өрнектеп қолданған қара сөзін «шекпен кигізіп, өзіне қайтарған» мұндай талап – дамуға бағыт алған әлемнің барлық тіліне де ортақ қасиет. Осы жолда Ахмет көшбасшы болса, Қаныштар мұны салаландарды. Жас ғалымның Алашқа, елге адалдығы осыдан көрінеді.

1924 жылы жазылған «Алгебраның» бүгінгі кирил қарпіндегі нұсқасын оқыған адам рим және араб цифрларынын кейінгі «ыншы», «інші» қосымшаларға да таңдануы мүмкін. Қазіргі жазуымызда «-» (дефис) осы қосымшалардың орнына жүреді және рим цифрынан кейін сызықша қойылмайды. Сондықтан жазуымыздың тарихын білу үшін де біз әлгі қосымшаларды сақтадық.

Бұл 85 жылдық тарихы бар кітапты жариялағанда академик Мұхтарбай Өтелбаев бастаған ЕҰУ ғалымдары (латын қарпіндегі нұсқасынан даярлағандар: Р.Ойнарұлы, Ә.Түнғатаров, О.Қашқынбаев, А.Ибатов, Қ.Мырзатаева, Л.Жапсарбаева, М.Алдай, А.Абылаева, З.Әбдіқалықова; араб қарпіндегі нұсқасымен салыстырған: Д.Мықтыбек), Кәкімбек Салықов ұйыстырған қаныштанушы қайраткерлер оның мәдени-тарихи құндылығын ескерді. Осы еңбек бүгінгі «нақты ғылымдар қазақшаланбайды!» дейтін кері кеткен пікірдің күлін көкке ұшырады деп есептейміз.

Ғылым мен білімдегі Алаш рухы, зиялыларымыздың сөзімен айтқанда, «күн сөнгенше сөнбейтіні» ақиқат. Осыны зерделеп, дамыту жолында Алла бәрімізге күш-қуат, парасат бергей.

***Шарипбай Алтынбек Амирович***
*Евразийский национальный университет им. Л. Н. Гумилева*
*д.т.н, профессор кафедры*
*Технологии искусственного интеллекта*
*Нур-Султан, Казахстан*
*sharalt@mail.ru*

# ПРОБЛЕМЫ СОЗДАНИЯ АЛФАВИТА КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ ЛАТИНСКОЙ ГРАФИКИ

**Аннотация:** В статье даются сведения о письменности казахского языка и её роли, доказывается некорректность действующего алфавита на основе кириллицы и обосновывается необходимость проведения реформы казахской письменности, показываются ошибки и противоречия предложенного на утверждение шестого варианта алфавита на основе латиницы, уточняется звуковая система казахского языка и предлагаются два варианта алфавита на основе латинской графике, первый на основе алфавита турецкого языка с диакритическими знаками, а второй на основе алфавита английского языка с диграфами.

**Ключевые слова:** аллофон, диакритика, дираф, орфография, фонема, которые определяются:

*Аллофон* – смыслонеразличительная и обозначаемая знаком, который обозначает в алфавите созвучную с ней фонему, минимальная фонетическая единица языка.

*Диакритика* – знак, обозначающий созвучные фонемы, передающий фонетические свойства звука, ударение и т. п. и помещаемый над, под или рядом с буквой.

*Фонема* – смыслоразличительная и обозначаемая отдельным знаком в алфавите минимальная фонетическая единица языка.

*Орфография* – раздел лингвистики, состоящий систему правил для единообразной передачи слов и грамматических форм речи на письме.

***Sharipbay Altynbek Amiruly***
*L.N.Gumilyov Eurasian National University*
*Doctor of Technical Sciences*
*Professor of the Artificial Intelligence Technologies Department*
*Nur-Sultan, Kazakhstan*
*sharalt@mail.ru*

# PROBLEMS OF CREATING THE ALPHABET OF THE KAZAKH LANGUAGE BASED ON THE LATIN GRAPHICS

**Abstracts:** In the article provides information about the writing of the Kazakh language and its role, proves the incorrectness of the current alphabet based on the Cyrillic alphabet and substantiates the need for a reform of the Kazakh writing system, shows errors and contradictions of the sixth version of the alphabet based on the Latin alphabet proposed for approval. Are clarified the sound system of the Kazakh language and suggestes two versions of the alphabet based on the Latin script: the first based on the Turkish alphabet with diacritics, the second based on the English alphabet with digraphs.

**Keywords:** Phoneme, allophone, diacritics, digraph, spelling.

## 1. Введение

Казахский язык является одним из тюркских языков, имеющий древнюю историю и формировавшийся на протяжении многих веков как и языки родственных тюркских племён, проживавших на территории современного Казахстана [1, 2]. Поэтому он, как и любой древний язык, нуждается в исследовании его истории, естественных лингвистических особенностей и возможностей.

Казахский язык является национальным языком казахов, проживающих в Республике Казахстан и во многих странах мира [3]. Поэтому его, как и любой национальный язык, служащий основным средством общения носителей языка, надо изучать и развивать для использования его в ежедневной жизни казахского народа.

Казахский язык является государственным языком Республики Казахстан. Поэтому он, как любой государственный язык, должен служить средством общения народов, населяющих территорию Казахстана, поддерживаться государством и использоваться во всех сферах его деятельности, в том числе и во внешних отношениях с другими странами.

Казахстан, интегрируясь в мировое сообщество, должен обеспечивать достаточный (по отношению к мировому уровню) уровень развития государственного языка.

В настоящее время казахи, живущие в разных странах, используют разные алфавиты, основанные на арабице, латинице и кириллице, что мешает их письменному общению и формированию своего единого культурного и информационного пространства.

В эпоху цифровизации во многих странах обсуждается проблема использования единого алфавита в глобальном информационном пространстве, в качестве которого предлагается принять английский алфавит, чтобы сократить затраты на поиск и обработку информационных ресурсов, объем которых растет очень быстро. Кроме того, все производимые в мире компьютеры и другие электронные устройства поддерживают базовый латинский алфавит. Если же алфавит некоторого языка имеет хотя бы одну букву, отличную от буквы базового латинского алфавита, то, чтобы работать на этих аппаратах с алфавитом этого языка, нужно дополнительно создавать шрифты, драйвера, программы сортировки и поиска данных, которые требуют немалых интеллектуальных, временных, трудовых и финансовых затрат.

## 2. Казахская письменность

Казахская письменность берет свое начало с древнетюркских рунических письменностей VI–X вв. н.э. Один из вариантов древнетюркских алфавитов, знаками которого высечена орхоно-енисейская надпись на надгробном памятнике Кюль-Тегина, впервые был расшифрован (определены значения знаков алфавита) в 1893 году шведским ученым Вильгельмом Томсеном [4].

Близкие к орхоно-енисейским надписям рунические надписи V–III вв. до н.э. были обнаружены в Казахстане в 50 км от г. Алматы, в Иссыкском кургане, в котором был захоронен «Золотой человек». Эти надписи называются «Иссыкское письмо», состоят из 26 знаков, внешне напоминающих орхоно-енисейскую письменность. Поэтому некоторые исследователи признают вероятность того, что потомком Иссыкского письма является более позднее классическое орхоно-енисейское письмо [5].

В связи с распространением ислама в письменном языке тюркского мира с IX по XX век н.э. на протяжении многих веков использовался арабский алфавит, так как на нём была написана священная книга мусульман – Коран. На этом же алфавите многие ученые, педагоги и

поэты более тысячи лет создавали свои бессмертные произведения, вошедшие в мировую сокровищницу науки, образования и культуры.

Однако арабский алфавит, созданный для семитских языков и хорошо приспособленный к требованиям арабского языка, не до конца отражал богатую фонетическую систему тюркских языков: ряд знаков в нем не был нужен тюркским языкам, и, наоборот, немало звуков, имеющиеся в тюркских языках, не находило в нем отражения. В результате появилась необходимость внесения изменений в арабский алфавит, используемый тюркскими языками.

Новый метод тюркского письма на основе арабской графики, называемый «Усуль аль-джадид», был впервые придуман выдающимся крымско-татарским просветителем Исмаилом Гаспралы. Суть его метода заключается в фонетической обработке арабского алфавита, которая ставила в соответствие звуки и буквы, в отличие от старого метода «Усуль аль-кадим», предлагавшего слоговое изучение языка, когда отдельные буквы сливались в слоги, а из слогов потом составлялись слова. Новый метод позволил минимизировать недостатки базового арабского произношения и в 2-3 раза сократить срок обучения грамоте [6].

В XX веке казахский язык поменял свой алфавит 3 раза. В первый раз в 1912 году с помощью метода «Усуль аль-джадид» основоположник теории казахского языка Ахмет Байтурсынов перевел казахскую письменность на новый алфавит на основе арабской графики. Для этого он систематизировал и уточнил звуковую систему казахского языка из 28 звуков, из которых 5 гласные и 19 согласные фонемы (*смыслоразличительная и обозначаемая отдельным знаком в алфавите минимальная фонетическая единица языка*), при этом 4 мягкие гласные считались *аллофонами* (*смыслонеразличительная минимальная фонетическая единица языка, обозначаемая знаком, который обозначает в алфавите созвучную с ней фонему*), на самом деле они являются *фонемами*. Затем он исключил все арабские буквы, обозначающие неказахские звуки и разработал новый алфавит казахского языка, который содержал 24 буквы на основе арабской графики и 1 специальный знак (апостроф). При этом для обозначения мягких гласных использовались диграфы, в которых первыми парами были буквы, обозначающие твердые согласные, а в каждом диграфе второй парой был апостроф. Этим алфавитом в нашей стране пользовались до 1929 года, а казахи, живущие в других странах, в частности, в Китае, Афганистане и Иране, пользуются им и по сей день, добавив еще 3 буквы для обозначения согласных фонем (*в*), (*ф*) и (*х*).

Во второй раз в 1929 году вместо алфавита на основе арабицы был принят алфавит на основе латиницы, который имел 31 букв.

В третий раз в 1940 году вместо алфавита на основе латиницы был принят 42 буквенный алфавит на основе кириллицы: из них 31 обозначают казахских звуков, которые были обозначены буквами алфавита основе латиницы; 9 используются для обозначения не свойственных казахскому языку звуков русского языка; 2 предназначены для обозначения мягкости и твердости согласных звуков. Этот алфавит действует до настоящего времени.

### 3. Необходимость реформы письменности казахского языка

Как уже сказано в пункте 2 в действующем алфавите казахского языка на основе кириллицы есть 42 букв, из них 15 (Аа, Әә, Ее, Ёё, Ии, Оо,Өө, Уу, Ұұ, Үү, Ыы, Іі, Ээ, Юю, Яя) обозначает гласные фонемы, 25 (Бб, Вв, Гг, Ғғ, Дд, Жж, Зз, Йй, Кк, Ққ, Лл, Мм, Нн, Ңң, Пп, Рр, Сс, Тт, Фф, Хх, Һh,Цц, Чч, Шш, Щщ) – согласные фонемы, а 2 буквы Ьь – мягкий знак и Ъъ – твердый знак.

Следует отметить, что 9 букв Ьь, Ъъ, Щщ, Ёё, Ээ, Юю, Яя, Ии, Уу не имеют никакого отношения к казахскому языку. В русском языке буквы Ьь и Ъъ используются для обозначения мягкости и твердости согласных звуков. В казахском языке мягкими и твердыми могут быть только гласные звуки. В исконно казахских словах нет звуков, которые могут быть обозначены буквами Щщ, Ёё, Ээ, Юю, Яя. Более того, при использовании в казахской письменности букв Ии, Уу, которые обозначают гласные звуки русского языка, нарушаются:

1) *закон сингармонизма*: в казахских корневых словах гласные звуки должны чередоваться с согласными и при этом в записи должны использоваться только мягкие, либо только твердые гласные, но при использовании указанных букв это условие нарушается. Например, в следующих словаах «қ**иа** **(**утес)», «иә (да)», «**ауа** (воздух)», «**әуе** (небо)», «**уа**қыт (время)», «**уә**де (обещание)», «к**ие** (святыня)», «қ**иуа** (вдали)», «қ**ия** (наискось)». «с**аяхат** (путешествие)», «с**үю** (любить)», «с**ұю** (разжижаться)» буквы гласных звуков встречаются подряд;

2) *орфографические правила*:

- если основа слова (корень или корень + суффикс) заканчивается на твердый (или мягкий) гласный звук, то к нему добаляется притягательное окончание в третьем лице «сы» (или «сі»). Например, «ан**а**+сы (мать)», «әж**е**+сі (бабушка)»;

- если твердая (или мягкая) основа слова заканчивается на согласный звук, то к нему добаляется притягательное окончание в

третьем лице «ы» (или «і»). Например, «отан+ы (родина)», «ел+і (страна)».

Но при использовании в записи основы слова букв Ии, Уу, которые обозначают гласные звуки русского языка, эти правила нарушаются. Например, реально пишутся «би+**і** (танец)», «ми+**ы** (мозг)», «ту+**ы** (флаг)» и «гу+**і** (гул)», вместо того, чтобы писать по правилу «би+сі», «ми+сы», «ту+сы» и «гу+сі», которые в казахском языке не имеют никакого смысла.

Указанные примеры доказывают ошибочность действующего алфавита казахского языка на основе кириллицы. Поэтому требуется проведение реформы письменности казахского языка.

## 4. Анализ предложенного на утверждение алфавита на основе латиницы

С цель реформы письменности казахского языка уже предложены 6 (шесть) версии алфавита казахского языка на основе латинской графики. Все они содержат ошибки и противоречия. Ниже анализтруется только шестой вариант.

В таблице 1 представлена шестая версия алфавита казахского языка на основе латиницы, которая размещена на сайте https://www.qazlatyn.kz/alphabet.

Таблица 1. Шестая версия алфавита казахскогоя языка на основе латиницы.

| № | Baspa túri | Árip ataýy | | № | Baspa túri | Árip ataýy |
|---|---|---|---|---|---|---|
| 1 | A a | [a] | | 17 | Ń ń | [yń] |
| 2 | Á á | [á] | | 18 | O o | [o] |
| 3 | B b | [by] | | 19 | Ó ó | [ó] |
| 4 | D d | [dy] | | 20 | P p | [py] |
| 5 | E e | [e] | | 21 | Q q | [qy] |
| 6 | F f | [fy] | | 22 | R r | [yr] |
| 7 | G g | [gi] | | 23 | S s | [sy] |
| 8 | Ǵ ǵ | [ǵy] | | 24 | T t | [ty] |
| 9 | H h | [hy] | | 25 | U u | [u] |
| 10 | İ i | [i] | | 26 | Ú ú | [ú] |
| 11 | I ı | [ıı] | | 27 | V v | [vy] |
| 12 | J j | [jy] | | 28 | Y y | [y] |
| 13 | K k | [ki] | | 29 | Ý ý | [uý] |
| 14 | L l | [yl] | | 30 | Z z | [zy] |
| 15 | M m | [my] | | 31 | Sh sh | [Shy] |
| 16 | N n | [ny] | | 32 | Ch ch | [chy] |

Здесь названия столбцов переводятся так «Baspa túri – Печатный вид», «Árip ataýy – Название буквы».

**Замечание**: Далее строчными кириллическими буквами в круглой скобке и буквами международного фонетического алфавита (МФА) в квадратной скобке будем обозначать соответствующие фонемы казахского языка.

Представленный в таблице 1 алфавит не понятен, так как в любом алфавите буквы обозначают не названия букв, а звуков языка, представляя их сначала символами МФА [7, 8] в квадратных скобках, как это принято во всех языках мира, или просто строчными буквами действующего алфавита, заключая их в круглые скобки.

В этом алфавите нет системности: некоторые созвучные твердые и мягкие гласные фонемы обозначены разными латинскими буквами, например, твердая фонема (ы) - [ɯ] обозначена буквой Yy, а созвучная с ней мягкая фонема (і) - [ɪ] буквой Iı. Звуки (ы) и (і) должны обозначаться одной буквы и для различичия для мягкого добавляется диакритический знак, например, Aa–(a)–[a], Áá–(ə)–[æ], Oo–(o)–[ɔ], Óó–(ө)–[ө], Uu–(ұ)–[ʊ], Úú–(ү)–[ʏ]. Использование знака *ударения* ухудшает читабельность текста.

Указанный алфавит неоднородный, так как для обозначения некоторых фонем используются и дикритические знаки и диграфы. Необоснованно использованы диграфы (ш) – [ʃ] – Sh sh, (ч) – [tʃ] – Ch ch. При этом в качестве пары диграфа использована латинская буква Cc, которая не включена в состав алфавита.

В пункте 3 доказано, что в казахском языке нет гласных звуков (и) и (у). Поэтому их обозначение буквами Iı и Ýý необоснованы, тем более буква Yy использована для обозначения несозвучного с ним твердого гласного звука (ы).

Главная ошибка в этом алфавите состоит в том, что в нем обоначены заимствованные из русского языка гласные звуки (и) – [i] и (у) – [u], которые порождают противоречия, указанные в пункте 3.

Учитывая вышесказанные замечания ниже в пункте 5 уточняется звуковая система казахского языка, а в пункте 6 предлагается новый вариант алфавит казахского на латинице, основанный на его уточненную звуковую систему.

## 5. Уточнение звуковой системы казахского языка

### 5.1. Гласные звуки

Известно, что в любом естественном языке устная речь первична, а письменная речь вторична. Поэтому для построения алфавита надо сначало уточнить звуковую систему казахского языка, и только потом подбирать буквы (знаки) для обозначения фонем, В некоторых случаях могут быть использованы и аллофоны.

Однако на текущий момент нет общепринятого мнения относительно фонетики казахского языка, до сих пор нет стандарта фонетики, где должны быть уточнены казахские звуки и их классификация. Подобного состояния, как в казахского языке, нет ни в одном другом государственном языке мира. Это связано с тем, что имеющееся учение о фонетике казахского языка унаследовано из русского языка и не отражает точную характеристику казахских звуков, что было показано в работах [9-11]. Поэтому не понятно, что делать с русскими гласными, которые были внедрены в казахский язык без учета его фонетических закономерностей. Один из способов решения этой проблемы - исключение этих звуков из звуковой системы казахского языка, как в азербайджанском языке, и разработка соответствующих орфографических правил.

В казахском языке гласные фонемы, в отличие от согласных фонем, играют важную роль в письменности, так как от их классификационных признаков зависит выполнимость закона сингармонизма и орфографических правил по присоединеию аффиксов (суффиксов и окончаний).

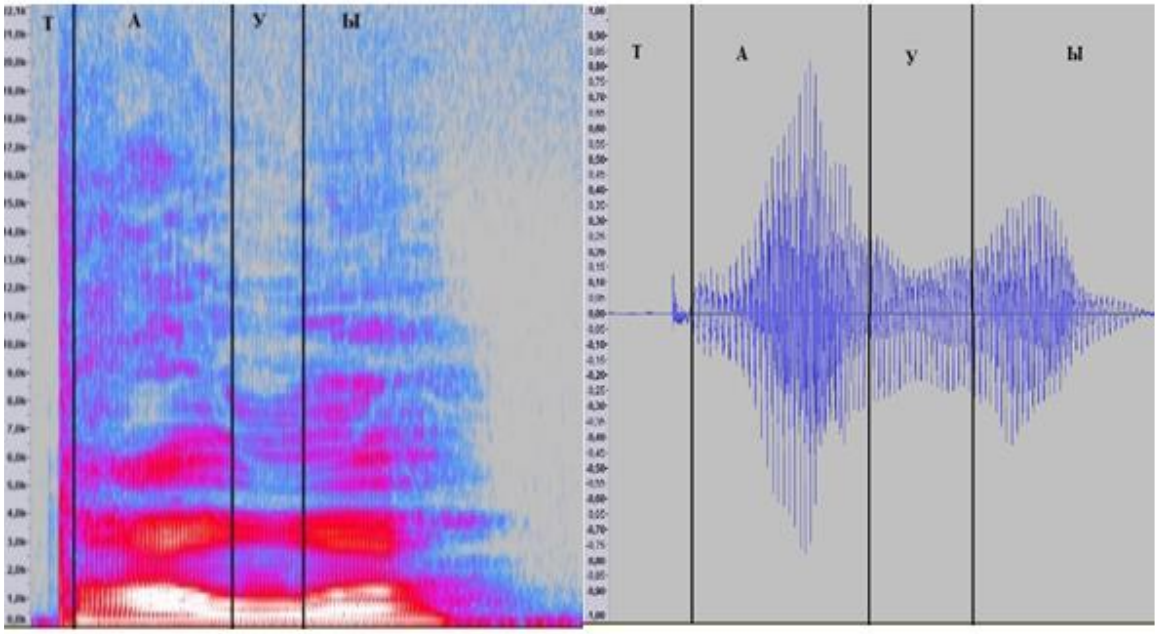
В действующем алфавите казахского языка обознаяены 11 гласных звуков (а), (ә), (е), (и), (о), (ө), (ұ), (ү), (у), (ы), (і). Из них звуки (и), (у) заимствованы из руского языка в 1940 году во время перевода казахской письменности на кириллицу. Чтобы проверить обоснованность их внедрения в казахский язык мы провели *перцептивный анализ* казахского языка с целью слуховой идентификации его звуковых единиц. Для этого составили фонетический насыщенный тест (звуки, слоги, слова, фразы), которого продиктовали носители казахского языка из разных регионов, разного пола и возраста.

В результате выяснили, что в исконно казахской речи нет фонемы (и) и (у). Так вместо звука (и) идетифицированы дифтонги (ый) и (ій) в зависимости от твердости и мягкости звучания прочитанного продиктованного теста.

Вместо гланой фонемы (у) идентифицирована поугласная фонема, которая нет в русском языке, но имеется в английском языке и обочена

она латинской буквой Ww, а полугласную фонему будем обозначать как (w).

На рисунке 1 показан результат обнаружения полугласной фонемы (w).




А В

Рисунок 1. Фонетическая сегментация слова «тауы», где А – спектрограмма, В – волна.

На рисунке 1. А можно заметить полугласного звука (w) между гласными звуками (а) и (ы). Звук (w) имеет очень ограниченные препятствия в процессе артикуляции и поэтому он похож на гласные звуки по спектрограмме и звуковой волне, но он отличается от них отсутствием энергии в районе частот 11-15 КГц, изменением формант в районе 2 КГц и 5 КГц и амплитудой, существенно меньшей, чем амплитуды гласных звуков (а) и (ы), что видно на рисунке 1.Б.

На основании проведенного эксперимента звуки (и) и (у) не будут включены в состав казахской звуковой системы как фонема, они могут использованы как аллофон при произношении некоторых заимствованных из других языков терминов.

Таким образом, в звуковую систему казахского языка будут включены 9 гласных фонем (а) - [ɑ], (ә) - [æ], (е) - [е], (о) - [ɔ], (ө) - [ɵ], (ұ) - [ʊ], (ү) - [ʏ], (ы) - [ɯ], (і) - [ɪ], которые будут обозначены в алфавите на основе латинской графики отдельными буквами

В некоторых источниках по фонетике казахского языка [7-10] для артикуляционной классификации казахских гласных звуков выделяют 3 бинарных признака.

Для корректного и полного описания характеристик девяти объектов трех бинарных признаков недостаточно, поскольку каждый признак имеет по 2 значения и максимальное количество различимых признаков равно $2^3 = 8$. Поэтому в указанных источниках класификация казахских гласных звуков некорректна и не соответствуют классификации звуков в МФА [11, 12], которая имеет 4 бинарных признаков: положение языка, положение челюсти и положение губ с показанными в таблице 1, а также вертикальное положение языка с двумя значениями (*верхние, нижние*). В МФА термин «верхние» называется «close», а «нижние» – «open», подразумевая близость при подъеме языка к небу [13, 14].

В таблице 2 показаны четыре бинарных артикуляционных признака казахских гласных звуков.

Таблица 2. Четыре артикуляционные признаки казахских гласных звуков

| **Гласные** | | **Артикуляционные признаки** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Обозначение в кириллице | Обозначение в МФА | Положение языка | | | | Положение губ | | Положение челюсти | |
| | | Вертикальное | | Горизонтальное | | | | | |
| | | Нижние | Верхние | Заднеязычные | Переднеязычные | Неокругленные | Округленные | Открытые | Закрытые |
| (а) | [ɑ] | + | – | + | – | + | – | + | – |
| (ә) | [æ] | + | – | – | + | + | – | + | – |
| (ы) | [ɯ] | + | – | + | – | + | – | – | + |
| (і) | [ɪ] | – | + | + | – | – | + | – | + |
| (ұ) | [ʊ] | + | – | – | + | + | – | – | + |
| (ү) | [ʏ] | – | + | – | + | – | + | – | + |
| (о) | [ɔ] | – | + | + | – | – | + | + | – |
| (ө) | [ɵ] | – | + | – | + | – | + | + | – |
| (е) | [е] | – | + | – | + | + | – | + | – |

Теперь, используя Булеву алгебру [12], мы построим математическую модель системы гласных звуков [13, 14] в виде алгебраического выражения на основе использования значений 4-х артикуляционных бинарных признака, показанных в таблице 2.

Обозначим 4 артикуляционных признаков *Вертикальное положение языка* и *Горизонтальное положение*, *Положение губ*, *Положение челюсти, языка* казахских гласных звуков через логические переменные $x_1$, $x_2$, $x_3$ и $x_4$, соответственно. Эти переменные принимают только значения 1 для истины и 0 для лжи:

- если язык находится в верхнем положении, то $x_1$ =1, иначе $x_1$ = 0;
- если язык находится в заднем положении, то $x_2$ =1, иначе $x_2$ = 0;
- если губы находится в округленном положении, то $x_3$ =1, иначе $x_3$ = 0;
- если челюсть находится в открытом положении, то $x_4$ =1, иначе $x_4$ = 0.

На основании установленных значений для переменных можно построить Булеву модель [12] системы казахских гласных звуков, представленной в таблице 3.

Таблица 3. Булева модель системы казахских гласных

| **Обозначение в кириллице** | **Обозначение в МФА** | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| (а) | [ɑ] | 0 | 1 | 0 | 1 |
| (ә) | [æ] | 0 | 0 | 0 | 1 |
| (ы) | [ɯ] | 0 | 1 | 0 | 0 |
| (і) | [ɪ] | 0 | 0 | 0 | 0 |
| (ұ) | [ʊ] | 1 | 1 | 1 | 0 |
| (ү) | [ʏ] | 1 | 0 | 1 | 0 |
| (о) | [ɔ] | 1 | 1 | 1 | 1 |
| (ө) | [ɵ] | 1 | 0 | 1 | 1 |
| (е) | [е] | 1 | 0 | 0 | 1 |

Далее для каждого из 9 гласных звуков, выписывая признаки в виде их конъюнкции и объединяя эти конъюнкции, получим следующую дизъюнктивную нормальную форму из 9 членов:

$$\left\{\begin{array}{c}(\overline{x}_1 \wedge x_2 \wedge \overline{x}_3 \wedge x_4) \vee (\overline{x}_1 \wedge \overline{x}_2 \wedge \overline{x}_3 \wedge x_4) \vee (\overline{x}_1 \wedge x_2 \wedge \overline{x}_3 \wedge \overline{x}_4) \vee \\ (\overline{x}_1 \wedge \overline{x}_2 \wedge \overline{x}_3 \wedge \overline{x}_4) \vee (x_1 \wedge x_2 \wedge x_3 \wedge \overline{x}_4) \vee (x_1 \wedge \overline{x}_2 \wedge x_3 \wedge \overline{x}_4) \vee \\ (x_1 \wedge x_2 \wedge x_3 \wedge x_4) \vee (x_1 \wedge \overline{x}_2 \wedge x_3 \wedge x_4) \vee (x_1 \wedge \overline{x}_2 \wedge \overline{x}_3 \wedge x_4)\end{array}\right\} \quad (2.1)$$

Применяя аксиомы Булевой алгебры, получим упрощенное выражение:

$$(x_1 \wedge x_3) \vee (\overline{x}_1 \wedge \overline{x}_3) \vee (x_1 \wedge \overline{x}_2 \wedge \overline{x}_3 \wedge x_4) \quad (2.2)$$

Выражение (2.2) будем называть *функцией принадлежности*, которая характеризует систему казахских гласных в теореме, приведенной ниже.

**Теорема принадлежности** [13, 14]**.** Гласный звук $\lambda$ принадлежит системе казахских гласных тогда и только тогда, когда его артикуляционные признаки $x_1$, $x_2$, $x_3$ и $x_4$, определенные выше, удовлетворяют дизъюнктивной нормальной форме

$$(x_1 \wedge x_3) \vee (\overline{x}_1 \wedge \overline{x}_3) \vee (x_1 \wedge \overline{x}_2 \wedge \overline{x}_3 \wedge x_4)$$

На основе четырех артикуляционных признаков построена в [16] геометрическая модель гласных звуков казахского языка, показанной на рисунке 1.

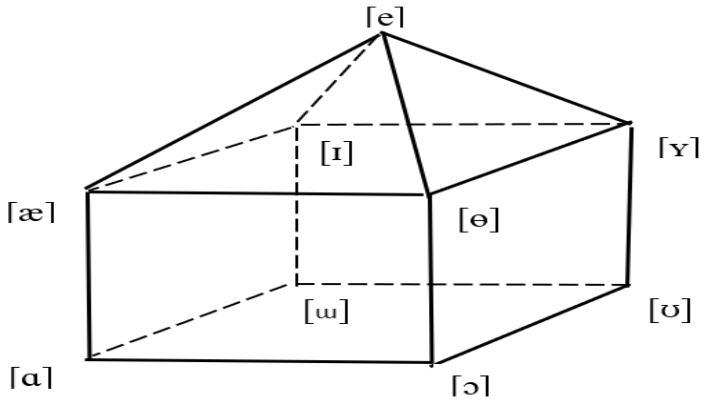



Рисунок 1. Геометрическая модель системы казахских гласных на основе 4-х бинарных артикуляционных признаков.

На рисунке 1 плоскость [ɑ] [ɔ] [ʊ] [ɯ] представляет заднеязычные (твердые) гласные звуки (а) (о) (ұ) (ы), плоскость [æ] [ɵ] [ʏ] [ɪ] – переднеязычные (мягкие) гласные звуки (ә) (ө) (ү) (і), плоскость [ɑ] [ɔ] [ɵ] [æ] – открытые гласные звуки (а) (о) (ө) (ә), плоскость [ɯ] [ʊ] [ʏ] [ɪ] – закрытые гласные звуки (ы) (ұ) (ү) (і), плоскость [ɔ] [ʊ] [ʏ] [ɵ] – округленные гласные звуки (о) (ұ) (ү) (ө), плоскость [ɑ] [ɯ] [ɪ] [æ] – неокругленные гласные (а) (ы) (і) (ә), а вершина [e] представляет особый звук (е).

По данной геометрической модели можно установить следующие факты (аксиомы):

1. Никакие вершины (гласные звуки) из нижней и верхней плоскостей, включая [e], не могут оказаться в исконно казахском слове. Этот факт называется небной (палатальной) гармонией гласных или сингармонизмом.

2. В казахском языке нет исконного слова, содержащего более трех различных гласных звуков, и если оно содержит три гласных звука, то это мягкие гласные, один из которых обязательно [е]. Это можно интерпретировать на рисунке 1 так: никакие 3 гласные нижней или верхней плоскостей не встречаются в казахском слове. Слово, содержащее 3 различные гласные, включает вершины плоскости, проходящей через [е] и любые 2 вершины верхней плоскости.

Теперь можно привести классификацию казахских гласных звуков на основе 4-х бинарных артикуляционных признака (по горизонтальному, по вертикальному положению языка, по положению челюсти, по положению губ), показанную в таблице 4.

Таблица 4. Классификация казахских гласных звуков на основе четырех бинарных артикуляционных признаков.


Здесь сверху показаны признаки по горизонтальному положению языка, слева - признаки по положению челюсти, справа - признаки по вертикальному положению языка, а признаки по положению губ учтены при расположении двух звуков в одной строчке – слева неокругленные, справа – округленные.

Акустический анализ гласных звуков казахского языка основан на на сингармонических тембрах: твердые нижние, мягкие нижние, твердые верхние. мягкие верхние и мягкие низкие. В таблице 5 приведена система сингармонических тембров гласных звуков казахского языка.

Таблица 5. Система сингармонических признаков гласных

| Обозначение гласных звуков | **Палатальный тембр** | | **Лабиальный тембр** | |
|---|---|---|---|---|
| | твердый | мягкий | низкий | высокий |
| (а) - [ɑ] | + | – | + | – |
| (ә) - [æ] | – | + | + | – |
| (о) - [ɔ] | + | – | – | + |
| (ө) - [ɵ] | – | + | – | + |
| (ұ) - [ʊ] | + | – | – | + |
| (ү) - [ʏ] | – | + | – | + |
| (ы) - [ɯ] | + | – | + | – |
| (і) - [ɪ] | – | + | + | – |
| (е) - [е] | – | + | – | + |

Мы провели эксперимент [14], микшируя осциллограммы заднеязычных (твердых) гласных звуков (а) - [ɑ], (о) - [ɔ], (ұ) - [ʊ] и (ы) - [ɯ] с осциллограммой звука [е] - (е) и обнаружили следующие свойства, представленные на рисунке 2:

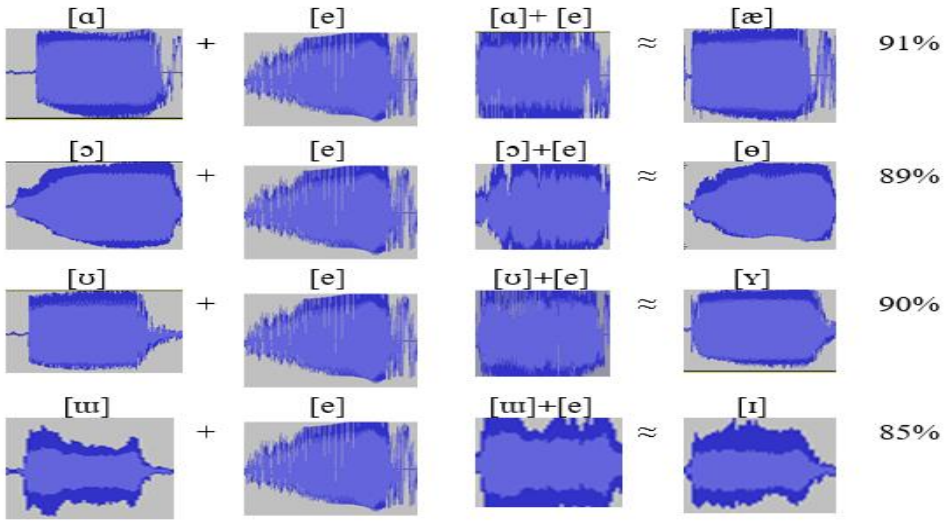



Рисунок 2. Свойства гласных звуков казахского языка

Указанные свойства доказывают, что в корневых словах казахского языка не должны встречаться пары, образованные из подряд стоящих гласных звуков (а) и (е), (о) и (е), (ұ) и (е), (ы) и (е).

Если указанные пары гласных звуков встречаются в составных словах между образующими их корневыми словами, то они должны разделяться пробелами.

В состав звуковой системы казахского не будет включаться только одна согласная фонема, которая обозначена в действующем алфавите буквой Щщ, а все остальные согласные фонемы: (б), (в), (г), (ғ), (д), (ж),

(з), (й), (к), (қ), (л), (м), (н), (ң), (п), (р), (с), (т), (ф), (х), (h), (ц), (ч), (ш) будут включены в неё. Среди фонемы (в), (ф), (х), (ц), (ч) являются заимствованными из русского языка, а остальные являются исконно казахскими фонемами.

Что касается заимствованных фонем (в), (ф), (х), (ц), (ч), то они используются для написания имен людей, наименований стран и местностей, а также международных непереводимых терминов. (Например, *Вагнер, Вьетнам, вагон, вакуум, вексель, вектор*, *вето, ввиза, викторина, вирус, вольт, Форд, Франция, факт, файл, факультет, федерация, физика, филармония, фонетика, формула, функция, Бухарест, Хельсинки, Хинчин, Хомский, Хонекер, Хорватия*, *Хуавей, химия, хлор, хром, хроника, Цезарь, Циолковский, Цой, Цейлон, Цюрих, цезий, цемент, цензура, циклотрон, цилиндр, циркуль, цунами, Чад, Чайковский, Чаплин, Челюскин, Черногория, Чехия, Чикаго, Чили, Чита, чеченец, чуваш*).

Указанные фонемы так глубоко внедрились в казахский язык, что даже у носителей казахского языка, не знающих русский язык, артикуляционные органы речи приспособились для их произношения. Эти фонемы не входят в состав суффиксов и окончаний, которые применяются при формировании словообразований и слоформ. Поэтому оставление их в составе звуковой системы казахского языка и использовать их при составлении нового алфавита на латинской графике как отдельные фонемы. От них казахский язык никак не пострадает, так как они не будут порождать противоречия в казахской орфографии и орфоэпии, наоборот позволят приблизить написания и звучания международных терминов к их оригиналам.

Таким образом, в реформе по переводу казахской письменности на латинскую графику будут участвовать **33** фонем, из них **9** гласные фонемы (а), (ә), (е), (о), (ө), (ұ), (ү), (ы), (і) и 24 согласные фонемы (б), (в), (г), (ғ), (д), (ж), (з), (й), (к), (қ), (л), (м), (н), (ң), (п), (р), (с), (т), (*w*), (ц), (ф), (х), (ч), (ш). В реформе будут участвовать также полугласные (и) и (у) в качестве аллафонов гласных фонем (і) и (ұ), соответственно.

## 6. Два варианта алфавита казахского языка на латинице

На основе уточнения в пункте 4 звуковой системы казахского языка в составлении его нового алфавита будут участвовать **33** фонем, из них **9** гласные фонемы и **24** согласные фонемы, а также полугласные звуки (и), (у) в качестве аллофонов гласных фонем (і) и (ұ), которые в МФА будут обозначаться как [i] и [w].

Заимствованные из русского языка гласные звуки (и) и (у), включенные в звуковую систему казахского языка как аллофоны

гласных фонем (і) и (ұ), будут обозначаться буквами, используемыми для обозначения фонем (і) и (ұ), соответственно. Например, если гласные фонемы (і) и (ұ) были обозначены латинскими буквами II и I, то не составило бы труда писать и читать такие термины, как *internet, institute, university, supremum*. Если в новом алфавите заимствованных гласных (и) и (у) обозначить отдельными буквами, то неизбежно появятся противоречия, упомянутые в пункте 2.

При определении состава нового казахского алфавита фонемы, для которых имеются адекватные латинские буквы будут обозначаться этими буквами без диакритических знаков и диграфов.

**Замечание**: Диграфы не входят в состав алфавита, они по существу являются орфографическими правилами того или иного языка, в котором они используются.

Фонемы, для которых не существуют адекватные латинские буквы будут создавать проблемы. К ним относятся мягкие гласные фонемы (ә), (ө), (ү) и (і), а также согласные фонемы (ғ), (қ), (ң), (ч) и (ш). Для решения этих проблем предлагаются два варианта алфавита на основе латинской графики:

1) В первом варианте будут использоваться надбуквенные и подбуквенные диакритические знаки, как в алфавите турецкого языка.

2) Во втором варианте будут использоваться диграфы, как в алфавите английского языка, ранее подобный алфавит был предложен в [14] без включения фонем (ц) – [ts] и (ч) - [tʃ] в звуковую систему казахского языка.

Теперь займемся выбором конкретных способов обозначения этих проблемных фонем. Так мягкие фонемы (ә), (ө), (ү) в первом варианте будут обозначаться с помощью надбуквенного диакритического знака «две точки»: Ä ä, Ö ö, Ü ü, при этом твердая фонема (ы) будет обозначаться буквой без точки I ı, а мягкая фонема (і) - надбуквенным диакритическим знаком «одна точка» I i. Во втором варианте алфавита эти фонемы будут представляться диграфами Ae ae, Oe oe, Ue ue.

Согласные фонемы (ң), (ғ), (ш) и (ч) в первом варианте будут обозначаться с помощью надбуквенного диакритического знака «тильда» Ññ и Ğğ, а во втором варианте – диграфами Ng ng, Gh gh.

В звуковую систему казахского языка включены согласные фонемы (ч) и (ш), которые будут обозначены в первом варианте подбуквенным диакритическим знаком «запятая» Ç ç и Ş ş, соответственно, а во втором варианте – диграфами Ch ch и Sh sh.

Предлагается включить в оба варианта алфавита букву базового латинского алфавита Xx, которая будет обозначать не конкретную фонему, а сочетания согласных (к) и (с). Использование этой буквы и не

будет порождать никаких противоречий в казахской письмепности и позволяет писать международные термины одинаково с оригиналом или ближе к нему. Например: *axelerat* - акселерат, *axis - ось, box - бокс, context - контекст, expert - эксперт, expor - экспортt, maxitum - максимум, mixer - миксер, Oxford - Оксфорд, taxi - такси, xerox - ксерокс*.

Первый вариант предлагаемого алфавита не порождает никаких противоречий в казахской орфоэпии и орфографии, так как он основан на уточненной звуковой системе казахского языка и на принципе «Один звук и одна буква».

При использования второго варианта предлагаемого алфавита возникают некоторые проблемы орфоэпии и орфографии казахского языка, связанные с использованием диграфов. Но от них можно избавиться, если учитывать следующие свойства и требования:

1) В исконно казахских словах перед согласным звуком (ң) не встречаются согласные звуки, поэтому в любом слове дограф «*ng*» должен писаться только в одном слоге.

2) При прямой кодировке разных по смыслу и записи на кириллице слов латинскими буквами происходит одинковое их написания:

$$kuengi = \begin{cases} \text{күнгі (дневной)} \\ \text{күңі (служанка)} \end{cases}$$

Поэтому, чтобы недопускать подобные случаи надо:

- учесть свойства ***ассимиляции*** (***уподобления***) звуков и ввести специальное орфографическое правило: «Если в слове за звуком (n) - [n] непосредственно следует звук (г), то при чтении такого слова звук (*н*) произносится как звук (ң) – [ŋ], т.е. звук (n) - [n] уподобляется (ассимилируется) звуку (ң) – [ŋ]. Например, **э*н*г**іме (беседа), түгі (ночной) при чтении произносится как **эң**гіме, түгі

- вместо одной латинской буквы ***n*** надо писать дограф ***ng***, например, «күгі = kue**n**gi = kue**ng**gi», что будет соответствовать закономерности произношения в казахском языке и не порождать неоднозначности в его письменности.

3) В слове “хана - помещение” фонему (х)- [h] надо заменить на фонему (қ) - [q]. Тогда записи комбированных с его участием не будут создавать никаких орфографических проблем. Например, «кітап**х**ана → кітап**х**ана = kitap**q**ana, қымыз**х**ана → қымыз**қ**ана = kitap**q**ana, кітап**х**ана → кітап**х**ана = kitap**q**ana».

4) В комбинированных словах или человеческих именн согласную фонему (х)- [h], которая непосредственно следует после согласной фонемы (с) - [s] или (з) - [z], надо заменить на фонему (қ) -[q]. Например, «асхана → асқана = asqana, Ас**х**ат → Ас**қ**ат = As**q**at, Дос**х**ан → Дос**қ**ан = Dos**q**an, қымызхана → қымызқана = qymyzqana, Оразхан → Оразқан = Orazqan. В противном случае, эти же слова «асхана = ashana = ашана, Ас**х**ат = Ashat = Ашат, Досхан = Doshan = Дошан, қымызхана = qymyzhana = қымыжана, Оразхан = Orazhan = Оражан».

Таким образом, устранив недостатки приведенного в таблице 1 шестого варианта алфавита казахского языка, предлагается на основе уточненной звуковой системы казахского два варианта алфавита на основе латинской графике: первый вариант на основе алфавита английского языка с использованием диграфов, второй варииант на основе алфавита турецкого языка с использованием диакретических знаков.

Предлаемые варианты алфавита казахского языка обладают следующими свойствами:

1 при создании и обработке информационных ресурсов на казахском языке не нуждаются в программных средствах, которые требуют дополнительные затраты на изготовление:

– шрифтов и драйверов, так как он состоит только из букв английского алфавита, размещенного на клавиатуре всех типов компьютеров с многочисленными шрифтами и драйверами для всех типов устройств ввода и вывода данных;

– программ сортировок и поиска информации, так как порядок следования букв в нём совпадает с порядком следования букв в английском алфавите;

2 алфавит позволяет сблизить написание непереводимых на казахский язык международных заимствованных терминов - слов с их написанием в оригинале без изменения звуковой системы казахского языка;

3 алфавит позволяет легко и быстро набрать казахский текст с помощью клавиатуры любого компьютера и смартфона.

В таблице 7 представлен алфавит казахского языка с использованием диакритических знаков на основе турецкого алфавита,

В таблице 8 представлен алфавит казахского языка с использованием диграфов на основе английского алфавита.

В таблице 9 приводятся орфографические правила представления фонем с помощью диграфов.

Таблица 7. Вариант алфавита казахского языка на основе турецкого алфавита

| № | Латиница | Кириллица | МФА | | № | Латиница | Кириллица | МФА |
|---|---|---|---|---|---|---|---|---|
| 1 | A a | А а | [ɑ] | | 18 | N n | Н н | [n] |
| 2 | Ää | Ә ә | [æ] | | 19 | Ñ ñ | Ң ң | [ŋ] |
| 3 | B b | Б б | [b] | | 20 | O o | О о | [ɔ] |
| 4 | C c | Ц ц | [tc] | | 21 | Ö ö | Ө ө | [ɵ] |
| 5 | Çç | Ч ч | [tʃ] | | 22 | P p | П п | [p] |
| 6 | D d | Д д | [d] | | 23 | Q q | Қ қ | [q] |
| 7 | E e | Е е | [e] | | 24 | R r | Р р | [r] |
| 8 | F f | Ф ф | [f] | | 25 | S s | С с | [s] |
| 9 | G g | Г г | [g] | | 26 | Ş ş | Ш ш | [ʃ] |
| 10 | Ğ ğ | Ғ ғ | [ɣ] | | 27 | T t | Т т | [t] |
| 11 | H h | Х х | [h] | | 28 | U u | Ұ ұ | [ʊ u] |
| 12 | I ı | Ы ы | [ɯ] | | 29 | Ü ü | Ү ү | [ʏ] |
| 13 | İ i | І і | [ɪ, i] | | 30 | V v | В в | [v] |
| 14 | J j | Ж ж | [ʒ] | | 31 | W w | | [w] |
| 15 | K k | К к | [k] | | 32 | X x | | [ks] |
| 16 | L l | Л л | [l] | | 33 | Y y | Й й | [y] |
| 17 | M m | М м | [m] | | 34 | Z z | З з | [z] |

Таблица 8. Алфавит казахского языка на основе английского алфавита

| № | Латиница | Кириллица | МФА | | № | Латиница | Кириллица | МФА |
|---|---|---|---|---|---|---|---|---|
| 1 | A a | А а | [ɑ] | | 14 | N n | Н н | [n] |
| 2 | B в | Б б | [b] | | 15 | O o | О о | [ɔ] |
| 3 | C c | Ц ц | [tc] | | 16 | P p | П п | [p] |
| 4 | D d | Д д | [d] | | 17 | Q q | Қ қ | [q] |
| 5 | E e | Е е | Е е | | 18 | R r | Р р | [r] |
| 6 | F f | Ф ф | [f] | | 19 | S s | С с | [s] |
| 7 | G g | Г г | [g] | | 20 | T t | Т т | [t] |
| 8 | H h | Х х | [h] | | 21 | U u | Ұ ұ | [ʊ u] |
| 9 | I i | І і | [ɪ, i] | | 22 | V v | В в | [v] |
| 10 | J j | Й й | [j] | | 23 | W w | | [w] |
| 11 | K k | К к | [k] | | 24 | X x | | [ks] |
| 12 | L l | Л л | [l] | | 25 | Y y | Ы ы | [ɯ] |
| 13 | M m | М м | [m] | | 26 | Z z | З з | [z] |

Таблица 9. Орфографические правила для диграфов

| № | Латиница | Кириллица | МФА | Пояснение |
|---|---|---|---|---|
| 1 | Ae ae | Ә ә | [æ] | |
| 2 | Oe oe | Ө ө | [ө] | |
| 3 | Ue ue | Ү ү | (ү) | |
| 4 | Ch ch | Ч ч | [tʃ] | Как в английском |
| 1 | Gh gh | Ғ ғ | [ɣ] | |
| 5 | Ng ng | Ң ң | [ŋ] | Как в английском |
| 6 | Sh sh | Ш ш | [ʃ] | Как в английском |
| 7 | Zh zh | Ж ж | [ʒ] | |

## 7. Заключение

В эпоху цифровизации общества многие естественные языки развиваются с помощью компьютерных программ: *созданы электронные словари, мультимедийные вопросно-ответные системы, интеллектуальные системы обучения и оценки знаний, машинные переводчики с одного языка на другой, системы распознавания и синтеза письменной и устной речи и т.д.* Основами этих раабот являются математические модели грамматических правил этих языков.

Подобные проблемы можно ставить и успешно решать и для казахского языка. Однако указанные ошибки не дают возможность формализовать морфологических правил казахского языка и автоматизировать морфологический анализ и синтез казахских слов. Предложенный новый алфавит позволяет успешно решать эти проблемы.

**Список литературы**

1. Казахский язык // Большая советская энциклопедия : в 30 т. / 3-е изд. – М.: Советская энциклопедия, 1969–1978.
2. https://ru.wikipedia.org/wiki/Казахский_язык
3. https://ru.wikipedia.org/wiki/Казахи.
4. https://ru.wikipedia.org/wiki/Древнетюркское_руническое_ письмо.
5. https://ru.wikipedia.org/wiki/Иссыкское_письмо.
6. http://www.niac.gov.kz/kz/expertiza/item/138-dzhadidizm-i-reforma-musulmanskogo-obrazovaniya.
7. International Phonetic Association. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge, U.K., 2003.
8. https://en.wikipedia.org/wiki/International_Phonetic_Alphabet.
9. Мырзабеков С. Қазіргі қазақ тілінің фонетикасы. Алматы, Дәуір, 2013. – 219 с.
10. Калиева Б.А. Қазақ тілінің фонетикасы. Алматы: Эверо, 2014. – 162 с.
11. http://onstudy.kz/dausty_dybys.

12. https://ru.wikipedia.org/wiki/Булева_алгебра.

13. Z. Yessenbayev, M. Karabalayeva and A. Sharipbayev, "Formant Analysis and Mathematical Model of Kazakh Vowels," 2012 UKSim 14th International Conference on Computer Modelling and Simulation, Cambridge, 2012. - P.427-431.

14. Шарипбай А.А. Проблемы перевода казахской письменности на латинский алфавит (на казахском и русском языках). Монография, Астана, ЕНУ имени Л.Н.Гумилева, 2017, -138 С.

*УДК 004.8+81’322*
***Сулейманов Джавдет Шевкетович***
*Научный руководитель института*
*прикладной семиотики Академии наук РТ, д.т.н., профессор*
*Казань, Татарстан, Россия*
*dvdt.slt@gmail.com*

# ИНФОКОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ И ТАТАРСКИЙ ЯЗЫК

**Аннотация:** В статье описывается многолетний опыт внедрения татарского языка в инфокоммуникационные технологии (ИКТ) в Республике Татарстан, включая локализацию компьютерных систем, создание стандартов, разработку прикладных программ и лингвистических ресурсов. Исследования и разработки осуществляются в институте прикладной семиотики Академии наук РТ в трех направлениях. Первое направление - национальная локализация компьютерных систем, включает разработку кодовых страниц, драйверов устройств, создание словарей и терминотворчество в области инфокоммуникационных технологий, а также татарскую локализацию известных программных продуктов, таких как ОС Windows и другие. Второе направление исследований посвящено созданию программных продуктов и инструментария для обработки татарского языка и лингвистических баз данных. Дается описание ряда важных программных продуктов: машинного переводчика, морфологического анализатора, национального корпуса татарского языка «Туган тел», Электронного Атласа диалектов татарского языка, Интернет-платформы «Тюркская морфема». Третье направление - когнитивные исследования в татарском языке и разработка лексико-грамматических моделей как формальной базы для создания интеллектуальных технологий обработки информации. В статье описывается ряд свойств татарского языка, делающих его привлекательным для систем искусственного интеллекта как языка человеко-машинного интерфейса, а также языка обмена информацией и общения между интеллектуальными системами. Показано, что наиболее привлекательными перспективными для создания интеллектуальных технологий являются следующие свойства татарского языка: 1) регулярность морфотактики; 2) фиксированность позиций соответствующих типов аффиксальных морфем; 3) морфологический эллипсис - возможность пропуска последовательности аффиксов при однородных именных словоформах с сохранением ее в последней словоформе; 4) возможность циклического

порождения нового значения путем последовательного применения одной и той же «формулы», то есть повторного присоединения одного и того же аффикса; 5) возможность рекурсивно задавать нечеткие команды и описывать нечеткие действия; 6) возможность рекурсивно описывать в рамках одной словоформы действия, относящиеся к целой ролевой ситуации; 7) построение предложения по схеме: S-O-V - сначала дается информация и ее анализ, и только после этого, возможно, с учетом реакции слушающего, определяется – положительное или отрицательное, само действие.

В системах искусственного интеллекта это называется активностью знаний, что является одним из важных признаков интеллектуальности системы. Для подобных систем естественным и основополагающим является стиль размышления: анализ-действие, размышление-цели-алгоритмы, а не командный стиль: действие-анализ, алгоритм-цель, как это реализовано в современных технологиях, основанных на менталитете индо-европейских языков.

**Ключевые слова:** национальная локализация, ИКТ, лингвистические ресурсы, когнитивный потенциал языка, морфологический эллипсис, активность знаний.

UDC 004.8+81’322

***Suleymanov Dzhavdet Shevketovich***
*Chief Researcher of the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan*
*Doctor of Engineering Sciences, professor*
*Kazan, Tatarstan, Russia*
*dvdt.slt@gmail.com*

## INFOCOMMUNICATION TECHNOLOGIES AND THE TATAR LANGUAGE

**Abstract:** The article describes of experience in introducing the Tatar language into information and communication technologies (ICT) in the Republic of Tatarstan, including the localization of computer systems, the creation of standards, the development of application programs and linguistic resources. Research and development are carried out at the Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan in three directions. The first direction - the national localization of computer systems, includes the development of code pages, device drivers, the creation of dictionaries and term creation in the field of infocommunication technologies, as well as the Tatar localization of well-known software

products, such as Windows OS and others. The second direction of research is devoted to the creation of software products and tools for processing the Tatar language and linguistic databases. The description of a number of important software products is given: a machine translator, a morphological analyzer, the national corpus of the Tatar language "Tugan tel", the Electronic Atlas of the Tatar language dialects, the Internet platform "Turkic Morpheme". The third direction is cognitive research in the Tatar language and the development of lexical and grammatical models as a formal basis for creating intelligent information processing technologies. The article describes a number of properties of the Tatar language that make it attractive for artificial intelligence systems as a human-machine interface language, as well as a language for information exchange and communication between intelligent systems. It is shown that the following properties of the Tatar language are the most attractive and promising for the creation of intellectual technologies: 1) regularity of morphotactics; 2) fixed positions of the corresponding types of affixal morphemes; 3) morphological ellipsis - the possibility of skipping a sequence of affixes with homogeneous nominal word forms while preserving it in the last word form; 4) the possibility of cyclic generation of a new meaning through the successive application of the same "formula", that is, the repeated attachment of the same affix; 5) the ability to recursively set fuzzy commands and describe fuzzy actions; 6) the ability to recursively describe, within the framework of one word form, actions related to the whole role situation; 7) building a sentence according to the scheme: S-O-V - first, information is given and its analysis, and only after that, perhaps, taking into account the reaction of the listener, is it determined - positive or negative, the action itself.

In artificial intelligence systems, this is called knowledge activity, which is one of the important features of the system's intelligence. For such systems, the thinking style is natural and fundamental: analysis-action, thinking-goals-algorithms, and not the command style: action-analysis, algorithm-goal, as implemented in modern technologies based on the mentality of the Indo-European languages.

**Key words:** national localization, ICT, linguistic resources, language cognitive potential, morphological ellipsis, knowledge activity

**Введение**

В современную эпоху глобализации и повсеместной цифровизации вопрос сохранения национальных языков обрел особую остроту и актуальность.

Очевидно, что одним из надежных и перспективных решений является включение языка в цифровое пространство как языка

накопления, обработки и передачи информации; языка коммуникации в глобальной компьютерной сети Интернет [CMLS, 2020].

В настоящее время практически полностью реализован комплекс организационных и директивных мер и создана необходимая информационно-технологическая база для обеспечения полноценного функционирования татарского языка в компьютерных технологиях.

Как правило, во «взаимоотношениях» компьютеров и ЕЯ исследования и разработки преследуют цель — обеспечить компьютерную поддержку языка, его изучение, сохранение и развитие. При этом, на наш взгляд, незаслуженно остается вне внимания такая весьма перспективная активность, как исследование и использование естественного языка как ресурса, имеющего необходимый потенциал для создания новых технологий обработки знаний и интеллектуального интерфейса Система искусственного интеллекта (СИИ) - Человек и СИИ - СИИ.

Исследования и разработки в Институте прикладной семиотики АН РТ осуществляются в трех направлениях: 1) Татарская локализация ИКТ, 2) Разработка инструментария и прикладных программ обработки естественного языка (ЕЯ) и создание лингвистических ресурсов, 3) Когнитивные исследования лексико-грамматического потенциала татарского языка (ТЯ) с целью создания новых интеллектуальных технологий обработки информации.

Первое направление непосредственно связано с проблемой повышения активности языка в сети Интернет; обеспечения функционирования татарского языка как государственного в Республике Татарстан, а также с национальной локализацией интерфейса и ресурсов ИКТ.

Следующее направление ориентировано на создание оригинальных технологий, программного инструментария и лингвистических ресурсов, предназначенных, главным образом, для обработки татарского языка. Также описан подход к решению одной из актуальных задач в области электронного обучения – задачи автоматизации контроля и оценки ответов обучаемого в свободной форме на ЕЯ.

Третье направление посвящено исследованию и решению задачи создания интеллектуальных технологий на основе использования когнитивного потенциала татарского языка. Среди важных признаков интеллектуальности систем, как правило, выделяются следующие: активность знаний, то есть первичность анализа данных и вторичность принятия решения; обработка нечеткой информации, и исполнение нечетких команд. В статье на примерах показывается перспектива

исследования в этих целях татарского языка как языка агглютинативного типа, обладающего такими важными свойствами, как регулярность, рекурсия, активность знаний и ряд других.

Национальная локализация ИКТ

В этом направлении исследований и разработок решаются следующие задачи: включение языка в компьютерные системы, сохранение языка, повышения его активности в сети Интернет; обеспечение функционирования татарского языка как государственного в Республике Татарстан, а также национальной локализации интерфейса и ресурсов ИТ; создание стандартов для использования татарского языка в компьютерных системах и Интернете, разработка и стандартизация системы татарских терминов и понятий в ИКТ.

В настоящее время задача национальной локализации для татарского языка решена полностью. Принято Постановление КМ РТ «О стандартах кодировки символов татарского алфавита для компьютерных применений» № 1026 от 9 декабря 1996 года. На основе стандартов учеными разработаны драйверы и шрифтовое обеспечение для татарского языка на кириллической основе. Создан пакет стандартных технологий и программ базовой татарской локализации «Татарский Офис 2001». На основе стандартов унифицированы драйверы устройств, которые в первое время создавались различными группами и отдельными специалистами по своему усмотрению и практически как вирус распространились по различным компьютерам.

На базе принятых стандартов по соглашению с фирмой Microsoft осуществлена полная татарская локализация операционной системы MS Windows и ее офисных приложений, включая интерфейс и справочные файлы, а также корректор татарских тестов (совместно с фирмой Microsoft). Разработано и внедрено более 5000 новых татарских терминов и понятий для компьютерных технологий. Текст переводов на татарский язык интерфейса и файлов помощи составил порядка 1 млн. словоформ. Работа по татарской локализации новых версий программных продуктов Microsoft продолжается и сегодня.

В настоящее время происходит активное внедрение татарского языка в мобильные устройства. Создаются локализованные сервисные приложения: клавиатура, словари, системы предиктивного набора текста (Predictive text input), игры, обучающие программы для различных систем (iOS 7, Android 4+, WP 8.0 – 8.1).

Одной из ключевых вопросов локализации компьютерных систем является терминотворчество.

В основу принципов перевода терминов при татарской локализации легли правила образования и применения терминов и понятий в

татарском языке, разработанные татарскими филологами [Закиев, 1995], а также результаты наших исследований, полученные при локализации компьютерных систем, при создании англо-татарско-русского толкового словаря по терминам информатики и при обучении студентов [Сулейманов, 2015].

При татарской локализация операционной системы MS Windows решались следующие задачи, требующие перевода терминов и понятий:

1) перевод интерфейсов операционных систем и ее офисных приложений – перевод текстов, которые отображаются на экране дисплея (например, текстов, используемых при работе с электронной почтой, с офисными приложениями Windows);

2) перевод на татарский язык текстов на кнопках меню (Save, Open, Close и т.д.); перевод файлов справок и помощи на татарский язык.

Татарская локализация операционной системы и офисных приложений – это не прямой перевод с английского или русского языков, а творческая адаптация программного продукта для комфортной работы татаро-язычного пользователя в соответствующей операционной среде. Как правило, при переводе текстов, терминов и понятий, относящихся к такого рода многофункциональным сложным разработкам, возникает необходимость использования знаний из разных дисциплин. Соответственно, национальная локализация выполняется в институте совместно с программистами, математиками, лингвистами, специалистами по татарскому языку.

Наряду с принципами, разработанными лингвистами, и хорошо известными в науке нами использовались два новых принципа. Подробное их описание дается в работе [Сулейманов, 2018].

1) Принцип «формального гнезда». Образование новых слов из матрицы татарского слова, в котором «убираются» все гласные буквы, и слово-схема заполняется другими татарскими гласными буквами.

Структуру татарских корневых слов можно представить как матрицу из согласных букв, заполненных гласными буквами. В современном татарском языке 9 гласных букв (а-ә, о-ө, у-ү, ы-е, и). Соответственно, заполняя, например, матрицу «т-з», можно образовать следующие слова: таз-тәз-тоз–төз-туз-түз-тыз-тез-тиз. Здесь 7 слов из 9-и представлены в татарском словаре: таз (тазик), тоз (соль), төз (стройный), туз (береста), түз (терпи), тез (строй), тиз (быстро). Такие схемы мы называем «формальными гнездами», это практически готовые структуры для порождения новых слов.

Тот же принцип «формального гнезда» можно применить и к структуре слова, образованного из слогов.

2) Принцип «возвращенных» слов. Возвращение в язык слова, которое этимологически является тюркским, использовалось в языке, сохранилось в других языках, или вышло из употребления в языке как «архаизм».

Одним из таких «возвращенных» слов является слово айкен (иконка, icone), обозначающее пиктограмму – небольшое растровое символическое изображение, используемое в графическом интерфейсе пользователя для выбора того или иного инструмента (программы) или файла.

Слово айкен-айкөн составлено из двух тюркских слов ай – луна и кен (көн) – солнце. Соответственно, вполне логично пиктограмму (знак) иконка по-татарски называть возвращенным словом айкен.

Программные системы и технологии и лингвистические ресурсы для татарского языка

Следующее направление исследований и разработок Института - это создание программного инструментария, прикладных программ и лингвистических ресурсов для татарского языка. Исследования теоретических и прикладных проблем компьютерной лингвистики применительно к татарскому языку, к его грамматике, лексикологии и лексикографии, к различным проявлениям в речи, с целью построения лингвистических моделей и создания на их базе систем обработки татарского языка и различных лингвистических ресурсов, являются весьма наукоемкой задачей и требуют интеграции знаний и умений специалистов в смежных областях – информатике, математике и лингвистике. Описанию современного состояния исследований и разработок в этом направлении и раскрытию приведенных ниже разработок в функциональном, содержательном и технологическом аспектах посвящена коллективная монография [Формальные модели, 2019].

Одной из фундаментальных задач, которая была поставлена и выполняется с1990-х годов, является задача создания и поддержки Машинного фонда татарского языка (МФТЯ) в сети Интернет со следующими корпусами: а) электронные неформатированные тексты (газеты, журналы, книги, документы и др.); б) размеченные тексты, словари, тезаурусы; в) рабочие версии программных систем и технологий: лингвопроцессоры (машинные переводчики, синтезатор речи, распознаватели текста и речи и др.), АРМы специалиста (учителя, редактора, лингвиста и др.), интеллектуальная многоязычная машина поиска.

Данная задача в настоящее время обрела более конкретные черты в рамках выполнения Государственной программы по изучению,

сохранению и развитию языков народов РТ, и привела к двум тесно связанным направлениям: 1) созданию программных средств и пакетов прикладных программ поддержки татарского языка в инфокоммуникационных технологиях и 2) созданию лингвистических ресурсов. В настоящее время в этих направлениях получены значительные результаты, ряд из которых приведен ниже.

Морфологический анализатор татарского языка и программный комплекс снятия морфологической неоднозначности (http://tatmorphan.pythonanywhere.com/). Как известно, татарский язык, как один из тюркских языков, обладает содержательно богатой и формально элегантной, регулярной, почти автоматной, морфологией [Heintz, 1989]. Морфологическая модель татарского языка является базовой составляющей практически во всех полнофункциональных лингвистических процессорах. Учитывая структурную специфику татарского языка и исходя из прикладных задач, к настоящему времени нами разработаны три различные модели морфологии. Генеративная модель морфологии, основанная на правилах словоизменения, обеспечивающая полноту анализа словоформы, распознавая словоформы потенциально неограниченной длины. Парадигматическая модель татарской морфологии обеспечивает быстрое распознавание словоформ и анализ корректности татарских словоформ с точностью до 95 % и используется в операционной среде MS Windows и ее офисных приложениях.

Кроме того, в рамках совместного проекта с Белкентским университетом (Турция) разработана двухуровневая модель морфологии татарского языка. На ее основе разработан морфологический анализатор татарского языка, используемый как средство аннотации текстов, и программный комплекс снятия морфологической многозначности в корпусе татарского языка.

Создана также гибридная модель морфологического анализа, использующая генеративный и парадигматический подходы и являющаяся частью информационно-инструментального комплекса «Татарская морфема» [Гатиатуллин, 2020].

Система машинного перевода (СМП) (https://translate.tatar). База параллельных текстов включает более 1.5 млн. русско-татарских пар предложений и двуязычные специализированные словари фамилий, имен, отчеств, государств, субъектов РФ, районов РТ, населенных пунктов, гражданств, национальностей в объеме более 95 тыс. словарных пар. Русско-татарский переводчик Татсофт, реализованный с использованием методов искусственного интеллекта, превосходит все аналоги по качеству перевода. Перевод осуществляется с «пониманием»

всего предложения, а не отдельных слов/фраз, реализованы возможности диктовки и озвучивания перевода, а также дальнейшего улучшения переводчика за счет моделей, созданных специально для татарского языка.

Татарский национальный корпус (ТНК) «Туган тел» (http://tugantel.tatar/). Формирование репрезентативной лингвистической ресурсной базы служит сохранению, изучению и развитию языка. В этом плане одним из важных разработок института является Татарский корпус «Туган тел» («Родной язык») [Khusainov, 2015]. Это лингвистический ресурс современного литературного татарского языка, адресованный широкому кругу пользователей: лингвистам, специалистам в области татарского, тюркского и общего языкознания, типологам, преподавателям татарского языка, деятелям культуры, а также всем, кто изучает и интересуется татарским языком. В настоящее время включает размеченные тексты различных жанров объемом более 200 млн. слов. Программная платформа корпуса языка представляет собой оригинальную разработку нашего Института и имеет инструментарий, практически «заточенный» под татарский язык, легко перенастраиваемый под другие тюркские языки.

Электронная версия Атласа татарских народных говоров (http://atlas.antat.ru/). Электронный Атлас татарских народных говоров построен совместными усилиями специалистов НИИ «Прикладная семиотика» АН РТ, ИЯЛИ АН РТ и КФУ, охватывает все основные районы расселения татар и отражает сведения по фонетике, морфологии, лексике и синтаксису татарского языка, собранные в 28 регионах Российской Федерации. Реализована специальная программа для сбора материалов и включения их в базу диалектологического Атласа силами пользователей. Электронный атлас татарских говоров позволяет получить информацию об имеющихся языковых явлениях в привязке к конкретным географическим объектам на картах. На данный момент используется информация с 215 карт языковых явлений.

Интернет-сервис «Многофункциональная модель тюркской морфемы» (МТМ). Еще одна разработка – Интернет-платформа «Тюркская морфема» [Gatiatullin, 2020] позволяет осуществить полную “инвентаризацию” тюркских морфем с описанием их характеристик на всех языковых уровнях (фонологическом, морфологическом, синтаксическом, семантическом). Основные функции сервиса: формирование ресурсной базы для программных продуктов, осуществляющих компьютерную обработку тюркских языков, таких как системы машинного перевода, информационно-поисковые системы, системы разметки электронных корпусов, извлечения данных и др.;

информационно-справочная система, содержащая практически полную информацию о тюркских морфемах; инструментарий для исследований ученых-тюркологов.

Важно отметить, что эта Интернет-платформа является одновременно и ценным программным инструментарием, а также лингвистической базой для сравнительного изучения тюркских языков и реализации совместных коллективных проектов с участием ученых, являющихся носителями языка этноса, с глубокой лингвистической интуицией.

В настоящее время к коллективной работе над заполнением баз данных Интернет-портала уже подключились 26 специалистов по 18 тюркским языкам, среди которых чувашский, крымскотатарский, тувинский, гагаузский, азербайджанский, башкирский, казахский, алтайский, киргизский, узбекский, якутский, татарский, уйгурский, кумыкский и др.

Когнитивные исследования лексико-грамматического потенциала татарского языка

Третье направление исследований института – когнитивные исследования лексико-грамматического потенциала татарского языка, связано с актуальной и перспективной задачей создания интеллектуальных операционных систем и интеллектуального программного инструментария на основе естественных языков, с учетом их структурных и концептуальных особенностей [Suleymanov, 2010].

Известно, что разработка моделей ЕЯ, исследование их возможностей для разработки языков искусственного интеллекта входят в число базовых проблем в области построения интеллектуальных систем. Такие задачи, как компьютерная обработка больших массивов ЕЯ-текстов, ЕЯ-диалог с системой, создание больших банков информации на основе ЕЯ, разработка языков посредников в многоязычной информационной среде, базирующихся на более развитых лингвистических моделях, приобретают особую актуальность в связи с развитием глобальных компьютерных сетей и формированием больших объемов распределенных данных.

В исследовании естественных языков можно выделить три аспекта: когнитивный, коммуникативный и технологический. Когнитивный аспект – это характеристика естественного языка с точки зрения возможностей описания модели мира, представления знаний. Коммуникативный аспект отражает потенциал естественного языка для кодирования, приема и передачи, семиотической обработки информации, организации диалога. Технологический аспект определяет формальный и концептуальный потенциал естественного языка для

реализации средств эффективной обработки, адекватного описания и компактного хранения информации на данном языке, создания эргономичных технических средств, учитывающих специфику языка (например, частотность букв при разработке клавиатуры), а также для разработки интеллектуального программного инструментария, включая операционные системы. Очевидно, в основе искусственных языков и систем программирования лежат глубинные структуры, ментальность естественного языка и, таким образом, эти системы реализуют описательный и вычислительный потенциал соответствующего ЕЯ. Как известно, современные средства накопления и обработки знаний на естественном языке малоэффективны и практически не справляются с такими задачами, как поиск и отбор информации в распределенных базах данных, извлечение знаний, семантический анализ текстовой информации. Причиной этому, прежде всего, служит то, что они изначально являются неинтеллектуальными, созданы на основе примитивных искусственных языков программирования, практически представляющих собой подмножество флективно-аналитических языков или искусственных конструкций, созданных на их основе.

В связи с этим перспективным представляется разработка нового программного инструментария по следующей технологии: 1) исследование и выявление естественных грамматических (морфологических, синтаксических, семантических) конструкций в различных языках, достаточно регулярных и обладающих естественной сложностью, в целях создания на их базе языков искусственного интеллекта нового поколения; 2) разработка языка-посредника на основе подмножеств и конструкций языков с определенными свойствами, позволяющими наиболее адекватно и сжато описывать контекст и быстро обрабатывать тексты на ЕЯ.

Как известно, для систем обработки знаний важны следующие характеристики, определяющие их эффективность и интеллектуальность: 1) время обработки; 2) объем памяти для хранения информации; 3) компактность хранения и передачи смысла; 4) возможность кодирования и обработки нечеткой информации; 5) активность знаний. Причем, первые три параметра описывают эффективность, а параметры 4 и 5 – интеллектуальность систем и технологий. Как показывают исследования [Suleymanov, 2010], татарский язык, являясь одним из тюркских языков, имеет богатую, сложную, но достаточно регулярную морфологию, обладает потенциалом, позволяющим эффективно кодировать и компактно хранить информацию, а также реализовывать на уровне аффиксальных морфем такие явления, как рекурсия, «нечеткость».

В объектно-предикативной модели мира именные группы, как правило, маркируют некое состояние объекта или объектов, в то время как действие, отношения между объектами и группой объектов описываются глагольной группой. Соответственно, выделяются когнитивные механизмы, реализуемые в рамках именной группы и когнитивные механизмы, реализуемые в рамках глагольной группы. Кроме того, сама структура текста, которая определяется синтаксическими закономерностями языка, служит одним из когнитивных механизмов языка, управляющим в тексте такой важной характеристикой, как активность знаний, естественным образом реализуя логическую схему: анализ-действие (известно, что активность знаний есть один из важных признаков интеллектуальной системы).

Далее в статье описываются и иллюстрируются на примерах соответствующие когнитивные формализмы, выделенные в татарском языке.

Когнитивные механизмы при описании состояния объектов. Как известно, татарская морфология является регулярной, почти автоматной [Heintz, 1983], и в то же время имеет естественную сложность, которая заключается, прежде всего, в следующем: 1)возможность присоединения определенных аффиксальных морфем, превращающих именную словоформу в глагольную или в форму прилагательного, и наоборот; 2) морфологическое (синтетическое) задание признаков модальности, настроения, эмоционально-личностного отношения к ситуации, объекту или процессу, описываемых данной словоформой; 3) контекстное разнообразие значений аффикса. Известно, что именная группа, как правило, кодирует некую семантическую ролевую ситуацию, а глагольная группа – контекстные отношения над этими ролями. Таким образом, возможность перехода с именной формы к глагольной и наоборот через присоединение соответствующих аффиксов позволяет описывать одновременно в пределах одной словоформы как сложную ролевую ситуацию, так и контекстные отношения между семантическими ролями. Тем самым обеспечивается компактность описания и хранения информации. Синтетический, аффиксальный способ словоизменения обеспечивает кодирование в рамках одной словоформы некоторого значения, описываемого на флективно-аналитических языках (например, на английском) несколькими словосочетаниями и даже предложениями.

Вместе с тем, морфология в большой степени регулярна, близка к автоматной, с небольшим количеством исключений из правил, что обеспечивает минимизацию емкостных и временных функций при обработке текстов на татарском языке, а также упрощает анализ

структуры и значения словоформы, несмотря на естественную сложность морфологии.

Важным свойством татарской морфологии, наряду с ее регулярностью, – фиксированное размещение аффиксов в последовательности аффиксальных морфем. Регулярность морфологии означает, что одна и та же схема сочетания морфем (морфотактика) присуща всем или почти всем именным и глагольным группам. Это дает возможность по одной и той же схеме, практически автоматно, образовывать словоформы с одинаковыми глубинными значениями аффиксов.

Например:

1) кран, краннар, краннарым, краннарыма – (‘кран, краны, мои краны, моим кранам’);

2) ат, атлар, атларым, атларыма – (конь, кони, мои кони, моим коням).

Именные корневые морфемы кран (кран), ат (конь) имеют одни и те же последовательности аффиксальных морфем с идентичными значениями. Обобщенно, приведенные парадигмы описываются следующими схемами:

Х(Имя сущ.), Х(Имя сущ.)+лар(афф.мн.), Х(Имя сущ.)+лар (афф.мн.) +ым (афф. притяж., 1 л., ед.ч.), Х(Имя сущ.)+ лар (афф.мн.) +ым(афф. притяж., 1 л., ед.ч.).+ а (афф. падежн., дат. падеж’).

Таким образом, мы определили два первых когнитивных механизма татарского языка: 1) регулярность морфотактики и 2) фиксированность позиций соответствующих типов аффиксальных морфем.

Как это следует из описания морфотактики, аффиксальная морфема присоединяется справа к именной словоформе, являющейся самой правой составляющей в последовательности словоформ, и относится ко всей именной группе.

Например:

Балачактан яраткан китапларым - ‘книги, любимые мной с детства’

(Балачактан яраткан китап) + лар (мн. )+ ым (притяж.)

Следующая возможность в татарской морфологии, которая может быть отнесена к третьему когнитивному механизму, называется морфологический эллипсис, это:

3) Возможность пропуска последовательности аффиксов при однородных именных словоформах с сохранением ее в последней словоформе.

То есть, возможность вывода любой последовательности аффиксов, общих для однородных членов, вправо, за последовательность

однородных членов, и присоединение их к последнему справа однородному члену.

Например:

Ишек алды тавыкларга, казларга, сарыкларга тулы = Ишек алды тавык, каз, сарыкларга тулы. (Двор полон кур, гусей, овец).

Мин кырларыбызга, урманнарыбызга, елагларыбызга шатланам = мин кыр, урман, елгаларыбызга шатланам . (Я радуюсь нашим полям, лесам, рекам).

Одним из важных и интересных когнитивных механизмов в татарском языке является рекурсия:

4) Возможность циклического порождения нового значения путем последовательного применения одной и той же «формулы», то есть повторного присоединения одного и того же аффикса.

Такими свойствами обладают аффиксальные морфемы –ДАгЫ (локатив2) и –нЫкЫ (притяжат.). Например, пусть задана лексема урман (‘лес’). Присоединение аффикса –дагы порождает новые объекты или свойства, являющиеся неопределенными: урмандагы – ‘нечто в лесу’; урмандагыдагы – ‘нечто в нечто в лесу’; ураманныкы – ‘то, что принадлежит лесу’; урманныкыныкы – ‘то, что принадлежит тому, что принадлежит лесу’.

Нетрудно заметить, что, эксплицитно задавая параметры после каждой морфемы, можно получить контекстную определенность словоформы. В реальных случаях такие параметры задаются имплицитно (т.е. неявно), наполняясь конкретным значением в зависимости от контекста речи. Рассмотрим следующий пример для иллюстрации изложенного утверждения. Пусть после каждого аффикса неопределенности стоят параметры: урман+ныкы(x0)+ ндагы(x1)+ныкы(x2)+ныкы(x3)+ндагы(x4)+ныкы(x5), где xi – контекстные объекты, т.е. объекты, которые либо приобретают конкретное значение из контекста, либо их задает пользователь (i = 1,…4). Тогда, придавая значения параметрам: x0 = «сосна», x1 = «ветка», x2 = «белка», x3 = «хвост», x4 = «шерсть», мы получаем следующее контекстное значение: «нечто (значение x5, придаваемое параметру последним аффиксом, осталось неопределенным) на шерсти, что принадлежит хвосту, что принадлежит белке, что на ветке, что принадлежит сосне».

5) Возможность рекурсивно задавать нечеткие команды и описывать нечеткие действия.

6) Возможность рекурсивно описывать в рамках одной словоформы действия, относящиеся к целой ролевой ситуации.

Свойство 5 кодируется глагольными аффиксами, занимающими позицию залога, т.е. сразу же после глагольной основы – ГАлА, - штЫр.

Например:

ю (‘мой’) – ‘мыть’ (3 лицо, ед.ч., повел. накл.)

югала (‘мой время от времени’)

ю(‘мой’)+гала (‘время от времени’)

югалаштыр (‘мой время от времени, время от времени – реже’)

ю(‘мой’)+гала(‘время от времени’)+штыр (‘время от времени’)

югалаштыргалаштыргала… (‘мой время от времени, время от времени, время от времени – и еще реже…’)

ю (мыть, корень, 3 лицо, ед.ч., повел.накл.)+гала (‘время от времени-изредка’)+штыр (‘время от времени-еще реже’)+гала (еще реже)+штыр(еще реже)+гала(еще реже)…

Сам факт, насколько редко требуется мыть – определяется, исходя из контекстной информации.

Реализация свойства 6 обеспечивается рядом специальных глагольных аффиксов, занимающих также залоговую позицию: -н, -Ыш, -т, -ДЫр.

Активность знаний. Известно, что английские предложения строятся по схеме S-V-O (subject-verb-object: субъект-глагол-объект)(“I’ll go to the cinema “Atilla” with my friend afternoon”). Очевидно, здесь действие управляет ситуацией. После того, как высказано однозначно намерение субъекта, дальнейшая информация становится пассивной, практически не влияет на выбор способа действия или усложняет его.

А тексте на татарском языке строится по схеме: S-O-V - сначала дается информация и ее анализ, и только после этого, возможно, с учетом реакции слушающего, определяется – положительное или отрицательное, само действие. («Min dustim belen toshten song bulasi “Atilla” kinosina baram/barmiym» – буквально: ‘Я со своим другом после обеда на фильм «Атилла» пойду/не пойду’).

В системах искусственного интеллекта это называется активностью знаний, что является одним из важных признаков интеллектуальности системы. Для подобных систем естественным и основополагающим является стиль размышления: анализ-действие, размышление-цели-алгоритмы, а не командный стиль: действие-анализ, алгоритм-цель, как это реализовано в современных технологиях, основанных на менталитете индо-европейских языков.

Такие возможности, естественные для татарского языка и закрепленные в его грамматике, позволяют ставить задачу о разработке

интеллектуальных технологий обработки информации на основе татарского языка.

4. Заключение

В статье изложены результаты деятельности института прикладной семиотики Академии наук РТ по созданию программных систем, технологий и лингвистических ресурсов с целью обеспечения паритетного функционирования татарского языка в инфокоммуникационных технологиях в качестве одного из государственных языков в Республике Татарстан. Разработанные институтом и представленные в статье программные продукты способствуют сохранению и развитию татарского языка, повышению его активности в киберпространстве и помогают татарскому языку стать языком компьютерных технологий, языком накопления, обработки, передачи информации в инфокоммуникационных технологиях, включая Интернет, социальные сети. Описан ряд потенциальных когнитивных возможностей татарского языка, которые показывают перспективу для татарского языка стать формальной базой для построения новых интеллектуальных технологий описания, хранения и обработки информации.

**Список литературы**

1. Правила образования, совершенствования и применения татарских терминов. Комитет при Кабинете Министров РТ по реализации Закона РТ “О языках народов РТ”. Зам. председателя Комитета профессор М.З. Закиев, председатель терминологической комиссии Комитета, доцент И.М. Низамов. – Казань, 1995. – 13с. (на татарском языке).

2. Сулейманов, Д.Ш. Англо-русско-татарско-чувашский словарь терминов по информатике и информационным технологиям (с толкованиями на татарском языке) / Д.Ш. Сулейманов, А.Ф. Галимянов, М.Х. Валиев, П.В. Желтов, М.П. Желтов, В.П. Желтов. // Приложение к Материалам III Международной конференции по компьютерной обработке тюркских языков (TurkILang 2015, Казань, 17-19 сентября 2015 г.). – Казань, Изд-во АН РТ, 2015. – 400 с.

**3.** Сулейманов Д.Ш., Галимянов А.Ф. Система татарских терминов в компьютерных технологиях и информатике // В сб. Трудов Первой международной конференции «Компьютерная обработка тюркских языков». – Астана: ЕНУ им. Л.Н. Гумилева, 2013. – С. 132-140.

4. Формальные модели и программные инструменты компьютерной обработки татарского языка / Р.Р. Гатауллин, А.Р. Гатиатуллин, Р.А. Гильмуллин, О.А. Невзорова, Д.Р. Мухамедшин, Д.Ш. Сулейманов, Б.Э.Хакимов, А.Ф, Хусаинов. Научное издание / Под редакцией Д.Ш. Сулейманова, А.Ф. Хусаинова. - Академия наук РТ, Институт прикладной семиотики АН РТ. – Казань: Изд-во Академии наук РТ, 2019. – 260 с.

5. Computational Models in Language and Speech/ Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020)/ Edited by Alexander Elizarov, Natalia Loukachevitch. - Kazan, Russian, November 12-13, 2020. (http://ceur-ws.org/Vol-2780/)

6. Ayrat Gatiatullin, Dzhavdet Suleymanov, Nikolai Prokopyev, Bulat Khakimov. About Turkic Morpheme Portal // in Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020). - Kazan, Russian, November 12-13, 2020. pp. 226-243 (http://ceur-ws.org/Vol-2780/).

7. Heintz J. and Schonig C. Turcic Morphology as Regular Language // Central Asianic Jornal (CFJ), 1989. -P.1-24.

8. Khusainov A., Suleymanov, D. An approach to automate process of creating speech analysis systems for under-resourced languages [Text] / A. Khusainov, D. Suleymanov // IEEE CPS Volume of Proc. of MICAI-2015. (Cuernavaca, October 25 to 31, 2015). – IEEE Computer Society, 2015. – P.28-34. [Tatar National Corpus] Tatar National Corpus [Электронный ресурс]. URL: http://tugantel.tatar/

9. Suleymanov D.Sh. Natural Cognitive Mechanisms in the Tatar language [Text] / D.Sh. Suleymanov // In the Collection of the Vienna Proceedings of the Twentieth European Meeting in Cybernetics and Systems Research. Ed. by Robert Trappel. Vienna, Austria, 6-9 April, 2010. – P.210-213.

***УДК. 004.8***

***Тукеев Уалшер Ануарбекович***
*Казахский Национальный Университет им. Аль-Фараби*
*Кафедра Информационных систем, д.т.н., профессор*
*Алматы, Казахстан*
*ualsher.tukeyev@gmail.com*

# ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ ПО ВЫЧИСЛИТЕЛЬНОЙ МОДЕЛИ МОРФОЛОГИИ НАОСНОВЕ ПОЛНОГО НАБОРА ОКОНЧАНИЙ

**Анотация.** В данной работе описывается вычислительная модель морфологии, основанная на полных наборах окончаний (CSE – Complete Set of Endings) для тюркских языков и текущее состояние использования данной CSE-модели морфологии для различных задач обработки тюркских языков. Предлагаемый подход позволяет пользователю (лингвисту) использовать универсальные (управляемые данными) алгоритмы и программы для ряда задач ОЕЯ, таких как определение основ слов (стемминг), морфологический анализ текста и сегментация текста. Одна из ключевых особенностей этого подхода заключается в том, что для нового языка только лингвистический ресурс этого языка в виде полной системы окончаний должен быть подготовлен в форме вычислительной реляционной модели данных. Затем используется универсальная программа для решения соответствующей задачи, основанная на разработанных данных. Рассмотрены такие задачи ОЕЯ как: стемминг, сегментация, морфологический анализ, машинный перевод текст-в-текст, машинный перевод речь-в-речь.

**Ключевые слова:** вычислительная модель, морфология, тюркские языки, обработка естественных языков

***UDC. 004.8***

***Ualsher Tukeyev***
*Al-Farabi Kazakh National University*
*Doctor of Technical Sciences*
*Professor of the Information Systems Department*
*Almaty, Kazakhstan*
*ualsher.tukeyev @gmail.com*

# TURKIC LANGUAGES PROCESSING ACCORDING TO THE COMPUTATIONAL MODEL OF MORPHOLOGY BASED ON A COMPLETE SET OF ENDINGS

**Abstract:** This paper describes a computational morphology model

based on Complete Set of Endings (CSE) for Turkic languages and the current state of use of this CSE morphology model for various tasks of processing Turkic languages. The proposed approach allows the user (linguist) to use universal (data-driven) algorithms and programs for a number of NLR tasks, such as stemming, morphological analysis and text segmentation. One of the key features of this approach is that for a new language, only the linguistic resources of this language in the form of a complete set of endings and corresponding tables should be prepared in the form of a computational relational data model. Then a universal program is used to solve the corresponding problem, based on the developed data. The following tasks of the NLR are considered: stemming, segmentation, morphological analysis, text-to-text machine translation, speech-to-speech machine translation.

**Keywords:** computational, model, morphology, Turkic, languages, processing, natural, languages.

**Введение**

Тюркские языки составляют семью, включающую более 35 языков [Дыбо, 2007], на которых говорят более 160 миллионов человек [Gutman and Avanzati, 2013] в нескольких странах. В тюркскую группу языков входят такие государственные языки, как азербайджанский, казахский, киргизский, узбекский, турецкий, туркменский. Языками субъектов государств являются алтайский, балкарский, башкирский, каракалпакский, крымскотатарский, кумыкский, ногайский, татарский, тувинский, уйгурский, хакасский, шорский, якутский.

В современном мире глобализация, языки и миграция создают серьезные проблемы для коммуникаций в обществе. Поскольку важность содействия эффективному общению между людьми резко возросла, искусственный интеллект в его различных формах рассматривается как ключевой фактор, способствующий распространению, обмену и доступу к знаниям в разных языках и культурах. Одной из передовых областей искусственного интеллекта является обработка естественного языка (ОЕЯ), в задачи которой, помимо прочего, входят: выделение основ слова (стемминг), морфологический анализ, сегментация текста, синтаксическая маркировка (POS – part of speech тегирование), машинный перевод, понимание языка, поиск информации, суммаризация (реферирование), извлечение информации (Information extraction).

Однако в мире есть тысячи языков с ограниченными ресурсами, не использующие технологии ОЕЯ. Сейчас в области ОЕЯ существуют две группы вычислительных моделей и методов для обработки языков: преобразователи с конечным числом состояний (FST – finite state

transducers) и методы машинного обучения. Группа методов FST требует использования ориентированного на пользователя языка программирования для описания исходных данных для новых языков, что непросто для лингвистов. Вторая группа методов, группа машинного обучения, требует большого объема электронных исходных данных для машинного обучения, чего нет для многих языков с ограниченными ресурсами.

В данной работе описывается новая вычислительная модель морфологии, основанная на полных наборах окончаний (CSE – Complete Set of Endings) для тюркских языков и текущее состояние использования данной CSE-модели морфологии для различных задач обработки тюркских языков. Предлагаемый подход позволяет пользователю (лингвисту) использовать универсальные (управляемые данными) алгоритмы и программы для ряда задач ОЕЯ, таких как определение основ слов (стемминг), морфологический анализ текста и сегментация текста. Одна из ключевых особенностей этого подхода заключается в том, что для нового языка только лингвистический ресурс этого языка в виде полной системы окончаний должен быть подготовлен в форме вычислительной реляционной модели данных. Затем используется универсальная программа для решения соответствующей задачи, основанная на разработанных данных.

**Обзор**

Существуют три общепринятые модели морфологии естественных языков [Spencer, 1991; Плунгян, 2003], а именно: «Item and Arrangement - Элемент и расположение» (IA- модель); «Item and Process – Элемент и процесс» (IP-модель); «Word and Paradigm - Слово и парадигма» (WP-модель).

IA-модель фокусируется на агглютинативном характере словоформ. Его основной инструмент моделирования выполняет линейную сегментацию словоформ на морфемы. Морфема — это минимальная значимая неделимая часть слова. Рассматривая морфемы как минимальные единицы грамматического описания, IA-модель хорошо подходит для описания морфологии агглютинативных языков.

IP-модель фокусируется на концепции динамической природы алломорфов, вводя один или несколько уровней представления словоформ. Каждая морфема словоформы обязательно имеет единственное глубокое представление, а также правила перехода к более поверхностным уровням представления с учетом контекста, при котором возможны алломорфные вариации представления морфемы.

WP-модель фокусируется на концепции флекции по парадигме. В этой морфологической модели слово рассматривается как единое целое, а не как комбинация основы и окончания. Флексия в WP-модели рассматривается по сходству, а минимальной единицей грамматического описания является словоформа.

Хорошо известными вычислительными моделями морфологии являются двухуровневые морфологии (TWOL) Киммо Косканиеми [Koskenniemi, 1983]. TWOL модель морфологии представляют слово на двух уровнях:

лексическое (глубокое) представление;

поверхностное представление.

Эта модель основана на TWOL правилах: преобразование лексического (глубокого) представления слова в поверхностное представление в зависимости от контекста. Для реализации этой технологии были разработаны программные средства, которые используются для многих языков. Для использования этих инструментов были разработаны специальные языки пользовательского интерфейса для исходных данных (правилатехнологии двухуровневой морфологии). Однако освоение и использование пользовательского языка для задания исходных данных для основанных на правилах методов, основанных на двухуровневой морфологии, - довольно трудоемкий процесс.

С точки зрения автора, это серьезное препятствие для широкого использования лингвистами технологий на TWOL правилах для стемминга, сегментации и морфологического анализа, особенно для языков с ограниченными ресурсами.

**Вычислительная CSE-модель морфологии агглютинативных языков**

Общеизвестные формы представления функции:

Аналитическое в виде формул: Y = F (X), F = 2.

Табличное в виде графика функции: F = 2

| Y = F (X) | |
|---|---|
| X | Y |
| 2 | 4 |
| 3 | 6 |
| … | … |

Наш подход к моделям основных задач ОЕЯ основан на табличном представлении функций.

Табличный подход лежит в основе реляционной модели данных, которая является универсальным подходом для моделирования реального мира в виде реляционных баз данных.

Построение CSE-модели морфологии и ее использование для анализа языка основаны на выводе полного набора окончаний для языка. Кроме того, в полной версии модели необходимо собрать набор основ слов данного языка, присоединяя к которым возможные окончания можно получить полный набор словоформ языка. Таким образом, для описания морфологии языка необходимо указать либо полный набор словоформ, либо полный набор окончаний и основ. Последний вариант, конечно, предпочтительнее, так как представление морфологии с полным набором окончаний и основ более экономично по объему описания, чем перечисление всех возможных словоформ [Булыгина, 1977]. Грамматический словарь Зализняка [Грамматический словарь Зализняка] можно отнести к морфологической модели перечисления словоформ.

Схема вывода полного набора окончаний для языка включает следующую четырехэтапную процедуру, основанная на комбинаторном подходе:

определение комбинации возможных размещений основных типов аффиксов;

выбор размещений основных типов аффиксов (осуществляется путем проверки их семантической приемлемости в языке);

перечисление возможных вариантов окончаний для каждого варианта семантически приемлемого размещения основных типов аффиксов;

объединение окончаний в полный набор окончаний для данного языка.

Рассмотрение построения CSE-модели морфологии на примере казахского языка и ряде других тюркских языков представлено в работах [Tukeyev and Karibayeva, 2020a; Tukeyev, Karibayeva, Zhumanov, 2020b; Toleush, Israilova, Tukeyev, 2021; Zhanabergenova and Tukeyev, 2021].

**Вычислительные модели данных для сегментации и морфологического анализа текстов на основе CSE-модели морфологии**

На основе предложенной CSE-модели для морфологии были построены вычислительные реляционные модели для сегментации текста и морфологического анализа.

Вычислительная реляционная модель данных для сегментации текста представляет собой реляционную (табличную) модель данных, состоящую из двух столбцов [Tukeyev, Karibayeva, Zhumanov, 2020b]. Первый столбец «окончаний» содержит полный набор языковых

окончаний, а второй столбец «сегментированных окончаний» содержит окончания языка, разбитые на аффиксы (Таблица 1).

Таблица 1. Вычислительная реляционная модель данных для сегментации казахских окончаний (фрагмент).

| **Окончания слов** | **Окончания как послед-тьаффиксов** | **Примеры** |
|---|---|---|
| gandarmensizder | gan@@dar@@men@@sizder | bar-gandarmensizder (you are withpeople who going) |
| largamyn | lar@@ga@@myn | apa-largamyn (I am to my sisters) |

Вычислительная модель данных для морфологического анализа представляет собой реляционную (табличную) модель данных, состоящую из двух столбцов. Здесь первый столбец «окончания» содержит полный набор окончаний языка, а второй столбец представляет собой последовательность морфологических характеристик, описывающих морфологический анализ соответствующих окончаний (Таблица 2). Для описания морфологических характеристик используются теги системы машинного перевода платформы Apertium [Apertium list of symbols].

Таблица 2. Вычислительная модель данных для морфологического анализа казахского языка (сегмент таблицы).

| **Endings** | **Morphological analysis** | **Comments** |
|---|---|---|
| largamyn | <NB>*lar<pl>*ga<dat>*myn<p1> | NB – nominal base type; pl - plural;dat - dative case; p1- 1-st person |
| gandarmensizder | <VB>*gan<pp>*dar<pl>*men<inst>*sizder<p2><frm> | VB-verbal base; pp-past participle;pl – plural; inst-instrumental case;p2 - 2nd person; frm- formality |

На основе предложенного подхода каждый новый язык будет иметь свои собственные вычислительные модели данных для сегментации и морфологического анализа.

**Вычислительные модели данных для машинного перевода на основе CSE-модели морфологии**

На основе предложенной CSE-модели для морфологии построены вычислительные реляционные модели и общий алгоритм для машинного перевода тюркских языков ($TURK^1$-> $TURK^2$).

Для данной задачи строится вычислительные реляционные модели (таблицы) соответствия стемов, стоп слов и окончаний языков $TURK^1$ и $TURK^2$.

Таблицы соответствия стемов и стоп слов языков $TURK^1$ и $TURK^2$ представляю собой соответствия лексиконов стемов и стоп слов (таблица 3 и 4).

Таблица 3. Таблица соответствия стемов слов языков $TURK^1$ и $TURK^2$

| Стемы языка $TURK^1$ | Стемы языка $TURK^2$ |
|---|---|
| $St1^S$ | $St1^T$ |
| $St2^S$ | $St2^T$ |
| … | … |

Таблица 4. Таблица соответствия стоп слов языков $TURK^1$ и $TURK^2$

| Стоп слова языка $TURK^1$ | Стоп слова языка $TURK^2$ |
|---|---|
| $Sw1^S$ | $Sw1^T$ |
| $Sw2^S$ | $Sw2^T$ |
| … | … |

Таблицы соответствия окончаний языков $TURK^1$ и $TURK^2$ представлены в таблице 5. Таблица 5. Таблица соответствия окончаний языков $TURK^1$ и $TURK^2$.

| Окончание языка $TURK^1$ | Морфологическое описание окончания $MDE^S$ | Морфологическое описание окончания $MDE^T$ | Окончание языка $TURK^2$ |
|---|---|---|---|
| $E1^S$ | $MDE1^S$ | $MDE1^T$ | $E1^T$ |
| … | … | … | ... |
| $En^S$ | $MDEn^S$ | $MDEm^T$ | $Em^T$ |

Морфологическое описание окончаний необходимо для проверки соответствия окончаний

TURK1 и TURK2, но в процессе вычислений не участвует.

Описание общего алгоритма для машинного перевода тюркских языков (TURK1 -> TURK2) представляется в следующем виде.

Дано: предложение тюркского языка TURK1 (S (source) – предложение исходного языка).

Получить: предложение тюркского языка TURK2 (T (target)– предложение целевого языка).

Шаги алгоритма:

1. Взять текущее слово wi предложения S.

2. Выполнить стемминг слова wi : StiS + EiS, где StiS – стем слова wi, EiS- окончание слова wi.

3. Окончание текущего слова исходного языка EiS ищется в столбце окончаний языка TURK1 и находятся все строки, где это окончание имеется.

4. Находятся стемы StiT целевого языка TURK2 , соответствующие стему StiS исходного языка TURK1 по таблице соответствия стемов языков TURK1 и TURK2.

5. Находится окончание EiT целевого языка TURK2, соответствующие найденному стему StiT, в соответствии правилам сингармонии целевого языка TURK2.

6. Стем StiT соединяется с окончание EiT, получая таким образом слово целевого языка TURK2, которое будет эквивалентом слова wi предложения исходного языка TURK1.

7. Если не встречается конец предложения, то перейти на п.1, иначе п.8.

8. Анализ следующего предложения текста. Если конец текста, то п.9.

9. Конец.

По данной технологии машинного перевода на основе CSE-модели морфологии для тюркских языков выполнены эксперименты для пар языков казахско-турецкий, казахско-татарский, казахско-узбекский, показавшие оценку по метрике BLEU в пределах 20%.

**Машинный перевод речь-в-речь на основе CSE-модели морфологии**

Проблема параллельных корпусов для обучения еще более актуальна для машинного перевода речь-в-речь S2ST (Speech to Speech Translation) так, как в Интернете такие электронные базы практически не существуют в отличии от текстовых параллельных корпусов, которые

изначально были на сайтах парламентов различных государств. Поэтому параллельные корпуса для S2ST (Speech to Speech Translation) создаются специально именно для этой задачи и требуют специального оборудования для записи речи и значительных финансовых затрат. В направлении машинного перевода речи в речь S2ST ведутся интенсивные исследования с применением технологий обучения нейронных сетей.

Для тюркских языков исследования машинного перевода речь-в-речь S2ST практически отсутствуют в силу вышесказанных трудностей создания параллельных корпусов речь-в-речь S2ST для обучения. Создание машинного перевода речь-в-речь S2ST по каскадной схеме (ST – TT – TS), где S – речь (speech), T – текст (text), также затруднено в силу отсутствия фазы TT для большинства тюркских языков. Поэтому в данном работе ставится задача построения машинного перевода речь-в-речь S2ST по каскадной схеме, где фазу TT предлагается решать для тюркских языков на основе новой модели морфологии по полной системе окончаний (CSE-модели), что практически эквивалентно полному перебору правил морфологии, позволяющее максимально гарантировать анализ любого слова текстов рассматриваемых тюркских языков. Так как особенность языков тюркской группы такова, что синтаксической структуры предложений в этих языках схожи, то основные проблемы машинного перевода языков данной группы будут находиться в области морфологии. Это направление описывается в виде постановки задачи для дальнейшего направления исследований применения технологии CSE-модели морфологии для актуальной проблемы обработки естественных языков.

**Универсальные алгоритмы и программы обработки тюркских языков на основе CSE-модели морфологии**

На основе CSE-модели морфологии разработаны алгоритмы и программы стемминга слов без словаря (lexicon-free stemming), алгоритмы и программы стемминга слов с словарем стемов, алгоритмы и программы стемминга слов с словарем стемов и словарем стоп-слов [Tukeyev, Karibayeva, Turganbayeva, Amirova, 2021]. На основе CSE-модели морфологии разработаны алгоритмы и программы сегментации текстов, морфологического анализа. Описание и исходные тексты разработанных алгоритмов и программ как open source ресурс находятся на github платформе, могут использоваться по лицензии CC BY-SA [NLP-KazNU].

**Результаты и эксперименты**

CSE-модель морфологии разработана для казахского, кыргызского, узбекского, древнетюркского, турецкого языков.

Эксперименты на казахском, киргизском и древнетюркском языках проводились для стемминга и сегментации, на узбекском и турецком - для стемминга.

В целом точность определения стемов и сегментации слов составила 85-95%.

В настоящее время наши магистры образовательной программы «Компьютерная лингвистика» проводят исследования на основе CSE-модели для башкирского, каракалпакского, татарского языков для задач стемминга, сегментации, морфологического анализа и машинного перевода

**Заключение и будущие работы**

Преимущество предлагаемой методики в том, что она ориентирована на лингвистов.

Для решения задач стемминга, сегментации, морфологического анализа требуется только:

- создание полного набора языковых окончаний для стемминга;
- построение таблицы сегментации концовок для задачи сегментации;
- построение таблицы морф-анализа концовок для задач морф-анализа;
- используется подходящая универсальная программа.

Результаты экспериментов показали по всем рассматриваемым задачам обработки сравнимые показатели с результатами методов, основанных на нейронных технологиях, требующих достаточно больших объемов корпусов для обучения.

Дальнейшие работы запланированы в направлении:

- повышение эффективности разработанных алгоритмов и программ;
- использования предложенной методики для других языков тюркской группы и задач ОЕЯ, в частности, машинного перевода «текст-текст» и «речь-в-речь».

**Список литературы**

1. Булыгина Т. В. (1977). Проблемы теории морфологических моделей. Наука. Москва, 288 с. Грамматический словарь Зализняка. https://gufo.me/dict/zaliznyak

2. Дыбо А.В. (2007). Хронология тюркских языков и лингвистические контакты ранних тюрков. http://s155239215.onlinehome.us/turkic/40_Language/Dybo_2007LingivistContactsOfEarlyTurksRu.htm

3. Плунгян В.А. (2003). Общая морфология: Введение в проблематику: Учебное пособие. Изд. 2-е, исправленное. — М.: Едиториал УРСС, 2003. - 384 с.

4. Apertium list of symbols. https://wiki.apertium.org/wiki/List_of_symbols

5. Gutman A. and Avanzati B. (2013). The languages gulper. Turkic languages. http://www.languagesgulper.com/eng/Turkic.html.

6. Koskenniemi K. (1983). Two-level morphology: A general computational model of word-form recognition and production. Tech. rep. Publication No. 11. Department of General Linguistics. University of Helsinki.

7. NLP-KazNU. - url: http://github.com/NLP-KAZNU

8. Spencer A. (1991). Morphological theory. An Introduction to Word Structure in Generative Grammar. Blackwell Publishers. pp.512

9. Tukeyev U., Karibayeva A. (2020a) Inferring the Complete Set of Kazakh Endings as a Language Resource. In: Hernes M., Wojtkiewicz K., Szczerbicki E. (eds) Advances in Computational Collective Intelligence. ICCCI 2020. Communications in Computer and Information Science, vol 1287, pp.741-751. Springer, Cham. https://doi.org/10.1007/978-3-030-63119-2_60

10. Tukeyev U., Karibayeva A., Zhumanov Zh. (2020b) Morphological Segmentation Method for Turkic Language Neural Machine Translation. Cogent Engineering, Volume 7, 2020 - Issue 1 https://doi.org/10.1080/ 23311916. 2020.1856500

11. Toleush A., Israilova N., Tukeyev U. (2021) Development of Morphological Segmentation for the Kyrgyz Language on Complete Set of Endings. In: Nguyen N.T., Chittayasothorn S., Niyato D., Trawiński B. (eds) Intelligent Information and Database Systems. ACIIDS 2021. Lecture Notes in Computer Science, vol 12672. Springer, Cham. pp.327-339. https://doi.org/10.1007/978-3-030-73280-6_26

12. Tukeyev U., Karibayeva A., Turganbayeva A., Amirova D. (2021) Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words. In: Nguyen N.T., Iliadis L., Maglogiannis I., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, vol 12876. Springer, Cham. https://doi.org/10.1007/978-3-030-88081-1_48

13. Zhanabergenova D., Tukeyev U. (2021) Morphology Model and Segmentation for Old Turkic Language. In: Nguyen N.T., Iliadis L., Maglogiannis I., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, vol 12876. Springer, Cham. https://doi.org/10.1007/978-3-030-88081-1_47

# ТҮРКІ ТІЛДЕРІ – ЖАҢА ИНТЕЛЛЕКТУАЛДЫ ТЕХНОЛОГИЯЛАР МЕН БІЛІМДЕРДІ ӨҢДЕУ ЖҮЙЕЛЕРІН ҚҰРУДЫҢ НЕГІЗІ РЕТІНДЕ

# ТЮРКСКИЕ ЯЗЫКИ КАК ОСНОВА ДЛЯ СОЗДАНИЯ НОВЫХ ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЙ И СИСТЕМ ОБРАБОТКИ ЗНАНИЙ

# TURKIC LANGUAGES AS THE BASIS FOR THE CREATION OF NEW INTELLIGENCE TECHNOLOGIES AND KNOWLEDGE PROCESSING SYSTEMS

*УДК 004.891*

***Исмаилов Исмаил Ариф оглы***
*Азербайджанский архитектурно-строительный университет,*
*Баку, Азербайджан*
*isi.isiev@mail.ru*

## ПРИМЕНЕНИЕ СТРУКТУРНОЙ ЭКСПЕРТНОЙ СИСТЕМЫ В ЭТИМОЛОГИЧЕСКИХ ИЗЫСКАНИЯХ

**Аннотация.** Предлагается новый метод структурной разработки групп экспертных под-систем (под-ЭС). В качестве приложения метода демонстрируется разработка структурной ЭС, для задачи“ Установления этимологий огузских этнонимов”. Для решения данной задачи нужно решать исторические, лингвистические, географические и другие подзадачи, т.е. нужны знания историков, лингвистов по разным языкам, специалистов по фольклору, мифологии, литераторов, географов. Предлагаемая система конструируется из трёх Под-ЭС. Это “Иноязычная под-ЭС” обрабатывающая Хотанские и Китайские тексты в предметной области, “Тюркская под-ЭС” (главная под-ЭС системы) и “Нелингвистическая нечёткая под-ЭС”, в которой оценивается нечёткость выводов Тюркской под-ЭС. Тюркская под-ЭС состоит из следующих блоков: База данных Тюркских этнонимов, База лингвистических знаний (с историко-фонетическими правилами), Машина логического вывода, Блок выходных результатов под-ЭС (Предлагаемые этимологии/этимоны целевых этнонимов – 22-х огузских этнонимов списка Махмуда Кашгари). В работе подробно

описаны также другие функциональные блоки Тюркской Под-ЭС с представлением некоторых результатов тестирования прототипа данной Под-ЭС.

**Ключевые слова:** этнонимы, экспертная система, база знаний, историческая фонетика, этимология

*UDC 004.891*

***Ismayilov Ismayil Arif oglu***
*Azerbaijan University of Architecture and Civil Engineering*
*Baku, Azerbaijan*
*isi.isiev@mail.ru*

## APPLICATION OF THE STRUCTURAL EXPERT SYSTEM IN ETYMOLOGICAL RESEARCH

**Abstract.** The paper proposes a structural model for the development of an expert system for solving the problem of establishing the correct etymologies of the Oghuz ethnonyms. To solve this problem, it is necessary to solve historical, linguistic, geographical and other subtasks, i.e. knowledge of historians, linguists in different languages, specialists in folklore, mythology, writers, geographers, etc. is needed. For the correct linguistic reconstruction of ethnonyms, a large amount of non- linguistic information is required. The names of 22 Oghuz ethnonyms from the list of M. Kashgari from the 11th century book Divanu Lugat at Turk are used as initial data. Extracted from medieval Khotanese texts and Chinese chronicles, Turkic ethnonyms are a source of information for the knowledge base of the “Foreign Sub-ES”. With the help of knowledge base rules, ethnonyms are reconstructed from foreign language forms to their original Turkic form. After the reconstruction, the Turkic ethnonyms come to the entrance of the “Turkic Sub-ES”. The main source of input information for the "Turkic Sub-ES" are the Old Turkic and Uighur texts of the 8th - 9th centuries AD. The Turkic ethnonyms extracted from the texts together with the reconstructed ethnonyms from the “Foreign Sub-ES” are facts for the knowledge base of the “Turkic Sub-ES” and are accumulated in the database of ethnonymic data, which then, after processing (dividing into proper ethnonyms and affixes or ethnonym-forming formants, and also if the ethnonyms are two or more composite, then dividing them into components) come in the form of subject facts-ethnonyms to the knowledge base of the “Turkic Sub-ES”. The knowledge base of the Turkic Sub-ES is equipped with historical phonetic rules, some of which are given in the work. To

represent knowledge in the knowledge base, we have chosen a model of reverse logical inference. The inference achine generates logical conclusions - supposed prototypes for twenty-two Oghuz ethnonyms. The results of the output engine in the form of etymological chains (sequences of phonetic and/or grammatical changes of the etymon on the way to the target ethnonym of the M. Kashgari list) are sent to the block “Output results of Sub-ES (Proposed etymologies/etymons of target ethnonyms)”. The non-linguistic information that is used in the system as arguments for or against the results obtained from the two previous linguistic Sub-ESs is accumulated in the “Non-Linguistic Data Base”. The fuzziness of the conclusions of the first two Sub-ES will be taken into account in the proposed structural ES in the “Non-Linguistic Sub-ES”.

**Keywords:** ethnonyms, expert system, knowledge base, historical phonetics, etymology

**Введение и постановка задачи**

В современной практике использования группы Экспертных Систем (ЭС - компьютерные программы, заменяющие экспертов в предметных областях) для решениятсложных междисциплинарных задач имплицитно присутствует идея разбивки сложной задачи разработки большой ЭС на множество частных разработок мелких Под-ЭС, для решения подзадач общей сложной глобальной проблемы в предметной области. В рамках структурной модели разработки облегчается процесс разработки ЭС, так как инженеры знаний могут абстрагироваться от главной цели и других под-целей и концентрироваться на задаче построения конкретной Под-ЭС. Кроме того возможна одновременная параллельная разработка нескольких Под-ЭС группой инженеров знаний.

Нами предпринята попытка структурной разработки экспертной системы в специальной междисциплинарной предметной области, где наиболее ярко проявляются преимущества структурного подхода к разработке ЭС, а именно для решения задачи установления правильных этимологий Огузских этнонимов. Для решения таких задач нужно решать исторические, лингвистические, географические и другие подзадачи, т.е. нужны знания историков, лингвистов по разным языкам, специалистов по фольклору, мифологии, литераторов, географов и других.

**Решение**

Этимологизировать – значит устанавливать первоначальное значение слова, т.е. отыскивать исходное слово (этимон), от которого

произошло рассматриваемое слово [Введенская, 2004, с. 10]. Этимология собственных имён, к которым в частности относятся этнонимы (наименования родов, племён и народов), отличается от этимологии нарицательных слов большей сложностью. Основная часть собственных имён образована не непосредственно от имён нарицательных, а от других собственных имён, более ранних по времени своего возникновения [Суперанская, 1986, с. 81]. Этимологическому исследованию свойственна множественность возможных решений, проблематичность, гипотетичность [Введенская, 2004, с. 13].

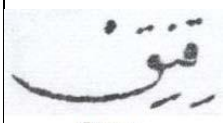
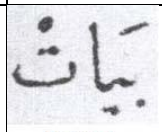
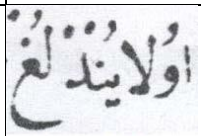
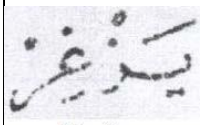
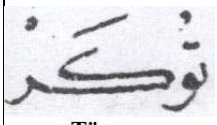
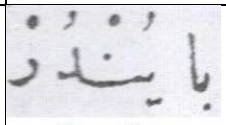
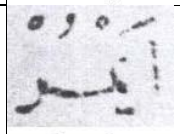
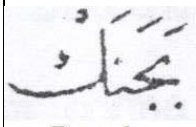
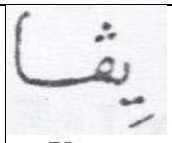
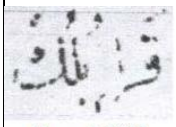
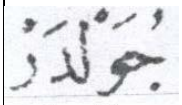
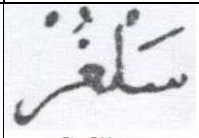
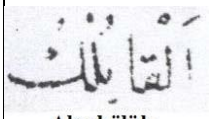
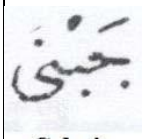
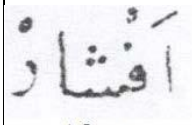
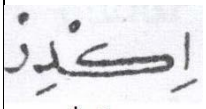
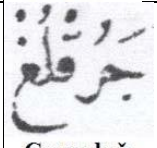
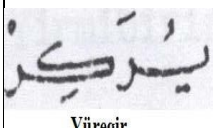
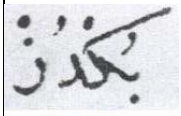
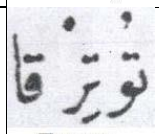
Для решения таких задач нужно решать исторические, лингвистические, географические и другие подзадачи, т.е. нужны знания историков, лингвистов по разным языкам, специалистов по фольклору, мифологии, литераторов, географов и т.д. Для правильной лингвистической реконструкции этнонимов требуется большая нелингвистическая информация (предлагается нами, как обобщающий термин для обозначения информации извлекаемой из данных множества других вышеперечисленных наук).

Таким образом, для построения ЭС с целью установления этимологии этнонимов недостаточно построения одной ЭС, необходима разработка группы Под-ЭС в рамках единой ЭС извлекающих и обрабатывающих информации из разных наук, а группа предполагает структуру, где отдельные локальные Под-ЭС должны каким-то образом взаимодействовать друг с другом для достижения общей цели – установления правильной этимологии огузских этнонимов.

Наименования 22 огузских этнонимов приводятся в нескольких источниках, самым ранним (70– е годы XI века нашей эры) из которых является список М. Кашгари [Ауэзова, 2005, с. 93-94], который в оригинальной арабской графике и латинской транскрипции приведён в таблице 1.

Относительно этимологий огузских этнонимов отдельными учёными (Баскаков Н.А., Кононов А.Н., Толстов С.П., Кумеков Б.Е., Плетнёва С.А., Махпиров В.У., Зуев Ю.А. и другие) (будем считать их экспертами в своей области) – историками, лингвистами, или этнонимиками и т.д. были предприняты узконаправленные исследования, и поэтому страдают однобокостью, не имеют достаточно веских аргументов в пользу предлагаемых этимологий, по мнению большинства других экспертов.

**Таблица 1.**
**Огузские этнонимы из списка М. Кашгари**
**Oghuz ethnonyms from the list of M. Kashgari**

| № | Этноним | № | Этноним | № | Этноним |
|---|---|---|---|---|---|
| 1 |  Qınıq | 9 |  Bayat | 17 |  Ulayundluğ |
| 2 |  Qayığ | 10 |  Yazğır | 18 |  Tüger |
| 3 |  Bayundur | 11 |  Eymür | 19 |  Beçenek |
| 4 |  Yıva | 12 |  Qarabölük | 20 |  Çuvaldar |
| 5 |  Salğur | 13 |  Alqabölük | 21 |  Cebni |
| 6 |  Afşar | 14 |  İgdir | 22 |  Çaruqluğ |
| 7 |  Begtili | 15 |  Yüregir | | |
| 8 |  Bügdüz | 16 |  Tutırqa | | |

Учитывая мысль известного эксперта-ономаста Суперанской «Основная часть собственных имён образована от других собственных имён, более ранних по времени своего возникновения» [Суперанская, 1986, с. 81], а также тот исторический факт, что ещё до огузов списка М. Кашгари по крайней мере с VII в. нашей эры известно существование союза девяти огузских племён и других тюркских племён, авторы статьи пришли к заключению, предположить главным этнонимообразующим принципом происхождения огузских этнонимов их происхождение от более ранних огузских и других тюркских этнонимов, т.е. “отэтнонимность”.

На рисунке 1 представлена детальная блок-схема предлагаемой структурной ЭС.

Извлечённые из хотанских текстов VIII – IX и китайских хроник той же эпохи тюркские этнонимы являются источником информации (предметными фактами) для базы знаний “Иноязычной ЭС”, которая снабжена специальными экспертными правилами. С помощью этих правил происходит реконструкция этнонимов из иноязычных (Хотанского и Китайского языков) форм к их исконной тюркской форме. После реконструкции тюркские этнонимы поступают на вход “Тюркской Под-ЭС”. Основным источником входной информации для “Тюркской Под-ЭС” являются древнетюркские и уйгурские тексты VIII – IX веков нашей эры. Извлечённые из текстов тюркские этнонимы вместе с реконструированными этнонимами из “ Иноязычной Под-ЭС” являются предметными фактами для базы знаний “ Тюркской Под- ЭС” и аккумулируются в базе этнонимических данных, которые затем после обработки (разделения на собственно этнонимы и аффиксы или этнонимообразующие форманты, а также если этнонимы дву или более составные, то разделение их на составляющие) поступают в виде предметных фактов-этнонимов в базу знаний “Тюркской Под-ЭС”.

База знаний Тюркской Под-ЭС снабжена историко- фонетическими правилами [Абдуллаева, Исмаилов, 2016, с. 127- 128], некоторые из которых приводятся далее. В работе для представления знаний в базе знаний Под-ЭС нами выбрана логическая модель, точнее как более рентабельный обратный логический вывод,[Уотермен, 1989, с. 75].

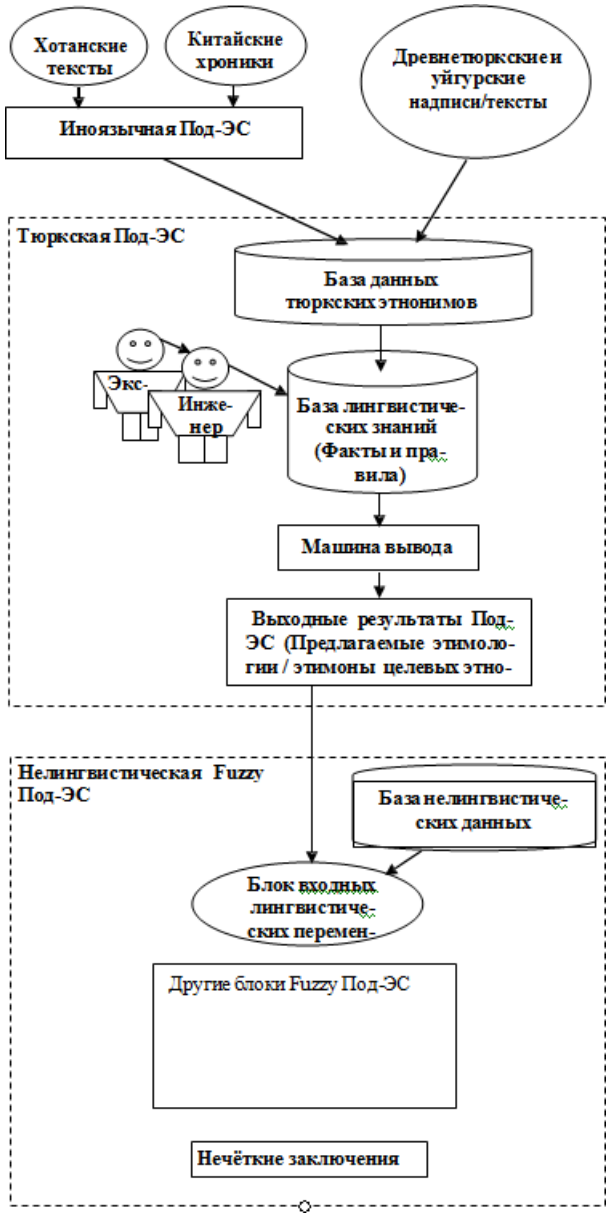



*Рис. 1. Блок-схема предлагаемой структурной ЭС. Block diagram of the proposed structural ES.*

Правило-1:
*THEN b ~ m*
*IF (“b” OR “m”) AND “середина слова”.*
где “ ~ ” соответствие или чередование звуков.

Правило-2:
*THEN d > y*
*IF “d” AND “середина слова” AND “d после гласного”.*
где “ > ” переход звука.

Правило-3:

*THEN u ~ ï*
*IF (“u” OR “ï”) AND “первый слог”.*

Правило-4:
*THEN ä ~ i*
*IF (“ä“ OR “i”).*

Правило-5:
*THEN r > l*
*IF “r” AND “конец слова”.*

Правило-6 (У Махмуда Кашгари):
*THEN t > d IF “t”.*

Правило-7:
*THEN ä ~ a*
*IF (“ä“ OR “a”).*

Этнонимообразующие аффиксы огузских этнонимов и этнонимы-композиты учитываются в базе знаний с помощью специальных процедур обработки этнонимов. На основе предметных фактов и экспертных правил, машина логического вывода формирует логические заключения – предполагаемые прототипы для 22 огузских этнонимов списка М. Кашгари. Результаты работы машины вывода в виде этимологических цепей (последовательностей фонетических и/или грамматических изменений этимона на пути к целевому этнониму списка М. Кашгари) поступают в блок «Выходные результаты Под-ЭС (Предлагаемые этимологии/этимоны целевых этнонимов)».

Нелингвистическая информация, которая используются в системе в качестве аргументов за или против результатов, получаемых из двух предыдущих лингвистических Под-ЭС накапливается в “Базе Нелингвистических данных”.

Как известно, этимология собственных имён, в частности этнонимов не является точной наукой, ей свойственны нечёткость, искажения, случайность, нерегулярность фонетических явлений и т.д. Нечёткость выводов двух первых Под-ЭС будет учитываться в предлагаемой структурной ЭС в “Нелингвистической Fuzzy Под- ЭС”. Нечеткие (Fuzzy) экспертные системы, как известно, не только весьма полезны в нечётких, приблизительных, гипотетических предметных областях, но также обладают лучшей по сравнению с обычными чёткими (Crisp) ЭС способностью представлять мнения многих

экспертов, порой не схожих и даже прямо противоположных [Negnevitsky, 2005, p.16]. В настоящее время “Fuzzy Под-ЭС” находится в стадии разработки. В процессе тестовых испытаний прототипа «Тюркской Под-ЭС» были получены представленные в таблице 2 предварительные результаты.

Таблица 2.Предварительные результаты тестирования Тюркскойпод-ЭС

**Preliminary test results of Turkic Sub-ES**

| **№ в списке Кашгари** | **Целевые эт-нонимы** | **Предлагаемый ЭС этимон** | **Правила Базызнаний** | **Компонент этнонима** | **Формант** |
|---|---|---|---|---|---|
| 1 | Qïnïq | Qunï | Правило 3 | - | -q |
| 5 | Salğur | Sir | Правило 4<br>Правило 7<br>Правило 5 | - | -ğur |
| 12 | Qarabölűk | bölűk | Спец. процедура | qara | - |
| 13 | Alqabölűk | bölűk | Спец. процедура | alqa | - |
| 14 | Ígdir | Ígdər | Правило 4 | - | - |

**Выводы**

1) Предлагается новый структурный метод разработки для группы ЭС для решения проблем в сложных, междисциплинарных предметных областях, который заключается в планировании разработки ЭС как совокупности взаимосвязанных Под-ЭС;

2) В качестве приложения структурного метода разработки ЭС, предлагается разработка “Структурной экспертной системы установления этимологий Огузских этнонимов”;

3) Описываются отдельные локальные Под-ЭС и базы знаний разрабатываемой экспертной системы;

4) Представлены предварительные результаты тестирования прототипа локальной “Тюркской Под-ЭС” для нескольких целевых этнонимов.

**Список литературы**

1. Л.А. Введенская, Н.П. Колесников, Этимология: Учебное пособие.-СПб.: Питер, 2004.-221 с.

2. А.В. Суперанская, Теория и методика ономастических исследований. “Наука”: Москва, 1986, 255 с.

3. Махмуд ал-Кашгари Диван Лугат ат-Турк / Перевод, предисловие и комментарии З.-А. М.Ауэзовой.-Алматы: Дайк-Пресс, 2005. - 1288с.

4. Абдуллаева Г.Г., Исмаилов И.А. Конструкция батареи экспертных систем для установления этимологий этнонимов (на примере огузских этнонимов) // Transactions of Azerbaijan National Academy of Sciences. Series of Physical-Technical and Mathematical Sciences. Informatics and Control Problems, V. XXXVI, 2016, № 3, p. 123 - 130.

5. Уотермен Д. Руководство по экспертным системам. М.: Мир,1989,388 с.

6. Michael Negnevitsky, Aritificial Intelligence, Addison-Wesley. England. 2005, 407 p.

7. Исмаилов И.А. Моделирование лингвистических явлений в экспертной системе с помощью исчислений предикатов первого порядка // Известия Азербайджанского Национального Аэрокосмического Агентства. 2018, № 3 (21), том 21, с. 34 – 43.

8. Исмаилов И.А. Разработка структурной экспертной системы // Вестник Компьютерных и Информационных Технологий. Москва, 2018, № 10, с. 48 – 58.

*UDC 004*
***Muratbekova Sh.***
*Tashkent State University of Uzbek language and literature named after Alisher Navoiy*
*Tashkent, Uzbekistan*
*shoira.m96@gmail.com*

# HOW TO CREATE TEXT VALIDATION SOFTWARE FOR A PLUGIN

**Abstract.**This article is written about the articles and programs used to determine the quality of plagiarism in the work of various diplomas and the stages of their work. In this article, various instructions are given about the differences of different programs from each other.

**Key words:** text, article, plagiarism, Anti-plagiarism programs

How difficult it is to bring the course work or essay in ETXT to the desired percentage. Students will "sweat" for hours on their texts in order to somehow increase their individuality and find their work successful. Well, if the "teacher" requires 60-70% - this can still be achieved, although you will have to spend enough time, and it depends on where the material for this work is obtained. What if the teacher makes 90% of the demand? It's a guard, friends!

How to cope with this disgrace and overcome the unfortunate anti-plagiarism? Who thought about all this and why? - such questions "arise" in the minds of thousands of students, the antiplagiate once again shows the uniqueness of 30-40%.

Remember that for each anti-plagiarism there is an anti-plagiarism! :)

Let's go directly to the topic of the question and determine it, how to cheat against plagiarism in 2019-2020 years on the examination of course work, diploma or essay. - it is very realistic to do and solve all this with "little blood".

It is necessary to activate the internal reserves of the dictionary and start "creating". It will be necessary to reconstruct each sentence in the text. We replace other people's thoughts with our own, choose synonyms. We leave only quotes and definitions from the original text. We practically create new work, change the structure. Yes, the work is laborious, but the specificity can be significantly increased. In this way, from the downloaded course or Diploma it is possible to make almost an author's work. The main thing is that the text in the work is meaningful and correctly formatted.

It takes a lot of time and effort.

The teacher can check the work for rewriting and find out where the material came from.

This method is one of the best if you approach text processing "with your head".

This is different from the first recommendation, because we mainly use synonyms and to a lesser extent use common sense. If in the full processing of the text you need to use your own opinion and knowledge on the subject of the work, then for superficial rewriting it is customary to use the following:

- replacement with synonyms
- access controls, use of revolutions
- replacement of phrases, words
- change the structure of text, paragraphs and sentences

Because this method of defeating Anti-plagiarism programs is not easy to call. They (programs) "become smarter" every year and very well consider rewriting. In order not to get caught up in the examination, it is necessary to apply the listed methods of making the course and other work unique together. For example, Text.ru or when checking the plugin on Content-Watch, only the change of phrases is immediately detected, and the text removed in the Advego program is highlighted in blue.

Below is a screenshot of the ETXT program, in which you can see that the rewrite is already being checked. [1]

If the work has at least 30-40% specificity and you need to reach 70-80%, this recommendation can work well - this is a very real opportunity to increase the specificity. Bypassing Anti-plagiarism in this way even applies to a weak C Class Student. But patience and willpower are required in any case.

Previously course works and essays were very easy a find.! I replaced Russian letters in words with English letters and now you have 100% uniqueness.

Now this does not work otherwise. The free gift is over :) anti-plagiarism free cheating is becoming increasingly difficult.

The method of diluting sentences with Epithets, introductory words and phrases also does not work (and the abundance of introductory words in the educational work does not always seem appropriate).

Choosing synonyms is a very good method, but with a lot of scientific and professional terms it is not so useful. And of course, if you use a stupid automatic replacement, the teacher will laugh at the "crazy" text.

Let's summarize what we do not need to spend time:

- Replacing Cyrillic letters with Latin letters
- Add introductory words

- Synonyms in the plural

When defining the MS WORD page, you can set automatic word wrapping. How does this help you cheat against plagiarism? If you antiplagiat.ru the automatic word wrap function on the site is enabled if you paste or copy from the text, the system will accept some words as unique because the words are partially cut.

Note: in this way, you can increase the specificity by no more than 2-7 percent. If you are a little lacking to go through the anti-plagiarism check, then you might want to try this method out. [2]

The essence of this method of bypassing Anti-plagiarism is that you need to search for material in English, Ukrainian or any other language, and then translate the material into Russian. In order to correctly format the text, it will be necessary to correct the material obtained in the Russian language. There is such a variant of anti-plagiarism free deception, but it is also difficult to call it free, because. The loss of time for translation and correction is also a type of payment. How to check the text for plagiarism is unknown, because you can find out. The correct case, and after the translation of the text, it will be mastered from another source in Russian. Each student chooses the method that suits him.

Gap is talking about scientific work that is not indexed by search engines. If already familiar with this work, you can "bite". But if you carefully copy and dilute the texts with your personal thoughts, the result can be very good.

Where to search for indexed texts? It can be:

- Foreign sites
- Translation of foreign works and articles
- Dissertations from paid directories
- Materials from the library have not yet been digitized

Note: Antiplagiarism. The VUZ system can check texts not only on Yandex and Google, but also from closed sources. Keep this in mind and try to find a way to pre-test your work against closed sources.

If you do not have time to independently correct the downloaded course works, diplomas, essays, you can use the services of our service.

The system technically increases the uniqueness of the work (at the document code level).

The text in the document does not change (visual), but Antiplagiat.ru, when checked by the university or ETXT program, the specificity will meet your requirements. [3]

For example, you can upload the downloaded finished work to the system with 5% uniqueness. To increase the specificity, the desired value, for example, up to 80-100%, it is necessary:

1. Upload the file to our system with the finished work
2. Select the desired processing system and specify
3. Wait for your text to be processed online

AntiplagiatExpress.ru - this is a student aid service. The developers note that it is able to increase the uniqueness of any text in order to pass anti-plagiarism.ru, ETXT, Advego and even university anti-plagiarism checks. If you urgently need to do course work or a unique diploma, they will help you in this regard with a modest fee. According to my observations, the prices here are cheaper than similar services, in addition, it works offline and you can increase the uniqueness by in files loading without external assistance, immediately after payment you will receive a link to download again. On it, if necessary, you can adjust% rarity in any direction. From specialists:

Flexible tariff for each page from 9 to 15 rubles (depending on the number of pages in the document);

Result guarantee;

Night time online customer support (they solve problems very quickly);

if it is certified and approved by payment systems, payment is officially accepted, that is, there are real guarantees of service delivery.

You can download the program from the link: Advego Plagiatus. Copy the text to it and perform a uniqueness check. I advise you to choose an in-depth examination, because it is a complete reflection of reality. Sometimes there may be requests for input, because search engines do not like scraping. Enter them manually or use the angitate service and so on.

This application is located at the following address. It is also sufficient to copy the text here and carry out the verification. It is possible to perform a batch scan of several files or the entire site.

For the most accurate result, you can use both programs together.

Today I'll tell you about how to check the anti-plagiarism text on the Internet. To start talking about this, you need to have an understanding of plagiarism in general. Plagiarism is the copying of someone's thoughts without a link to the author.

In the works written by you, plagiarism cannot be allowed, whether it will be a course, a diploma, an article. If someone else's opinion is used in the work, you can simply refer to it. Antiplagiate is a service that, if interpreted, struggles with non-unique data.

**CONCLUSION**

An anti-plagiarism system is an interesting thing and many do not like it. The fact is that many people do not like to write on their own, it's easier for them to take and copy. With such texts it does not work, so not everyone likes it. In fact, there are a lot of such services on the Internet. They check the text and give a certain percentage of originality.

**Reference**

1. Edelstein O.A., Shobolova L.P., Makarov Yu.S. Changes in the properties of rocks under the influence of adsorption-active media. On Sat. Issues of managing the state of the mountain range. 1984, M, You p.224, S.80-86.

2. Khalimov I., Sharafutdinov U.Z., Yuldoshev A.S., Avezova D.A. Modeling of processes of underground leaching, hydraulic fracturing and clogging. Mountain Bulletin of Uzbekistan. No. 2 (81) 2020 From 8-11.

3. Alikulov Sh.Sh. Khalimov I.U., Khamidov S.B., Alimov M.U. Intensification of the parameters of underground leaching of uranium from low-permeability ores on the example of uranium deposits. UNIVERSUM: TECHNICAL SCIENCES. 2020, pp. 57-62.

*УДК 621.374*
***Балсаидов А. Ш.***
*Алматинский Технологический Университет*
*Алматы, Казахстан*
*balsaidov@gmail.com*

# ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ИЗОБРАЖЕНИЙ ДЛЯ НАИЛУЧШЕГО РАСПОЗНАВАНИЯ ТЕКСТА

**Аннотация.** В данной статье я рассматриваю процесс бинаризации изображений. Приведены различные методы бинаризации, такие как, метод Otsu, бинаризация нижнего порога, бинаризация верхнего порога, бинаризация с двойным ограничением, метод Ниблека, метод Саувола, метод Бернсена. Также рассматривал вопрос о достоинствах и недостатках каждого из них, а также примеры использования, того или иного метода бинаризации на исходном изображении. Также я провел распознавание текста с различными уровнями помех. Для распознавания текста с обработанного изображения использовалась библиотека Tesseract.

**Ключевые слова:** Бинаризация, глобальные методы, локальные методы, метод Ниблека, метод Саувола, метод Бернсена.

*UDC 621.374*
***Balsaidov A. Sh.***
*Almaty Technology University*
*Almaty, Kazakhstan*
*balsaidov@gmail.com*

# PRE-PROCESSING SHOWN FOR BEST TEXT RECOGNITION

**Abstract.** In this article, I consider the process of image binarization. Various binarization methods are given, such as the Otsu method, lower threshold binarization, upper threshold binarization, double constraint binarization, Niblack's method, Sauvol's method, Bernsen's method. He also considered the issue of the advantages and disadvantages of each of them, as well as examples of using one or another binarization method on the original image. I also performed text recognition with various levels of noise. The Tesseract library was used to recognize text from the processed image.

**Keywords:** Binarization, global methods, local methods, Niblek's method, Sauvol's method, Bernsen's method.

**Введение.** Задача конвертации информации в текстовом виде с бумажных на электронные носители имеет большую потребность в наши дни. Наиболее легким способом конвертации информации является фотографирование либо сканирование бумажных документов. Несмотря на простоту этого метода, отсканированный или сфотографированный графический файл требует относительно больших трат на хранение и распространение информации. А так же, стоит отметить что на данный момент поисковые системы не индексируют текст на изображении. Что в свою очередь пагубно влияет на популяризацию и распространению казахского языка.

В таком случае, предпочтительным вариантом является конвертация бумажных носителей в электронный документ.

Наличие размытости, шумов, а также низкой контрастности изображения используемого для распознавания, значительно усложняет процесс алгоритма. Поэтому перед самим распознаванием изображение должно пройти процесс препроцессинга. Процесс нацелен на повышение качества распознаваемого изображения. Происходит удаление шумов, повышение контрастности, а также его бинаризация.

**Бинаризация изображения.** Процесс бинаризации - это конвертация изображения в черно-белое. Целью процесса является уменьшение объема информации на изображении для улучшения. Методы бинаризации можно разделить на две группы: глобальные и адаптивные. Главным параметром является порог - t, значение с которым сравнивается остальные пиксели изображения. После сравнения пикселю приравнивается значение 0 или 1. Процесс бинаризации представил на рисунке 1.

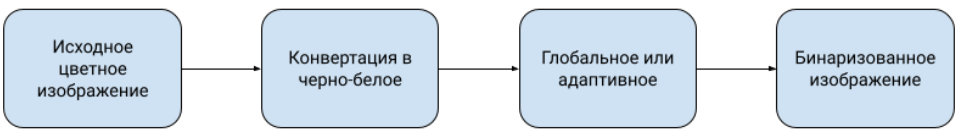



*Рисунок 1. Процесс бинаризации*
*Figure 1. process of binarization*

**Глобальные методы бинаризации.** В данном методе бинаризации работа происходит со всем графическим объектом сразу. К глобальным методам относятся:

– бинаризация с двойным ограничением;

– бинаризация верхнего порога;
– бинаризация нижнего порога;

-метод OTSU;

Бинаризация с нижним порогом является одним из самых простых методов преобразования графического объекта, где рассматривается лишь одно значение порога:

$$F'(m,n) = \begin{cases} 0, F(m,n) & x \geq t \\ 1, F(m,n) & x < 0 \end{cases}$$

В некоторых случаях используется метод бинаризации нижнего потока. В результате графический объект конвертируется в негативное изображение.

$$F'(m,n) = \begin{cases} 0, F(m,n) & x \leq t \\ 1, F(m,n) & x > 0 \end{cases}$$

Если необходима обработка выделенной части графического объекта, значение яркости пикселей которые могут изменяться в определенном диапазоне, то применяется метод бинаризации с двойным ограничением.

$$F'(m,n) = \begin{cases} 0, F(m,n) \geq t1 \\ 1, t1 < F(m,n) \leq t2 \\ 0, F(m,n) > t2 \end{cases}$$

Если есть необходимость получить наиболее простое решение для дальнейшего анализа изображения, то стоит применить алгоритм неполной пороговой обработки.

$$F'(m,n) = \begin{cases} F(n,m), F(m,n) > t \\ 0, F(m,n) \leq t \end{cases}$$

Бинаризованное изображение глобальным методом до и после представил на рисунке 2.

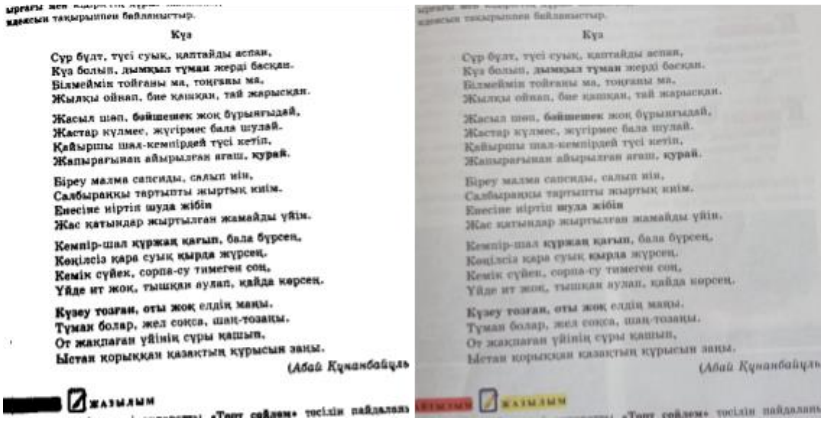


*Рисунок 2. Бинаризованное изображение глобальным методом до и после*
*Figure 2. Binarized image by global method before and after*

С другой стороны процесс бинаризации может привести к дефектам изображения, таким как, потеря информации, разрывы в линиях, искажение символов, нарушение целостности объектов а также, появление шумов. Это связано с тем что, входные изображения не всегда отличаются лучшим качеством. Освещение, ракурс, вспышка при фотографировании, а также другие факторы сильно влияют на работу алгоритмов бинаризации. Если рассмотреть фотографию(см. рисунок 3) сфотографированную на обычный смартфон со вспышкой, можно заметить что, края фотографии менее контрастны. При глобальной бинаризации данной фотографии(см. рисунок 4) края становятся искаженными, что в свою очередь делает невозможным распознавание данных частей фотографию.

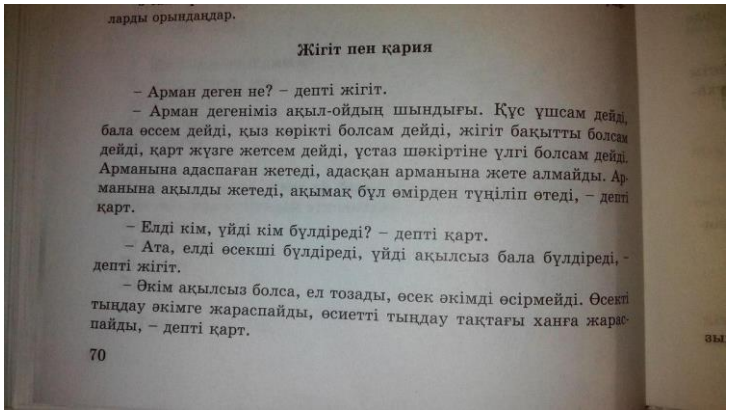
Отсканированным либо отфографированным изображениям присуще в целом одни и те же проблемы при распознавании текста. Для решения этой проблемы существуют различные методы локальной бинаризации изображения.

**Локальные методы бинаризации**. Локальные методы бинаризации производят разделение изображения на блоки где, производится вычисление необходимого порога, основываясь на информации об интенсивности пикселей. При разделении изображения, размер блока должен быть минимальным, но достаточным для сохранения исходных особенностей объекта. При адаптивной бинаризации можно получить удовлетворительный результат без использования фильтров. Рассмотрим локальные методы для бинаризации:

метод Ниблека;

метод Саувола; метод Бернсена;

Для тестирования каждого метода я использовал исходное изображение указанное на рисунке 3.


ларды орындандар.

**Жігіт пен қария**

– Арман деген не? – депті жігіт.

– Арман дегеніміз ақыл-ойдың шындығы. Құс ұшсам дейді, бала өссем дейді, қыз көрікті болсам дейді, жігіт бақытты болсам дейді, қарт жүзге жетсем дейді, ұстаз шәкіртіне үлгі болсам дейді. Арманына адаспаған жетеді, адасқан арманына жете алмайды. Арманына ақылды жетеді, ақымақ бұл өмірден түңіліп өтеді, – депті қарт.

– Елді кім, үйді кім бүлдіреді? – депті қарт.

– Ата, елді өсекші бүлдіреді, үйді ақылсыз бала бүлдіреді, – депті жігіт.

– Әкім ақылсыз болса, ел тозады, өсек әкімді өсірмейді. Өсекті тыңдау әкімге жараспайды, өсиетті тыңдау тақтағы ханға жараспайды, – депті қарт.

70

*Рисунок 3. Растровое исходное изображение*
*Figure 3. Source bitmap*

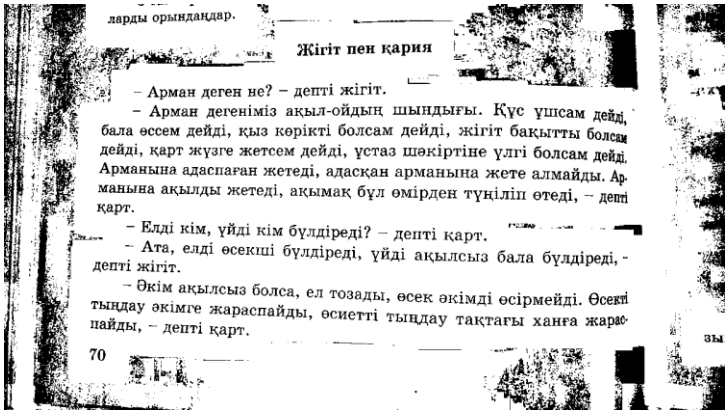
В методе Ниблэка идет обработка каждого пикселя изображения и их получение значение порога. Величина определяется вычислением локального среднего и локально среднеквадратического отклонения. Значение порога для точки с координатами(m,n) вычисляются по следующей формуле:

$$t(m,n) = \mu(m,n) + k \star \sigma(m,n)$$

Где, μ(m,n) представляет собой среднее, σ(m,n) -среднеквадратичное отклонение, а значение k определяет, какую именно часть границы объекта необходимо взять в качестве объекта.

Данный метод за счет своей простоты позволяет достичь высокую скорость бинаризации графического объекта.

На рисунке 4 показано применение метода Ниблэка.


ларды орындандар.

**Жігіт пен қария**

– Арман деген не? – депті жігіт.

– Арман дегеніміз ақыл-ойдың шындығы. Құс ұшсам дейді, бала өссем дейді, қыз көрікті болсам дейді, жігіт бақытты болсам дейді, қарт жүзге жетсем дейді, ұстаз шәкіртіне үлгі болсам дейді. Арманына адаспаған жетеді, адасқан арманына жете алмайды. Арманына ақылды жетеді, ақымақ бұл өмірден түңіліп өтеді, – депті қарт.

– Елді кім, үйді кім бүлдіреді? – депті қарт.

– Ата, елді өсекші бүлдіреді, үйді ақылсыз бала бүлдіреді, – депті жігіт.

– Әкім ақылсыз болса, ел тозады, өсек әкімді өсірмейді. Өсекті тыңдау әкімге жараспайды, өсиетті тыңдау тақтағы ханға жараспайды, – депті қарт.

70

*Рисунок 4. Результат бинаризации изображения с помощью метода Ниблэка.*
*Figure 4. The result of image binarization using the Niblack method.*

К методам локальной бинаризации относит и метод Саувола.

Определение порога бинаризации вычисляется с помощью прохождения изображения окном *w*w*. Порог *t(x,y)* определяется следующей формулой:

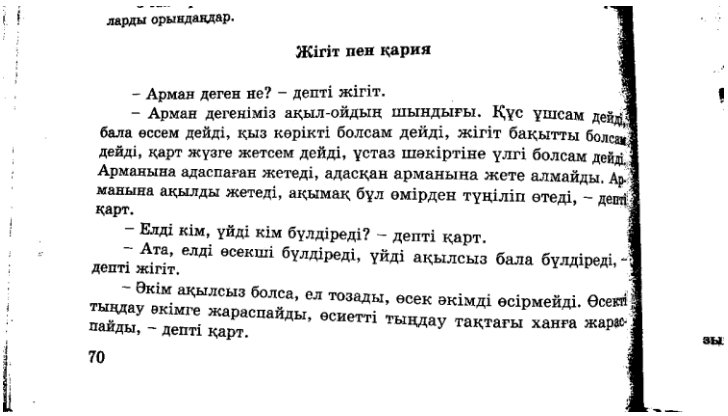
$$t(x,y) = m(x,y)\left[1 + k(\frac{s(x,y)}{R} - 1)\right]$$

Где, *m(x,y)* это среднее значение, *s(x,y)* среднеквадратическое отклонение интенсивности пикселя в окне *w*w* вокруг пикселя*(x,y)*.

В данной формуле R представляет максимальное отклонение, а k является параметром, который принимает значение в диапазоне[0.2, 0.5].

Метод Саувола применяется в основном для изобрежний с неравномерной яркости. Алгоритм менее устойчив к шуму, к низкому освещению изображения.

Результаты бинаризованного изображения методом Саувола указана в рисунке 5.


ларды орындаңдар.

Жігіт пен қария

– Арман деген не? – депті жігіт.

– Арман дегеніміз ақыл-ойдың шындығы. Құс ұшсам дейді, бала өссем дейді, қыз көрікті болсам дейді, жігіт бақытты болсам дейді, қарт жүзге жетсем дейді, ұстаз шәкіртіне үлгі болсам дейді. Арманына адаспаған жетеді, адасқан арманына жете алмайды. Арманына ақылды жетеді, ақымақ бұл өмірден түңіліп өтеді, – депті қарт.

– Елді кім, үйді кім бүлдіреді? – депті қарт.

– Ата, елді өсекші бүлдіреді, үйді ақылсыз бала бүлдіреді, – депті жігіт.

– Әкім ақылсыз болса, ел тозады, өсек әкімді өсірмейді. Өсекті тыңдау әкімге жараспайды, өсиетті тыңдау тақтағы ханға жараспайды, – депті қарт.

70

ары

*Рисунок 5. Результат бинаризации с помощью метода Саувола*
*Figure 5. The result of image binarization using the Sauvola method*

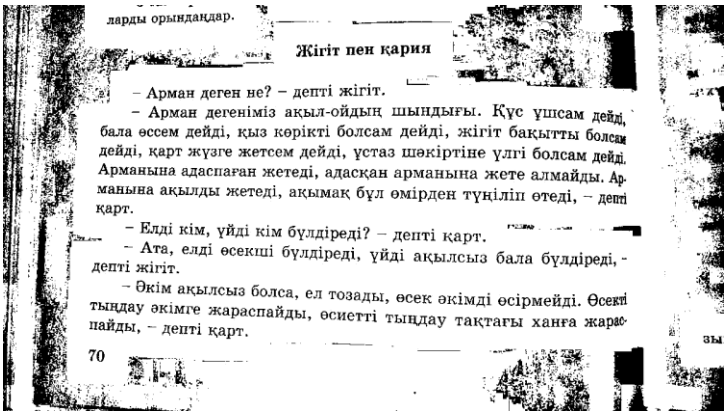
Следующим методом бинаризации является метод Бернсена. В данном методе изображение делится на квадраты, имеющие центр в точке (m,n). Каждый пиксель изображения имеет порог в пределах квадрата *r*r*, вычисляющийся по следующей формуле:

$$t(m,n) = \frac{jhigh + jlow}{2}$$

Где, *jhigh* наибольший уровень яркости квадрата, а *jlow* наименьший. Если обрабатываемый пиксель больше порога задаваемый пользователем, то он будет относится либо к черному, либо к белому.

Из минусов метода Бернсена можно отметить что, при обработке однородных областей формируются ложные черные пятна на бинаризованном изображении. Метод широко используется для схематических и картографических графических объектов.

Бинаризованное изображение методом Бернсена указан в рисунке 6.


ларды орындандар.

Жігіт пен қария

– Арман деген не? – депті жігіт.

– Арман дегеніміз ақыл-ойдың шындығы. Құс ұшсам дейді, бала өссем дейді, қыз көрікті болсам дейді, жігіт бақытты болсам дейді, қарт жүзге жетсем дейді, ұстаз шәкіртіне үлгі болсам дейді. Арманына адаспаған жетеді, адасқан арманына жете алмайды. Арманына ақылды жетеді, ақымақ бұл өмірден түңіліп өтеді, – депті қарт.

– Елді кім, үйді кім бүлдіреді? – депті қарт.

– Ата, елді өсекші бүлдіреді, үйді ақылсыз бала бүлдіреді, – депті жігіт.

– Әкім ақылсыз болса, ел тозады, өсек әкімді өсірмейді. Өсекті тыңдау әкімге жараспайды, өсиетті тыңдау тақтағы ханға жараспайды, – депті қарт.

70

*Рисунок 6. Результат бинаризации с помощью метода Бернсена.*
*Figure 6. The result of image binarization using the Bernson method*

**Заключение.** В данной статье описана система, которая распознает текст с изображения. Методом препроцессинга изображений является бинаризация. Были рассмотрены разные методы бинаризации, а также их формулы с применением. Система была реализована на языке программирования C# с использованием библиотек OpenCV. На данный момент программа активно

тестируется. В будущем планируется реализовать автоматический

выбор метода бинаризации для изображений с помощью машинного обучения.

**Список литературы**

1. Contrast adaptive binarization of low quality document images. - Electronic Express 2004.

2. Tesseract OCR/Github репозиторий открытого программного обеспечения. 2008-2022 URL:https://github.com/tesseract-ocr/tesseract(дата обращения:11.05.2022);

3. OpenCV / Github - репозиторий открытого программного обеспечения. 2006-2021 URL: https://github.com/opencv/opencv

4. Niblack W. An introduction to digital image processing // Prentice-Hall, Englewood Cliffs. – 1986. – P. 115–116.

5. Sauvola J. Adaptive document image binarization / J. Sauvola, M. Pietikäinen // Document Analysis and Recognition. – 1997. – Vol. 1. – P. 147–152.

6. Bernsen J. Dynamic thresholding of gray level images // Proceedings of International Conference on Pattern Recognition (ICPR). – Paris. – 1986. – P. 1251–1255.

*УДК 004.8, 004.94*

***[1]Сулейманов Д.Ш., [2]Гильмуллин Р.А., [3]Мухаметзянов И.Р.***
*Институт прикладной семиотики*
*Академии Наук Республики Татарстан*
*Казань, Татарстан, Россия*
*[1]dvdt.slt@gmail.com, [2]rinatgilmullin@gmail.com, [3]to.ilnur@gmail.com*

# О ПОТЕНЦИАЛЕ ГРАММАТИКИ ТАТАРСКОГО ЯЗЫКА ДЛЯ РАЗРАБОТКИ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

**Аннотация**. В статье рассматривается ряд лексико-грамматических признаков, определяющих технологичность татарского языка, как языка агглютинативного типа, представляющих определенный методологический и практический интерес для создания программных средств эффективной обработки естественно-языковой информации.

**Ключевые слова**: когнитивные показатели, технологичность языка, интеллектуальная система, морфологический эллипсис, рекурсия, нечеткие команды, активность знаний

*UDC 004.8, 004.94*

***[1]Suleymanov D.Sh., [2]Gilmullin R.A., [3]Muhametzyanov I.R.***
*Institute of Applied Semiotics of the*
*Academy of Sciences of Tatarstan Republic*
*Kazan, Tatarstan, Russia*
*[1]dvdt.slt@gmail.com, [2]rinatgilmullin@gmail.com, [3]to.ilnur@gmail.com*

# ON THE POTENTIAL OF THE GRAMMAR OF THE TATAR LANGUAGE FOR THE DEVELOPMENT OF THE INTELLIGENT SYSTEMS

**Abstract**. The article describes a number of lexico-grammatical features witch determine technological effectiveness of the Tatar language as an agglutinative type language. These features present a certain methodological and practical interest for creating software tools for effective processing of natural language information.

**Keywords**: cognitive indicators, technological effectiveness of the language, intellectual system, morphological ellipsis, recursion, fuzzy commands, knowledge activity

**Введение**

Большой практический интерес для построения интеллектуальных систем обработки информации представляют исследования естественных языков в трех аспектах: когнитивный, коммуникативный и технологический [Сулейманов, 2010]. Когнитивный аспект определяет потенциал естественного языка для описания модели мира, процессов мышления и представления знаний. Коммуникативный аспект отражает набор лексико-грамматических средств естественного языка для кодирования, приема и передачи, семиотической обработки информации, организации диалога. Технологический аспект определяет формальный и концептуальный потенциал естественного языка для реализации средств эффективной обработки, адекватного описания и компактного хранения информации, создания эргономичных технических средств, учитывающих специфику языка, а также для разработки интеллектуального программного инструментария, включая языки программирования, операционные системы.

В силу того, что в основе искусственных языков и систем программирования лежат глубинные структуры, соответственно, ментальность естественного языка, эти системы реализуют описательный и вычислительный потенциал соответствующего естественного языка.

Как известно, современные средства накопления и обработки знаний на естественном языке не интеллектуальны и слабо справляются с задачами поиска, извлечение знаний, семантический анализ текстовой информации. Это, прежде всего связано с тем, что они созданы на основе примитивных искусственных языков программирования, представляющих собой подмножество флективно-аналитических языков или формализмов, созданных на их основе. Еще одна причина сложностей в системах обработки ЕЯ связана с организацией их моделей, строящихся на основе формальных систем, в частности, порождающих грамматик (например, [Chomsky, 1957]), что создает две принципиальные проблемы: монотонность результатов логического вывода и пассивность инструментов логико-семантического анализа информации. Такая организация моделей ЕЯ названа в работе [Цейтин, 1980] *глобальным подходом* к организации исследований ЕЯ.

Современные системы искусственного интеллекта, основанные на технологиях нейронных сетей, машинного обучении и больших объемах знаний (bigdata), как известно, не способны порождать новые знания и объяснить, интерпретировать, представленное им решение задачи и быть «понятным» для человека.

В связи с этим перспективным представляется исследование технологического аспекта естественных языков с целью выявления лексико-грамматических (морфологических, синтаксических, семантических) структур, достаточно регулярных и обладающих естественной сложностью, обеспечивающих поверхностное кодирование технологических характеристик, с целью создания на их базе новых языков программирования с развитыми возможностями интеллектуальной обработки информации. Такие исследования особенно актуальны для агглютинативных языков, к которым относятся тюркские языки, которые характеризуются достаточно сложной и, одновременно, практически регулярной морфологией, позволяющей в одной словоформе закодировать практически целую субъектно-предикативную ситуацию, описываемую в флективно-аналитических языках, таких как английский язык, несколькими предложениями.

Для систем обработки знаний определяющими их интеллектуальность являются следующие характеристики: 1) возможность кодирования и обработки нечеткой информации; 2) активность знаний.

В статье описываются исследования технологического аспекта естественных языков на примере татарского языка и раскрывается ряд показателей, определяющих эффективность лексико-грамматической модели татарского языка с точки зрения создания интеллектуальных систем обработки информации.

## 1. Когнитивные и лексико-грамматические показатели технологичности татарского языка

### 1.1. Регулярность и естественная сложность морфологии

Как показывают исследования, татарский язык, являясь одним из тюркских языков, имеет богатую, сложную, и одновременно, достаточно регулярную морфологию [Сулейманов и др., 2003], обладает потенциалом, позволяющим эффективно кодировать и компактно хранить информацию, а также реализовывать на уровне аффиксальных морфем такие явления, как рекурсия, «нечеткость». Как известно, автоматные грамматики обладают минимальными характеристиками временной и емкостной функций, то есть на порождение и обработку информации требуется меньше времени и промежуточной памяти по сравнению с контекстно-свободными и контекстно-зависимыми грамматиками. Таким образом, регулярность, почти автоматность морфологии татарского языка обеспечивает минимизацию емкостных и временных функций при обработке текстов на татарском языке, а также достаточно простой анализ структуры и значения словоформы, несмотря на естественную сложность морфологии.

Рассмотрим далее более детально ряд признаков, характеризующих регулярность и естественную сложность морфологии татарского языка на примере именных и глагольных форм.

Регулярность морфологии означает, что одна и та же схема сочетания морфем (морфотактика) присуща всем или почти всем именным и глагольным формам, соответственно. Такая возможность позволяет по одной и той же схеме практически автоматно образовывать словоформы с одними и теми же глубинными значениями аффиксов. Важным свойством татарской морфологии, наряду с ее регулярностью, является фиксированность позиций аффиксов в последовательности аффиксальных морфем и связанность позиций аффиксов с их типами. Например, в последовательностях словоформ: 1) елга, елгалар, елгаларым, елгаларыма – (‘река, реки, мои реки, моим рекам’); 2) кыр, кырлар, кырларым, кырларыма – (‘поле, поля, мои поля, моим полям’) именные корневые морфемы елга (‘река’) и кыр (‘поле’) имеют одни и те же последовательности аффиксальных морфем с идентичными значениями. Позиции аффиксальных морфем, составляющих словоформу, связаны с определенными типами морфем и неизменны относительно друг друга. Аффиксальные морфемы определенного типа могут появляться только в соответствующей позиции, либо выпадать вместе с позицией.

Свойства регулярности морфотактики и фиксированности позиций соответствующих типов аффиксальных морфем в татарском языке присущи также и глагольным формам.

Последовательность аффиксов, служащая для описания соответствующих значений ролевой ситуации, кодируемой глагольной группой, также как и в случае именной группы, определяется для глагольной словоформы, следующей самой правой в последовательности словоформ, входящих в глагольную группу.

Например, в глагольной группе:

*Йөгереп барып карап алып кайттыгызмы?* (букв.: *Бегом+сходив+посмотрев+взяв возвратились ли?*) последовательность аффиксов *–ты+гыз+мы* присоединяется к последней глагольной морфеме кайт (пов. накл., 2 л., ед.ч.) (*‘возвратись’*), очевидно, являясь некоторой заскобочной цепочкой, завершающей глагольную группу и относящейся ко всей глагольной группе.

Рассмотрим ряд признаков, определяющих естественную сложность татарской морфологии: 1) возможность присоединения к словоформе определенных аффиксальных морфем, изменяющих тип слова, превращающих, например, именную словоформу в глагольную или в форму прилагательного и наоборот; 2) морфологическое

(синтетическое) задание признаков модальности, настроения, эмоционально-личностного отношения к ситуации, объекту или процессу, описываемым данной словоформой; 3) контекстное разнообразие значений аффикса. Известно, что именная группа, как правило, кодирует некую семантическую ролевую ситуацию, а глагольная группа – контекстные отношения над этими ролями. Таким образом, возможность перехода с именной формы к глагольной и, наоборот, с глагольной формы к именной, через присоединение соответствующих аффиксов, позволяет описывать одновременно в пределах одной словоформы как сложную ролевую ситуацию, так и контекстные отношения между семантическими ролями.

Тем самым обеспечивается компактность описания и хранения информации. Синтетический, аффиксальный способ словоизменения обеспечивает кодирование в рамках одной словоформы некоторого значения, описываемого на флективно-аналитических языках (например, на английском) несколькими словосочетаниями и даже предложениями.

В качестве примера реализации признака (1) рассмотрим следующую словоформу, являющуюся корректной для татарского языка: *Татарчалаштыргалаштыручылардагыныкыларгамыни?* (Разве тем (к тем/на тех), что принадлежит тому (той), что на тех, кто (что) время от времени занимаются татаризацией (переводом на татарский язык)?). Данная словоформа имеет следующую структуру: Татар (Имя сущ.) + ча (Наречие) + ла (Глагол) + штыр (Глагол, залог) + гала (Глагол, залог)+штыр (Глагол, залог)+у (Субстантив., имя действ.)+чы (Имя сущ.) + лар (Множ.) + дагы (Субстантив., локатив) + ныкы (Субстантив., притяжат.) + лар (Множ.) + га (Директив) + мыни (Вопрос, удивление).

Возможность аффиксального задания признака модальности (2), в отличие от других языков, в которых данный признак отображается либо эмоционально-просодически, либо с помощью дополнительной словоформы, также является свойством, способствующим адекватной интерпретации значения словоформы и минимизирующим время его распознавания.

Третий признак сложности татарской морфологии определяет контекстное разнообразие значений аффикса. Практически все аффиксальные морфемы обладают свойством полисемии. В частности, как показывают наши исследования [3], аффиксальная морфема –ГА имеет порядка 20 значений, то есть используется для кодирования до 20 различных контекстных значений.

## 1.2. Морфологический эллипсис и явление рекурсии в татарском языке

Явления морфологического эллипсиса и рекурсии в татарском языке также могут быть отнесены к показателям, повышающим технологичность татарского языка.

Морфологический эллипсис определяется как возможность пропуска последовательности аффиксов при однородных именных словоформах с сохранением ее в последней словоформе. То есть, возможность вывода последовательности аффиксов любой длины, общей для однородных членов, вправо, за последовательность однородных членов, и присоединение их к последнему справа однородному члену.

Например: Ишек алды тавыкларга, казларга, сыерларга тулы = Ишек алды тавык, каз, сыерларга тулы. ‘Двор полон кур, гусей, коров’.

Мин кырларыбызга, урманнарыбызга, елгаларыбызга, тауларыбызга шатланам = мин кыр, урман, елга, тауларыбызга шатланам . ‘Я радуюсь нашим полям, лесам, рекам’.

Явление рекурсии определяется как возможность циклического порождения нового значения путем последовательного применения одной и той же «формулы», т.е. повторного присоединения одной и той же аффиксальной морфемы.

Такими свойствами обладают аффиксальные морфемы –ДАгЫ (локатив2, место-временной падеж 2) и –нЫкЫ (притяжат.падеж), которые можно назвать также аффиксами неопределенности, т.е. аффиксами, придающими неопределенность к присоединенным лексемам.

Например, пусть задана лексема тау (‘гора’). Присоединение аффикса –дагы порождает новые объекты или свойства, являющиеся неопределенными: таудагы – ‘нечто на горе’; таудагыдагы – ‘нечто на нечто на горе’; тауныкы – ‘то, что принадлежит горе’; тауныкыныкы – ‘то, что принадлежит тому, что принадлежит горе’.

По такой формуле может быть образована словоформа практически неограниченной длины. Естественно, такие длинные последовательности морфем в речи практически не используются. Это, прежде всего, как мы считаем, связано с проблемами глубины памяти, удобства общения между людьми. Тем не менее, подобное словоизменение является совершенно корректной с точки зрения грамматики татарского языка и словоформа, образованная присоединением последовательности любой длины, гипотетически всегда имеет смысл, конкретное значение приобретается при «погружении» словоформы в определенный контекст.

Приведем пример со следующей словоформой: тауныкын-дагыныкыныкындагы, которая однозначно раскладывается на следующие составляющие - тау+дагы+ныкы+ндагы+ныкы+ндагы -‘тау’ (имя сущ.+локатив2+притяж.пад.+локатив2+притяж.пад. +локатив2).

Данная словоформа означает следующее:

‘нечто, находящееся на/в нечто, принадлежащее нечто, находящееся на/в нечто, принадлежащее нечто, находящееся на/в горе’.

Нетрудно заметить, что, задавая параметры после каждой морфемы эксплицитно, в явном виде, можно добиться определенности значения словоформы. То есть словоформа, после подстановки конкретных значений вместо аффиксов неопределенности, также приобретает конкретное значение. В реальных случаях, в речи, такие параметры наполняются конкретным значением от контекста речи, дискурса.

Рассмотрим следующий пример для иллюстрации изложенного утверждения. Пусть после каждого аффикса неопределенности стоят параметры: тау+дагы(x1)+ндагы(x2)+ныкы(x3)+ныкы(x4)+ ндагы(x5) + ныкы(x6), где xi – контекстные объекты, т.е. объекты, приобретающие конкретное значение либо из контекста, либо их задает пользователь (i=1,6). Таким образом, придавая значения параметрам: x1= «куыш», x2= «аю», x3= «лапа», x4= «коготь», x5= «мед», мы получаем следующее контекстное значение: «нечто (значение x6, придаваемое параметру последним аффиксом, осталось неопределенным), что присуще меду, что на когте, что принадлежит лапе, что принадлежит медведю, что находится в пещере, что находится на горе».

На месте корневой морфемы также может стоять неопределенный параметр:

X+дагы(x1)+ндагы(x2)+ныкы(x3)+ныкы(x4)+ндагы(x5)+ныкы(x6).При этом на месте X может быть любое понятие, задаваемое имплицитно, и раскрываемое через контекст, либо задаваемое эксплицитно (т.е. явно) пользователем. Например, для нашего случая: X=тау (‘гора’).

Рассмотрим проявление свойства рекурсии на примере целых предложений.

Кыр куяны колакларындагы кара тапларда матурлык бар. Урман куяныныкылардагыларныкыннан башкарак. (‘Есть красота в черных пятнах на ушах полевых зайцев. Несколько иная, чем та красота, которая в черных пятнах на ушах лесного зайца’).

Здесь в словоформе куянынык ылардагыларныкыннан = куяны(‘заяц’)+ныкы(x0)+лар(множ.)+дагы(x1)+лар(множ.)+ныкы(x2)+ннан (исх.падеж) ряд понятий (x0, x1, x2) задан имплицитно, однако, однозначно раскрывается по предыдущему контексту (т.е. по пресуппозиции): x0=колак (‘ухо’); x1= кара тап(‘черное пятно’); x2 = матурлык(‘красота’).

Второе предложение при полном эксплицитном написании выглядит следующим образом: Урман куяны колакларындагы кара таплардагы матурлыктан башкарак (‘Несколько иная, чем та красота, которая в черных пятнах на ушах лесного зайца’).

Даже на этом коротком примере элементарный расчет показывает, что применение рекурсивных аффиксов приводит к сжатию информации и существенной экономии памяти. В случае применения рекурсии в приведенном примере количество слов сокращается в два с лишним раза и число используемых символов уменьшается на 23 (В варианте без рекурсии: 7 слов, 64 знака; в варианте с рекурсией: 3 слова, 41 знак). При этом по контексту осуществляется достаточно простая и однозначная экспликация неопределенностей, которые известны в лингвистике как явление анафоры. В нашем случае этот тип анафоры можно назвать анафорой рекурсии.

### 1.3. Нечеткость описания команд и действий и описание одной глагольной словоформой ролевой ситуации

Известно, что поверхностное, лексическое описание предикатов (команд, действий, отношений), как правило, осуществляется глагольными словоформами.

Далее рассмотрим следующие два признака технологичности, отражающие естественные когнитивные механизмы, проявляющиеся в глагольных словоформах:

1) Возможность рекурсивно задавать нечеткие команды и описывать нечеткие действия и связи между объектами.

2) Возможность рекурсивно описывать в рамках одной словоформы действия, относящиеся к целой ролевой ситуации.

Признак (1) кодируется глагольными аффиксами, занимающими позицию залога, т.е. сразу же после глагольной основы, – ГАлА, -штЫр.

Например:

у (‘стирай’) – ‘стирать’ (3 лицо, ед.ч., повел. накл.);

угала (‘стирай время от времени’) - у(‘мой’)+гала (‘время от времени’);

угалаштыр (‘стирай время от времени, время от времени: реже’) - у(‘мой’)+гала(‘время от времени’)+штыр (‘время от времени’);

угалаштыргала (‘стирай время от времени, время от времени, время от времени: еще реже’) - у(‘стирай’) +гала (‘время от времени’) + штыр (‘время от времени’) + гала (‘время от времени’); угалаштыргалаштыргала… (‘стирай время от времени, время от времени, время от времени: и еще реже…’)

у (стирать, корень, 3 лицо, ед.ч., повел.накл.)+гала(‘время от времени: изредка’)+штыр(‘время от времени: еще реже’)+гала(еще реже)+штыр(еще реже)+гала(еще реже)…

Сам факт, насколько редко требуется стирать, определяется исходя из контекстной информации или из модели мира. Например, футболканы уыштыргала – ‘стирай футболку время от времени’, может означать команду: стирать футболку один раз после нескольких дней одевания, в то время как, команда галстукны уыштыргала - ‘стирай галстук время от времени’, скорее всего, означает: стирать галстук один раз в несколько месяцев или даже еще реже.

Реализация признака 2 обеспечивается рядом специальных глагольных аффиксов, занимающих также залоговую позицию: -н, -Ыш, -т, -Дыр.

Рассмотрим изменения ролевой ситуации при присоединении соответствующих аффиксов на примере с глагольной словоформой ташла (‘бросай’).

Участники действия: субъект S, объект-предмет $O_k$, где $k \geq 1$.

Для словоформы ташла (‘бросай’) ролевая ситуация следующая: S воздействие на $O_k$

Присоединение аффиксов -н, -Ыш, -т, -ДЫр приводит к изменениям, описанным ниже.

1) –н:

ташлан – ташла+н (‘бросайся’)

Ролевая ситуация: S воздействие S (рефлексия)

2) -Ыш:

ташлаш – ташла+ш (‘помогай бросать/бросай вместе’)

Участники действия: субъект S, объект-актор $A_{i,j}$, объект-предмет $O_k$, где i – номер группы объекта-актора, $i \geq 1$; j – число участников в группе i, $j \geq 1$.

Ролевая ситуация:

S воздействие (помощь) $A_{i,j}$ и (S & $A_{i,j}$) воздействие (бросить) $O_k$.

3) -т, -ДЫр:

ташлат – ташла+т (‘сделай так, чтобы бросил/бросили’)

Ролевая ситуация:

S воздействие $A_{i,j}$ -> $A_{i,j}$ воздействие (бросить) $O_k$. Здесь стрелка -> означает импликацию.

ташлаттыр – ташла+т+тыр (‘сделай так, чтобы сделали так, чтобы бросили’)

Ролевая ситуация:

S воздействие $A_{i,j}$ -> $A_{i,j}$ воздействие $A_{l,m}$ -> $A_{l,m}$ воздействие (бросить) $O_k$. ташлаттырт - ташла+т+тыр+т (‘сделай так, чтобы сделали

так, чтобы сделали так, чтобы бросили')Ролевая ситуация: S воздействие $A_{i,j}$ -> $A_{i,j}$ воздействие $A_{l,m}$ -> $A_{l,m}$ воздействия $A_{s,t}$ -> $A_{s,t}$ воздействие (бросить) $O_{k.}$

По такой формуле, подставляя новые определенные аффиксы, можно создавать все новые и новые ролевые ситуации и описывать процессы на лексическом уровне. Например, добавление аффикса –Ыл к последней полученной словоформе: ташлаттыртыл превращает сам субъект в объект-предмет, объект воздействия, т.е. S= $O_{k.}$

Получается следующая ролевая ситуация:

S воздействие $A_{i,j}$ -> $A_{i,j}$ воздействие $A_{l,m}$ -> $A_{l,m}$ воздействие $A_{s,t}$ -> $A_{s,t}$ воздействие (бросить) $S_{.}$

### 1.4. Активность знаний

Английские предложения строятся по схеме S-V-O (subject-verb - object: субъект-глагол-объект), а татарские - по схеме: S-O-V. То есть, англичанин, если говорит, например, о намерении сходить в кино, сначала скажет, пойдет или не пойдет, и только после этого выдает информацию – куда, какой, зачем, с кем, когда и т.д. ('I'll go to the cinema with my friend afternoon'). Как видно из примера, здесь действие управляет ситуацией. После того, как высказано однозначно намерение субъекта, дальнейшая информация становится пассивной, практически не влияет на выбор способа действия или усложняет его. А на татарском языке сначала дается информация и её анализ, и только после этого, возможно, с учетом реакции слушающего, определяется – положительное или отрицательное, само действие. ('Мин дустым белән төштән соң кинога барам/бармыйм'» – буквально: 'Я со своим другом после обеда на фильм пойду/не пойду'). В системах искусственного интеллекта это называется активностью знаний, что является одним из важных признаков интеллектуальности системы [Поспелов и др., 1999]. Для интеллектуальных систем естественным и основополагающим является стиль размышления: анализ-действие, размышление-цели-алгоритмы, а не командный стиль: действие-анализ, алгоритм-цель, как это реализовано в современных технологиях, основанных на менталитете английского языка. То есть сначала анализируется, обрабатывается информация, а затем осуществляется некое действие - подбирается соответствующая адекватная модель представления знаний или выбираются соответствующие алгоритмы и схемы реализации, оптимальность и эффективность которых во многом определяется корректностью и полнотой анализа информации. Это можно назвать событийным программированием.

Такие возможности, являющиеся естественными для татарского языка и закрепленные в грамматике татарского языка, позволяют ставить задачу о разработке интеллектуальных программ накопления и извлечения знаний в глобальных компьютерных сетях, что, как известно, становится сверхактуальной задачей в современном информационном мире.

**3. Заключение**

В статье описан ряд потенциальных когнитивных возможностей татарского языка, которые показывают перспективу для татарского языка стать формальной базой для построения новых интеллектуальных технологий описания, хранения и обработки информации.

В дальнейшем нами планируется исследование и построение математических моделей, отражающих лексико-грамматический потенциал татарского языка, как основы интеллектуальных технологий, включая такие свойства морфологии, как рекурсия, морфологический эллипсис, функциональное многообразие и семантическая многовалентность аффиксов (в том числе, аффиксов кодирования неопределенной информации и нечетких команд). Весьма перспективным представляется также исследование синтаксической структуры, обеспечивающей реализацию свойства активности знаний, являющегося важным показателем интеллектуальности прикладной системы.

**Список литературы**

1. Сулейманов Д.Ш. К вопросу исследования технологического аспекта естественных языков // Обработка текста и когнитивные технологии: Труды XI Междунар. науч. конф. (Констанца, 7–14 сентября 2009 г.). Казань: Изд-во Казан. гос. ун-та, 2010, с. 232-245.

2. Chomsky N. Syntactic Structures. The Hague: Mouton, 1957.

3. Цейтин Г.С. О соотношении естественного языка и формальной модели. Архив АН СССР. Работа в Научном совете по комплексной проблеме "Кибернетика", 1980 г.

4. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А. Концептуальные и лингвистические аспекты разработки корпуса татарского языка // Фэнни Татарстан. 2017. № 2. С. 7-16.

5. Сулейманов Д.Ш., Невзорова О.А., Галиева А.М., Гатиатуллин А.Р., Гильмуллин Р.А., Хакимов Б.Э. Размеченный корпус татарского языка "Туган тел": аспекты реализации. В сборнике: Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2014. Научные редакторы: Д.Ш. Сулейманов, О.А. Невзорова. 2014. С. 88-93.

6. Поспелов Д.А., Осипов Г.С. Прикладная семиотика. (Из неизданных книг). Доступно: http://raii.org/library/ainews/1999/1/OSPOS.ZIP

7. Suleymanov D.Sh. Natural Cognitive Mechanisms in the Tatar language [Text] / D.Sh. Suleymanov // In the Collection of the Vienna Proceedings of the Twentieth European Meeting in Cybernetics and Systems Research. Ed. by Robert Trappel. Vienna, Austria, 6-9 April, 2010. – P.210-213.

*ӘОК 81.35*
***Сыздықова Г. О.***
*Л.Н.Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*go.syzdykova@mail.ru*

## А. БАЙТҰРСЫНҰЛЫ ТЫНЫС БЕЛГІЛЕРІНІҢ ТҮРЛЕРІ МЕН ҚОЛДАНЫСЫ ТУРАЛЫ

**Аңдатпа.** Қазақ тіліндегі тыныс белгілерінің қалыптасуы мен дамуы ғалым А.Байтұрсынұлының зерттеулерінен бастау алады. Ғалым тыныс белгілерінің түрлерін жіктеп, әрқайсысына қолданыс сипатына қарай атау береді. Тыныс белгілерінің жұмсалымы мен қызметін белгілейтін ережелер жүйесін ұсынады.

Мақалада А.Байтұрсынұлының қазақ тіліндегі тыныс белгілерінің түрлері мен қызметі, сөйлемдегі қолданысы туралы ғылыми көзқарастары сараланады. Ғалымның тыныс белгілеріне берген атаулары қазіргі қолданыста қалыптасқан тыныс белгі атауларымен салыстырыла талданады.

Жасалған талдаулар негізінде А.Байтұрсынұлының ұсынған тыныс белгілерінің басым бөлігінің қазіргі тіл жүйесінде қолданылатындығы анықталды. Бұл А.Байтұрсынұлының қазақ пунктуациясының негізін салушы ғалым ретіндегі орнын тағы да айқындай түседі.

**Түйін сөздер**: пунктуация, тыныс белгілері, жазу, емле, тіл жүйесі.

*УДК 81.35*
***Сыздыкова Г. О.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*go.syzdykova@mail.ru*

## А. БАЙТУРСЫНУЛЫ О ВИДАХ И УПОТРЕБЛЕНИИ ЗНАКОВ ПРЕПИНАНИЯ

**Аннотация.** Формирование и развитие знаков препинания в казахском языке берет свое начало в исследованиях ученого А. Байтурсынова. Ученый классифицирует типы знаков препинания и дает каждому знаку название в зависимости от характера применения. Представляет систему правил, устанавливающих действие и функцию знаков препинания.

В статье анализируются научные взгляды А. Байтурсынова о видах и функциях, применении в предложении знаков препинания казахского языка. Названия знаков препинания ученого сопоставляются с названиями знаков препинания, сложившимися в современном употреблении.

На основании проведенного анализа установлено, что большинство знаков препинания, предложенных А. Байтурсыновым используются в системе современного языка. Это еще раз подчеркивает место А. Байтурсынова как ученого-основателя казахской пунктуации.

**Ключевые слова:** пунктуация, знаки препинания, графика, орфография, языковая система.

*UDC 81.35*

***Syzdykova G.***
*L. N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*go.syzdykova@mail.ru*

## A. BAITURSYNULY ON THE TYPES AND USE OF PUNCTUATION MARKS

**Abstract:** The formation and development of punctuation marks in the Kazakh language originates in the research of the scientist A. Baitursynov. The scientist classifies the types of punctuation marks and gives each sign a name depending on the nature of the application. Represents a system of rules that establish the action and function of punctuation marks.

The article analyzes the scientific views of A. Baitursynov on the types and functions, the use of punctuation marks in the sentence of the Kazakh language. The names of punctuation marks of the scientist are compared with the names of punctuation marks that have developed in modern use.

Based on the analysis, it was found that most of the punctuation marks proposed by A. Baitursynov are used in the modern language system. This once again emphasizes the place of A. Baitursynov as the scientist-founder of Kazakh punctuation.

**Keywords:** punctuation, punctuation marks, graphics, spelling, language system.

Пунктуация – тіл білімінің тыныс белгілер жүйесін зерттейтін саласы. Ал тыныс белгісі сол тілдің жазу және емле жүйесімен тығыз байланысты. Бұл байланыс тыныс белгілерінің жазудың пайда болуымен бірге қалыптасып, жазудың даму барысында жүйеленуімен

анықталады. Пунктуацияның жазу және емлемен байланысы оған берілген анықтамаларда да нақты тұжырымдалады. Мәселен, Ғ.Қалиевтің «Тіл білімі терминдерінің түсіндірме сөздігінде» пунктуация былай түсіндіріледі: «Тыныс белгілері (пунктуация), (лат. *punctuotio, punctum* нүкте) – 1. Графика, орфографиямен бірге жазба тілдің негізгі құралы болып табылатын, әліпбиден тыс графикалық белгілер жүйесі. Тыныс белгілерінің мақсаты – жазба мәтіннің айтудағы бөлшектенуіне сай оны графикалық тұрғыда дұрыс беру. 2. Тіл білімінің тыныс белгілері және оларды дұрыс қолдану ережелері туралы тарауы. Тыныс белгілері жазуда қолданылатын тыныс белгілерінің құрамын, атқаратын қызметі мен білдіретін мағыналарын және оларды қолдану ережелерін зерттейді» [Қалиев, 2005, 340-б]. Бұл анықтамада пунктуация, біріншіден, әліпбиден тыс графикалық белгілер жүйесі, екіншіден, тіл білімінің дербес бір тарауы ретінде сипатталған.

Лингвистикалық түсіндірме сөздікте пунктуация (лат. *punctuohatio, punctum* – нүкте) – «1) графика, орфография және алфавиттен тыс болатын белгілер жүйесі. Бұлар жазба тілдің негізгі құралдары болып табылады. Пунктуацияның негізгі мақсаты – жазба текстi бөлшектеу және графикалық құрылымды жасау; 2) тарихи қалыптасқан жазба текстің кодификациялық нормалары мен ережелері; 3) тыныс белгілердің қолдану нормалары мен пунктуация жүйесіндегі заңдылықтарды зерттейтін тіл білімінің саласы» [Салқынбай, Абақан, 1998, 170-171-б.] ретінде анықталады. Алдыңғы сөздіктегі анықтамадан ерекшелігі: мұнда пунктуацияны «тарихи қалыптасқан жазба текстің кодификациялық нормалары мен ережелері» ретінде де анықтайды.

Ғалым Ф.Мұсабекова «Қазіргі қазақ тілінің пунктуациясы» оқу құралында тыныс белгісі таңбасының алғашқы элементтері түркі тілдерінің бұрынғы ескерткіштерінде кездескенін және ол кездегі бар белгілер, көбінесе, тыныс белгісі қызметінен гөрі бір сөзді екінші сөзден айыру қызметін атқарғанын айтады. Бертін келе, қазақ тіліндегі алғашқы басылым «Дала уәлаяты» газетінде (1888-1902) 1894 жылдан бастап кейбір сөйлемдердің жігін ажыратуға әредік сызықша (–), әр түрлі жұлдызшалар (*) сияқты шартты таңбаларды қолданады [Мұсабекова, 1991, 6- б.].

Қазақ тілінің тыныс белгілері жөнінде жазылған еңбектерді хронологиялық жағынан жүйелеуде Ф.Мұсабекова осы бағыттағы алғашқы еңбектердің бірі ретінде Ш.Сарыбаевтің 1936 жылы жазылған «Қазақ сөйлемінде үтірдің жазылатын орындары» деп аталатын мақаласын көрсетеді. Бірақ қазақ тілінің дыбыс, сөз, сөйлем, сөйлеу жүйесімен бірге тыныс белгілері жүйесінің зерттелу тарихы әріден, қазақ тіл білімінің негізін салушы, ұлт ұстазы Ахмет Байтұрсынұлының

еңбектерінен бастау алатындығы қазіргі кезде анықталған. Ғалым 1925 жылы түзетіліп, толықтырылып, жаңа емлемен қайта басылған «Тіл-құрал» деп аталатын 3-тілтанытқыш кітабында қазақ тіліндегі тыныс белгілерін жүйелеп, атау беріп, әр атауды ұлттық танымммен байланыста саралаған. Одан кейін 1928 жылы шыққан «Тіл жұмсар» еңбегінде ғалым мынандай пікір айтады: «Әріптен басқа жазуда қолданылатын белгілерге *бүгінге дейін анықтап белгілі ат қойылған жоқ* еді. Сүгіретке қарап біреу олай, біреу былай деп атайтын еді. Мұнда соларға ат қойылды. Мұндағы *ат тек сүгіретіне қарай емес, жұмсалатын орнына қарай қойылады*. Мәселен: **тыныстық (.), жапсарлық (,), қосарлық (=), тастарлық (–), дәлдеулік (« »), сұраулық (?), лептеулік (!)»** [Байтұрсынов, 1992, 337-б.]. Ғалым тыныс белгілеріне «*бүгінге дейін нақты ат қойылмағанына»* баса назар аударып, олардың әрқайсысына атау береді. Атау беруде әр белгінің сыртқы формасын ғана емес, қолданатын орнын, қызметін де негізге алады.

Бұл мақалада А.Байтұрсынұлының ұсынған тыныс белгілері түрі, атауы, жұмсалымы және қызметі тұрғысынан сараланып, қазақ тілінің қазіргі қолданыстағы тыныс белгілерімен салыстырыла талданады.

Қазақ тілінде қазіргі қолданыстағы тыныс белгілер саны – 10 [Сыздық, 2000, 107-116 б.]. А.Байтұрсынұлы «Тіл-құрал» еңбегінде 11 тыныс белгісін ұсынады [Байтұрсынов, 1992, 310-312 б.] (1-кесте).

*1-кесте*

*А.Байтұрсынұлының «Тіл-құрал» еңбегі мен Р.Сыздықтың «Қазақ тілінің анықтағышындағы» тыныс белгілер жүйесі*

| № | *А.Байтұрсынұлы. Тіл тағылымы. 1991* | | *Р.Сыздық. Қазақ тілінің анықтағышы. 2000* | |
|---|---|---|---|---|
| | *Тыныс белгі атауы* | *Таңбасы* | *Тыныс белгі атауы* | *Таңбасы* |
| 1 | Кіші сызықша (тіркестіру белгісі) | = | - | - |
| 2 | Үлкен сызықша (жұмақтау белгісі) | – | Сызықша | – |
| 3 | Ноқат (нүкте), (ұлы тыныс) | . | Нүкте | . |
| 4 | Үтірлі ноқат | ; | Нүктелі үтір | ; |

| | | | | |
|---|---|---|---|---|
| | (нүкте), (орта тыныс) | | | |
| 5 | Теріс үтір (кіші тыныс) | **,** | Үтір | **,** |
| 6 | Қос ноқат (қос нүкте), (бәшелеу белгісі) | **:** | Қос нүкте | **:** |
| 7 | Сұрау белгісі | **?** | Сұрау белгісі | **?** |
| 8 | Леп белгісі | **!** | Леп белгісі | **!** |
| 9 | Қабат үтір (қабатша қос тырнақ) | **« »** | Тырнақша | **« »** |
| 10 | Жақша (қамау белгісі) | **( )** | Жақша | **( )** |
| 11 | Көп ноқат (көп нүкте) | **...** | Көп нүкте | **...** |

Кестедегі салыстырудан қазіргі тіл жүйесінде А.Байтұрсынұлының 20-ғасырдың 20-жылдары ұсынған тыныс белгі атауларының басым бөлігінің сол қалпында, өзгеріссіз қолданыста жүргенін байқаймыз. Ғалымның ұсынған *нүкте, қос нүкте, сұрау белгісі, леп белгісі, жақша, көп нүкте* тәрізді тыныс белгі атаулары еш өзгеріссіз қабылданса, «*үлкен сызықша» сызықша, «теріс үтір» үтір,* «*үтірлі ноқат» (нүкте) нүктелі үтір* түрінде тек құрамындағы анықтаушы сыңарын (*үлкен*, *теріс*) түсірумен немесе анықтаушы-анықталушы сыңарларының (*үтірлі нүкте – нүктелі үтір*) орнын ауыстырумен өзгешеленсе, *«қабат үтір» тырнақша* болып басқа сөзбен аталған. Ал тыныс белгі таңбалары еш өзгеріссіз қолданылған. А.Байтұрсынұлының ұсынған тыныс белгілерінің ішінде *кіші сызықша (тіркестіру белгісі)* Р.Сыздықтың «Анықтағышындағы» тыныс белгілер қатарында көрсетілмейді. А.Байтұрсынұлы *кіші сызықшаны «тіркестіру белгісі»* деген қосымша атаумен атап, оның қос сөздерде және сөздің бір жолға сыймай қалған бөлігін келесі жолға тасымалдауда қолданылатын белгі екендігін түсіндіреді. Ғалымның бұл тұжырымы Р.Сыздықтың «... белгісіздік-жалпылық және болжалдық мағынасындағы қос сөз, сан есімдер ... дефис арқылы жазылады» деген пікірінде дамытылады [Сыздық, 2000, 113-б.]. Сондықтан кіші сызықша орфографиялық таңба ретінде танылады.

А.Байтұрсынұлы тыныс белгілеріне ат қоюда, ғалымның өз сөзімен айтсақ, олардың «жұмсалатын орнын», яғни белгілі бір мәтін ішіндегі атқаратын қызметін негізге алады. Кейбір тыныс белгілеріне берген жанама атаулары арқылы олардың қызметін айқындап береді. Мәселен, сызықшаны *«үлкен сызықша»* деп атап оған *«жұмақтау белгісі»,* қос нүктеге *«бәшелеу белгісі»,* жақшаға *«қамау белгісі»* деген қосымша атаулар береді. Сол сияқты *«ұлы тыныс» (нүкте), «орта тыныс» (нүктелі үтір), «кіші тыныс» (үтір)* атауларынан тыныс белгілерінің жұмсалу орны мен қызметі айқын көрінеді.

Ғалым тыныс белгілерінің қызметі мен қолданатын орындарды олардың ережелерінде анықтайды [Байтұрсынұлы, 2013, 221-224 б.] (2-кесте).

*2-кесте*

*А.Байтұрсынұлының тыныс белгілері ережесі*

| № | **Тыныс белгі атауы** | **Ережесі** |
|---|---|---|
| 1 | **Кіші сызықша (тіркестіру белгісі)** | 1) Бір сөзді екінші сөзге тіркестіретін орында қойылады, сөзді олай тіркестіру екі сөзді қосақтап, бір сөз (қос сөз) қылып айтатын орындарда келеді. Мәселен, *төсек-орын, жүк-аяқ, құрт-құмырсқа*. 2) сөздің бір бөлімін екінші бөліміне тіркестіру керек болған орындарда қойылады, ондай тіркестіру көбінесе сөздің бөлімдері жазып келе жатқан жолға түгелімен сыймай, сыймаған бөлімін екінші жолға шығару қажет болған жерде келеді. Мысалы: *Жақсы-ақ-пысың? Жалғыз-ғана-мы-сың?* |
| 2 | **Үлкен сызықша (жұмақтау белгісі)** | Бұл белгі бытыранды ұғымдарды шоғырландырып, жалқыланған ұғымдарды жалпыландырып, теңелерлік нәрселерлі теңеп, балап айтатын орындарда қойылады. Мәселен, *Көк шалғын, ағаш, бұлақ – бәрі жақсы. Қасқыр – аң. Ақымақты үйрету – өлгенді тірілту.* |
| 3 | **Ноқат (нүкте) (ұлы тыныс)** | 1) әбден біткен ойлы сөйлемдерді бір-бірінен айыратын орында қойылады; 2) мақала, кітап басының атауларынан соң қойылады; 3) қысқартқан сөздің соңынан қойылады; 4) келте-келте қысқартылып айтқан сөйлемдердің соңынан қойылады. |
| 4 | **Үтірлі ноқат (нүкте) (орта тыныс)** | 1) сөйлемдердің ішкі жақындығы күшті болған жерде қойылады; 2) сыйысулы құрмалас сөйлемдердің баяндауыштары демеулермен |

| | | |
|---|---|---|
| | | құраспай, алды-алдына тұрған жана да қастарында тұрлаусыз сөйлем мүшелері немесе бағыныңқы сөйлем болған орында қойылады. |
| 5 | **Теріс үтір (кіші тыныс)** | 1) бұратана сөздерді бөлек шығару үшін қойылады; 2) сыйысулы сөйлемдердің бірөңкей мүшелерінің арасына қойылады; 3) бағыныңқы сөйлемнің жігін басыңқы сөйлемнен айыру үшін қойылады; 4) лепті сөйлемдерде одағайдан кейін қойылады; 5) бір сөз не болмаса сөйлем қайта-қайта айтылғанда қойылады. |
| 6 | **Қос ноқат (қос нүкте) (бәшелеу белгісі)** | 1) алдыңғы сөйлемдегі пікірді кейінгі сөйлем ыдыратып сөйлейтін орындарда қойылады. *Мысалы: Біздің екі үйіміз бар: жаз тігетін киіз үй, қыс кіретін там үй*; 2) келтірінді сөйлем алдында қойылады (*Сонда есек сөз айтады бұлбұл құсқа: Мақтаулы бар ғой әнпаз әрбір тұста...*). |
| 7 | **Сұрау белгісі** | Бұл белгі жауап сұраған сөйлемдердің соңынан қойылады. |
| 8 | **Леп белгісі** | 1) лепті сөйлемдерден соң қойылады; 2) тілекті сөйлемдердің бұйрық, өтініш түрлерінің соңынан қойылады; 3) тілекті сөйлемнің үгіт түріндегісінің қаратпа сөзділерінің соңынан қойылады. *Бар! Жүгір! Ұш!* |
| 9 | **Көп ноқат (көп нүкте) (қалдыру, тастау белгісі)** | Сөзді бүкпелеп жорта қалдырып сөйлегенде немесе сөздерді тізіп тұтастыра сөйлемей, үзіп бөлек-бөлек сөйлеген орындарда қойылады. |
| 10 | **Қабат үтір (қабатша қос тырнақ)** | 1) төл сөздердің жігін ашу үшін алды-артынан қойылады *(«Бекер ме» қасқыр айтты «менің сөзім»?);* 2) кітап, журнал, газет аттарына қойылады; 3) бір нәрсені сөз қылғанда, айырып айқын көрсету үшін қойылады (*леп белгісі «!» лепті сөйлемге, сұрау белгісі «?» сұраулы сөйлемге қойылады*); 4) бір нәрсені теріс мағанада атау үшін қойылады. Мәселен, *Жолдас емес адамды «жолдас» деп атау теріс атау болады.* |
| 11 | **Жақша (қамау белгісі)** | 1) бір нәрсені қабарында болдыра кетейін деген орындарда қойылады; 2) алдағы сөзді артқысы түсіндіре өтетін жерлерде қойылады. |

А.Байтұрсынұлы тыныс белгілерінің қолданылатын орындарын ережелердегі *«... орындарда келеді», «...қойылады»* сөздері арқылы нақтылап отырған.

Тыныс белгілерінің қойылатын орындары ғалымның «Тіл жұмсар» кітабында да қарастырылады. Мәселен, **тыныстық** (.) – дауыстың тынатын жеріне қойылатын белгі. **Жапсарлық** (,) – дауыс тынбай, тек сөз арасы көбірек ашыла айтылатын жерге қойылатын белгі. **Қосарлық** (-) – екі сөзді қосу керек болған орында, бір сөздің екі бөлімін қосу керек болған орында немесе біріне-бірі жедел екі сөз қосарынан айтылатын орында, не болмаса бір сөздің өзі қайта-қайта жедел айтылатын орында қойылатын белгі. **Тастарлық** (–) – жазбай тасталған сөздің орнына қойылатын белгі. **Дәлдеулік** (« ») – бұлжытпай дәлдеп жазып көрсетерлік жерде қойылатын белгі [Байтұрсынұлы, 2013, 496-497 б.]. Мұндағы назар аударатын нәрсе: «Тіл-құрал» мен «Тіл жұмсарда» кейбір тыныс белгілерінің түрліше аталуы. Мәселен, нүкте *«ұлы тыныс» - «тыныстық»*; үтір *«кіші тыныс» - «жапсарлық»;* кіші сызықша (дефис) *«тіркестіру белгісі» - «қосарлық»*; сызықша *«жұмақтау белгісі» – «тастарлық»;* тырнақша *«қабат үтір» – «дәлдеулік»*.

«Тіл-құрал» мен «Тіл жұмсардағы» тыныс белгі атауларының әркелкі берілуін ғалым өзінің пікірінде былай түсіндіреді: «Тіл-құрал» қазақ тілі қандай құрал екендігін тұтас түрінде таныту үшін түрлі бөлшектерін, тетіктерін ұсағын ұсағынша, ірісін ірісінше жүйелі тұрған орнында алып көрсетіп танытады. «Тіл жұмсар» сол үлкен құралдың бөлшектерін, тетіктерін балаға шағындап, бөлек-бөлек ойыншық сияқты құрал жасап, соларды танытып, соларды жұмсату арқылы барып үлкен құралды танытады. Бұл айтылған «Тіл-құралдан» «Тіл жұмсардың» *негізгі басқалығы*, мұнан өзге де *бала шамасына шағындағандықтан туған басқалықтар бар.* Оларды мұнда түгендеудің қажеті жоқ, үйткені «Тіл жұмсармен» үйреткенде, әр қайсысы өзінің орны-орнында көрінбекші. Мұнда тек басты-бастыларын ғана көрсете өтсек болады» [Байтұрсынұлы, 2013, 496-б.]. «Тіл жұмсар», атынан да көрініп тұр, «сөйлеу, оқу, жазу тілін жұмыс тәжірибесі арқылы танытатын» құрал. Басты мақсаты – «Тіл-құралдағы» білімді жаңа жолмен тәжірибе арқылы үйрету, баланың білімді өздігінен алуына ықпал ету. Сондықтан бұл еңбектегі «*басқалықтар»*, оның ішінде мақаланың зерттеу нысанына алынған тыныс белгі атауларындағы ерекшеліктер де осындай қажеттіліктен, баланың шамасына қарай шағындаудан туған.

Жалпы алғанда, қазақ тіліндегі тыныс белгі атаулары, олардың қызметі мен қолданысы, пунктуация ережелерінің негізі алғаш А.Байтұрсынұлының зерттеулерінде қалыптасты. Ғалым тыныс

белгілерін жазуда қолданатын әріптен басқа белгілер қатарында қарастырып, әрбір тыныс белгісінің таңбасы, қолданатын орны, сөйлем ішіндегі атқаратын қызметіне дейін нақтылап берді. А.Байтұрсынұлының ғылыми тұжырымдары, оның ішінде тыныс белгілері және оның ережелері де, қазіргі қазақ пунктуациясының ғылыми негізіне айналып отыр.

Бұл зерттеуді Қазақстан Республикасы Білім және ғылым министрлігінің Ғылым комитеті қаржыландырады (№BR11765535 грант).

**Әдебиеттер тізімі**

1 Байтұрсынов А. Тіл тағылымы. – Алматы: Ана тілі, 1992. – 448 б.

2 Байтұрсынұлы А. Қазақ тіл білімінің мәселелері. – Алматы: Абзал-Ай, 2013. – 640 б.

3 Қалиев Ғ. Тіл білімі терминдерінің түсіндірме сөздігі. – Алматы: Сөздік-Словарь, 2005. – 440 б.

4 Мұсабекова Ф. Қазіргі қазақ тілінің пунктуациясы. – Алматы: Ана тілі, 1991. – 128 б.

5 Салқынбай А., Абақан Е. Лингвистикалық түсіндірме сөздік. – Алматы: Сөздік-словарт, 1998. – 304 б.

6 Сыздық Р. Қазақ тілінің анықтағышы. – Астана: Елорда, 2000. – 532 б.

*ӘОК 004.932.72*
***[1]Амангелді Н , [2]Кудубаева С.А., [3]Турсынова Н.А.,***
***[4]Баймаханова А., [5]Ерболатова А., [6]Абдиева С.***
*Л. Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]nurzadaamangeldy@gmail.com, [2]kudubayeva_ sa @ enu.kz,*
*[3]ntursynova000@gmail.com*

# ҚАЗАҚ ЫМ ТІЛІНДЕГІ СӨЗДЕРДІҢ КӨРСЕТІЛУ ПІШІНДЕРІН ӨЗГЕ ЫМ ТІЛДЕРІМЕН САЛЫСТЫРМАЛЫ АНАЛИЗІ

**Аңдатпа.** Бұл мақалада қазақ ым тілін жеке ым тілі ретінде өмір сүре алатынын дәлелдеу мақсатында, қазақ, орыс, ағылшын, түрік ым тілдерімен салыстырмалы анализ жасалды. Ымды көрсету формасы конфигурация (қол / білек), орындау орны (локализация), қозғалыс бағыты, қозғалыс сипаты және қолмен жасалмайтын компонент (бет әлпеті мен артикуляциясы) тұрғысынан зерттеу жасалды. Зерттеу нәтижесінде 1045 сөздің 4 тілде көрсетілу формаларын бақылау нәтижесінде Қазақ ым тілі жеке ым тілі ретінде өмір сүре алатыны дәлелденді.

**Түйін сөздер:** ымдарды тану, адами-машиналық интерфейс, Виола-Джонс, корреляциялық талдау, сурдоаударма, эталон, ым тілі, артикуляция, локализация, конфигурация.

*УДК 004.932.72*
***[1]Амангелді Н, [2]Кудубаева С.А., [3]Турсынова Н.А.,***
***[4]Баймаханова А., [5]Ерболатова А., [6]Абдиева С.***
*Евразийский национальный университет имени Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]nurzadaamangeldy@gmail.com, [2]kudubayeva _ sa @ enu.kz,*
*[3]ntursynova000@gmail.com*

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПО ФОРМЕ ДЕМОНСТРАЦИИ СЛОВ КАЗАХСКОГО ЖЕСТОВОГО ЯЗЫКА С ДРУГИМИ ЖЕСТОВЫМИ ЯЗЫКАМИ

**Аннотация.** В данной статье проведен сравнительный анализ казахского, русского, английского, турецкого жестовых языков с целью доказательства того, что казахский язык жестов может существовать как отдельный язык жестов. Было проведено исследование формы отображения с точки зрения конфигурации (рука / предплечье), места

выполнения (локализация), направления движения, характера движения и компонента, который не может быть выполнен вручную (выражение лица и артикуляция). В результате исследования было доказано, что казахский жестовый язык может существовать как отдельный язык жестов наблюдая за формами демонстрации 1045 слов в 4 языках.

**Ключевые слова:** распознавание жестов, человеко-машинный интерфейс, Виола-Джонс, корреляционный анализ, сурдоперевод, стандарт, сурдоперевод, артикуляция, локализация, конфигурация.

**Кіріспе**. Ым (лат. gestus – дене қозғалысы) – белгілі бір мағынаға ие, яғни символ немесе эмблема болатын адам денесінің немесе оның бөліктерінің қозғалысы.

Ыммен сөйлеу дегеніміз - арнайы лексикалық және грамматикалық заңдылықтармен сипатталатын, есту қабілеті бұзылған адамдардың қимылдары қолдайтын тұлға аралық қарым-қатынас әдісі.

Ым тілі- бұл есту қабілеті қалыпты адамдар мен есту қабілеті бұзылған адамдар арасындағы вербалды емес байланыс жүйесі, ал соңғысы - бұл іс жүзінде әр сөзге сәйкес қимылды табуға болатын қарым-қатынастың негізгі әдісі ретінде қолданылады [1]. Ым тілінің негізгі бірлігі - бұл ым, яғни қолдың қимыл-қозғалысы, бет әлпет мимкасы мен артикуляциясы, басты бұру және т.б. көмегімен затты белгілеу мүмкіндігі, объектінің параметрлерін визуализациялау.

Көп жағдайда ым тілінің көмегімен есімдер мен тектерді, шетелдік, техникалық және медициналық анықтамаларды жеткізу мүмкін емес. Сондықтан ым тілімен қатар, оған қосымша ретінде саңыраулар (есту қабілеті мүгедектері) дактильді әліппені кеңінен қолданады. Дактильді тіл грамматикасы саңыраудың туған тілі грамматикасына ұқсайды. Дактилогоияны көбінесе саусақтармен ауада жазу деп айтуға болады: визульды қабылданады және жазбаша сөйлеу сияқты орфографияның барлық ережелеріне сүйенеді [2].

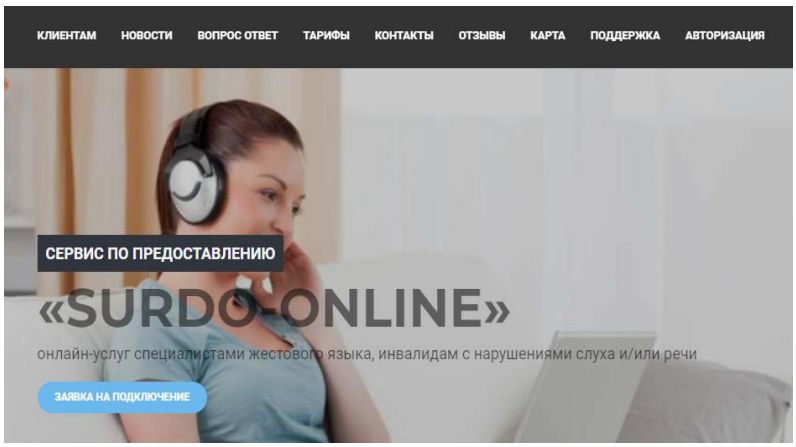
Бірақ тыныс белгілері емес: леп белгісі мен сұрақ белгісі тиісті мимика арқылы жеткізіледі; нүкте мен көп нүкте паузамен; сызықша, қос нүкте жіне басқа да тыныс белгілерінің өзіндік көрсету түрлері болғанымен, дактильді жазбада көрсетілмейді.

Ыммен көрсетуді параметрлеу үшін қимылды бес компонент ажыратады: конфигурация (қол / білек), орындау орны (локализация), қозғалыс бағыты, қозғалыс сипаты және қолмен жасалмайтын компонент (бет әлпеті мен артикуляциясы) [3].

**Салыстырмалы анализ:**

Бүгінгі таңда Қазақстанда есту қабілеті бұзылған науқастардың жалпы саны 200 мыңға жуық адамнан асты, оның ішінде 700 баланың есту мүшелері дамуының туа біткен ақаулары бар. Жыл сайын науқастар саны 5% - ға артады. 1000 жаңа туған нәрестеге бір саңырау баладан келеді, ал бір жасқа толған сайын бұл сан артады [4].

«Азаматтарға арналған үкімет» мемлекеттік корпорациясы» КЕАҚ фронт-офистерінде 2016 жылдан бері есту мен сөйлеу қабілеттері бұзылған азаматтарға мемлекеттік қызметтерді көрсету бойынша «Surdo-Online» сервис жұмыс істейді. Сервис есту қабілетімен проблемасы бар қызмет алушы, ыммен сөйлеу оператор және Мемлекеттік корпорация фронт-офисінің операторы арасындағы видеобайланыс болып табылады (сурет 1).




Сурет 1 «Surdo-Online» сервисінің интерфейсі

Емханаларда, жұмыспен қамту орталықтары мен басқармаларында, кейбір қонақ үйлерде, университеттерде, сақтандыру компанияларында, бизнес құрылымдарда, Алматы халықаралық әуежайында, қонақ үйлер мен «Магнум» дүкендер желісінде Surdo-Online сервисі жұмыс істейді [5].

Surdo-Online (SOL) – онлайн-режимде сурдоаудармашылармен бейне байланыс үшін бұлтты сервис. Оның көмегімен ұйымдар есту және/немесе сөйлеу қабілеті бұзылған мүгедектер үшін ең аз шығынмен қолайлы орта құра алады, ал мүгедектер – коммуникациялар мен қызметтерге қол жеткізе алады. Нақты уақыт режимінде есту және сөйлеу қабілеті зақымдалған адам сурдоаудармашымен байланысқа шығады, ол оның сөздерін хабарласып жатқан адамға дыбыстап айтады. Және, керісінше, мүгедек хабарласып отырған компания қызметкерінің сөздері ым-қимыл тіліне аударылады.

Сурдоаудармa саласын толығырақ зерттеу үшін есту қабілеті бұзылған адамдардың қалыпты өмірге бейімделуіне көмектесетін құралдарды қарастыру қажет. Солардың бірі Қазақстандағы есту қабілеті шектеулі азаматтарға арналған «Қол қимыл әлемі» деп аталатын сөздік (сурет 2).


Сурет 2 «Қол қимыл әлемі» кітабы

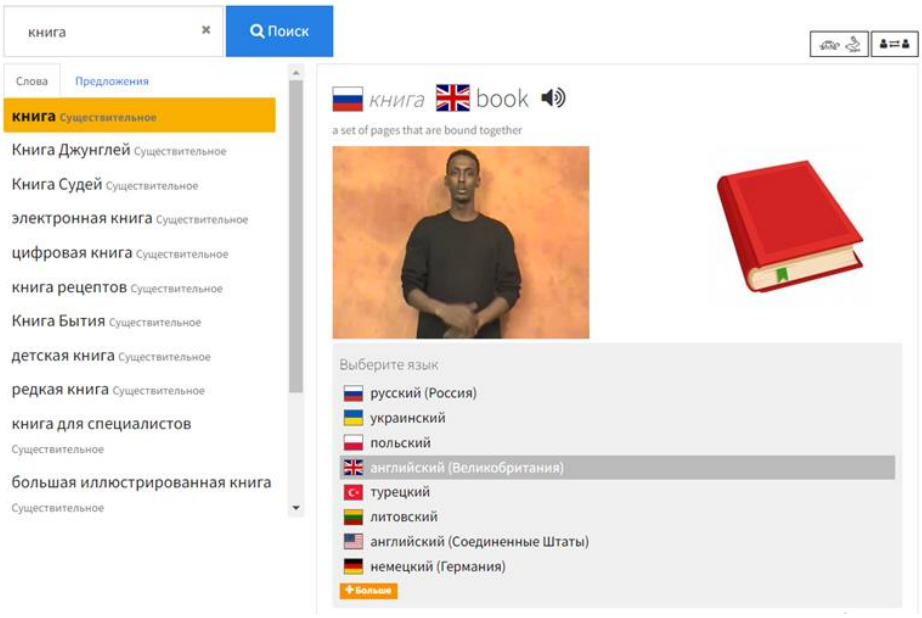
«Үміт» есту қабілеті бойынша мүгедектерді қолдау орталығының көркем суретті, фотокадрмен бейнеленген, түсініктеме сапасы жоғары, мазмұнды қол қимыл сөздігін қазақ, орыс және ағылшынша тілінде шығарған дүниесі. Бұл сөздік кітабы – есту қабілеті шектеулі азаматтарға арналған. Кітаптің мақсаты – ымды тіл белгісі келген адамдарға тәжірбиелік көмек [8]. Бұл кітапта Қазақстан мемлекетіндегі есту қабілеті шектеулі азаматтардың күнделікті пайдаланатын ым тілі жазылған. Негізгі мақсаты – осы тілді пайдалатын азаматтар арасында тілді дамыту, дұрыс ым белгілерімен араласуына, қол саусақпен сөйлесуді оны өркендетуіне көмек көрсету. Кітапта есту қабілеті шектеулі азаматтарға, есту мүмкіншілігі шектеулі адамдармен араласуға жеткілікті ымды сөздер сөздікте жазылып, суретте бейнеленген [5].

Сурдоаудармаға арналған қосымшалардан басқа веб-ресурстар да бар. Веб-ресурстар – саңырау адамдарға қатысты ресурстарды орналастыратын мамандандырылған сайттар: мақалалар, әсіресе саңырау адамдарға қатысты заң жобалары, ымдау тілін үйретуге арналған материалдар мен сабақтар және дактил алфавиті. Ең танымал ресурстардың бірі,ол–surdo.kz сайты [6]. Сурдосервер қазақ ым тілінің және әлемнің ым тілдері онлайн қол жеткізудің ресурстарына саңырауларға және нашар еститін адамдарына және барлық қалағандарға көмегі үшін құрастырылған.

Қазақ ым тілінің жеке ым тілі болып өмір сүре алатынын дәлелдеу үшін орыс, ағылшын, түрәк тілдерімен салытырмалы анализ жасау керек. Қазақстанда саңырау адамдардың немесе есту қабілеті нашар адамдардың саны шамамен жалпы халық санының 1.6% құрайды. Бұл дегеніміз шамамен 300 мың адам. Олар басқа адамдармен қарым-қатынас жасауды өз бетімен үйрене алмайды, өйткені оқу процесі қарым-қатынаспен байланысты. Кітап, газет оқу да белгілі бір қиындықтар туғызады. Адаммен байланыспаған кезде, ол белгілі бір сөздердің қалай айтылатынын естімейді, содан кейін ол оларды оқи алмайды. Телебағдарламаларды субтитрсіз көру де мүмкін емес. Бұл мәселемен айналысатын көптеген мамандандырылған мектептер мен мекемелер бар. Әңгімелесуде саңырау-мылқау адам барлық назарын әңгімелесушіге аударады, өйткені ол сөздерді аузынан оқиды. Есту және саңырау адамдар арасында сөйлесу үшін ымдау тілі қолданылады және ол өте қажет. Ол мимиканы, ым-ишараны біріктіреді. Ым тілі мамандарын даярлау өте еңбекті қажет ететін процесс, ол маманнан жауапкершілікті талап етеді [9]. Бірақ ғалымдардың зерттеулері тілдің туғаннан бастап бізге тән инстинкт екенін растайтындықтан, декомпенсация, белгілі бір екінші ақауға бейімделу жүреді. Бұл декомпенсацияның нәтижесі: инстинкт (қарым-қатынастың туа біткен қажеттілігі) + қимыл + мимика + көру = ымдау тілі! Коммуникативті функциялар дыбыстық тілдің орнына ымдау тілін алады [7].

Ымдау тілінде, мысалы, феминизмді білдіретіндей қимыл жоқ. Бұл тақырыпты түсінетіндер арасында түсінік болуы мүмкін. "Парламент" және "депутат" сияқты сөздерде жоқ. Бұл сөздерді түсіндіру үшін бірнеше қимылдарды қолдану керек, мысалы, "министрлік бар, адамдар отырады және талқылайды" деген секілді бірнеше қимылдың арқасында түсіндіреді.

Салыстырмалы анализ жүргізу барысында «Қол қимыл әлемі», surdo.kz және де әлемдік ым тілдерге арналған spreadthesign.com ресустарын қолданып, сөздің қызық ағылшын орыс түрік тілінде көрсетілу формалары зерттелді. Себебі көптеген интернеттегі ресуртардың айтуынша «Қазақтың ым тілі – миф. Қазақта дактиль – әріпті білдіретін ым бар. Жаңа сөздер жоқ» деген пайымдаулар көптеп жүреді. Қазақ ым тілі орыс ым тілінен шыққан, жек тіл емес деген пайымдаулар да бар.




Сурет 3 spreadthesign.com ресурсы

Spreadthesign ресурсы көптеген тілдерде бір сөздің көрсетілу фрормасын бақылауға ыңғайлы болғандықтан аталмыш ресурс зерттеу жүргізуге пайдаланылды.

Осындай пайымдауларға жауап ретінде және қазақ ым тілінің жеке тіл бола алатынын дәлелдеу мақсатында жалпы көлемі 1050-дей сөзге келесі кестедегідей салыстырмалы анализ жасалды.

Бастапқы кесте

| № | Сөздер | Қазақ | Орыс | Ағылшын | Түрік |
|---|---|---|---|---|---|
| 1. | Артикуляция | + | + | - | - |
| 2. | Ауру | + | + | - | - |
| 3. | Бас | + | + | - | - |
| 4. | Бетжүзі | + | + | + | + |
| 5. | Дене | + | + | - | + |
| 6. | Денсаулық | + | + | + | - |
| 10. | Жүрек | + | + | + | + |
| 11. | Жұқпалы ауру | * | * | - | |
| 12. | Аяқ-киім | - | - | - | - |
| 13. | Әйелдер қалпағы | - | - | - | - |
| 14. | Белдемше (юбка) | + | - | - | + |
| 15. | Бетперде (маска) | - | - | - | - |
| 16. | Бәтенке | - | - | - | - |
| 17. | Етік | - | - | - | - |
| 18. | Киім | - | - | - | - |
| 19. | Костюм | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| 20. | Қалпақ | + | - | - | + |
| 21. | Қолғап | - | - | - | - |
| 22. | Мода | - | - | - | - |
| 23. | Сүлгі | - | - | - | - |
| 24. | Өкше | - | - | - | - |
| 25. | Спорт жейдесі | - | - | - | - |

+ бәрінде көрсетілу формасы бірдей болған жағдайда;
- мүлдем ұқсастық болмаған жағдайда;
***** - өте қатты ұқсас, бірақ соңында немесе басында сәл өзгеріс бар

Мысалы жоғарыдағы кестеде “Аяқ-киім”, “Қолғап”, “Мода”, “Сүлгі”, “Өкше” деген сөздер төрт тілде көрсетілу формасы бірдей емес, “Артикуляция”, “Ауру”, “Бас”, “Бетжүзі”, “Дене”, “Денсаулық”, “Жүрек” деген сөздердің қызық тілі мен орыс тіліндегі көрсетілу формасы бірдей. “Жұқпалы ауру” деген сөздің көрсетілу формасы ұқсас, бірақ элементтерінде кішкене өзгерістер бар. “Белдемше”, “Қалпақ” сөздері қазақ, түрік тілдерінде бірдей.

1050 сөздің 300 жуығы, 38 пайызы ешбір тілге ұқсастығы жоқ сөздер. Еш тілге ұқсамайтын сөздер кестесінен үзінді.

Ұқсастығы жоқ сөздер кестесі

| № | Сөздер | Қазақ тілі | Орыс тілі | Ағылшын тілі | Түрік тілі |
|---|---|---|---|---|---|
| 3 | Күз | - | - | - | - |
| 4 | Қараңғы | - | - | - | - |
| 5 | Өлшем | - | - | - | - |
| 6 | Сағат | - | - | - | - |
| 7 | Секунд | - | - | - | - |
| 8 | Сәрсенбі | - | - | - | - |
| 9 | Шілде | - | - | - | - |
| 10 | Жүк | - | - | - | - |
| 11 | Қатты | - | - | - | - |
| 12 | Тар(узкий) | - | - | - | - |
| 13 | Алтын | - | - | - | - |
| 14 | Альбом | - | - | - | - |
| 15 | Ақжелкен (петрушка) | - | - | - | - |
| 16 | Бомба | - | - | - | - |
| 17 | Жақсы | - | - | - | - |
| 18 | Жақын | - | - | - | - |
| 19 | Зиян | - | - | - | - |

| | | | | | |
|---|---|---|---|---|---|
| 20 | Келесі | - | - | - | - |
| 21 | Маңызды | - | - | - | - |
| 22 | Мәңгі | - | - | - | - |
| 23 | Соңғы | - | - | - | - |
| 24 | Ілгіш | - | - | - | - |

Осы еш тілге ұқсамайтын, тек қана көрсетілу қазақ тілінде бар сөздер, қазақ ым тілінің жеке тіл болуының дәлелі бола алады.

Сонымен қатар барлық тілде көрсетілу формалары бірдей сөздер болды, олар жалпы сөздердің 15 пайызын құрды.

Барлық тілде көрсетілуі бірдей кесте

| № | Сөздер | Қазақ тілі | Орыс тілі | Ағылшын тілі | Түрік тілі |
|---|---|---|---|---|---|
| 3 | Сорпа | + | + | + | + |
| 4 | Қайық | + | + | + | + |
| 5 | Мақтаныш | + | + | + | + |
| 6 | Сену | + | + | + | + |
| 7 | Сот | + | + | + | + |
| 8 | Баскетбол | + | + | + | + |
| 9 | Билеу | + | + | + | + |
| 10 | Бокс | + | + | + | + |
| 11 | Волейбол | + | + | + | + |
| 12 | Стадион | + | + | + | + |
| 13 | Теннис | + | + | + | + |
| 14 | Бояу | + | + | + | + |
| 15 | Араласу | + | + | + | + |
| 16 | Бас тарту | + | + | + | + |
| 17 | Есту | + | + | + | + |
| 18 | Жоғары | + | + | + | + |
| 19 | Кездесу | + | + | + | + |
| 20 | Кезек | + | + | + | + |

Жалпы сөздердің 10 пайызыныі орыс тілімен мүлдем сәйкестігі жоқ екені анықталды

Қазақ ым тілінің сөздерінің орыс ым тілімен сәйкестік кестесі

| № | Сөздер | Қазақ | Орыс |
|---|---|---|---|
| 1. | Дәрі | * | * |
| 2. | Емдеу | * | * |
| 3. | Жұқпалы ауру | * | * |

| | | | |
|---|---|---|---|
| 4. | Кефир | * | * |
| 5. | Жаңғақ | * | * |
| 6. | Қаймақ | * | * |
| 7. | Чеснок | * | * |
| 8. | Сүт | * | * |
| 9. | Аққайын | * | * |
| 10 | Жапырақ | * | * |
| 11 | Жұлдыз | * | * |
| 12 | Орман | * | * |
| 13 | Өзен | * | * |
| 14 | Терең | * | * |
| 15 | Теңіз | * | * |
| 16 | Телефон | * | * |

Жоғарыда айтып өткендей қазақ ым тілі орыс ым тіліне өте ұқсас болып келеді. Бірақ – та біздің қазақ тіліне ғана қатысты сөздер бар, орыс тілінде кездеспейтін. *Қарастырған 1050 сөзге әр қайсысына талдау жасалды*, сол сөздердің ішінде тек қазақ тіліне тиісі сөздерде болды соларға мысал келтіріп кететін болсақ:

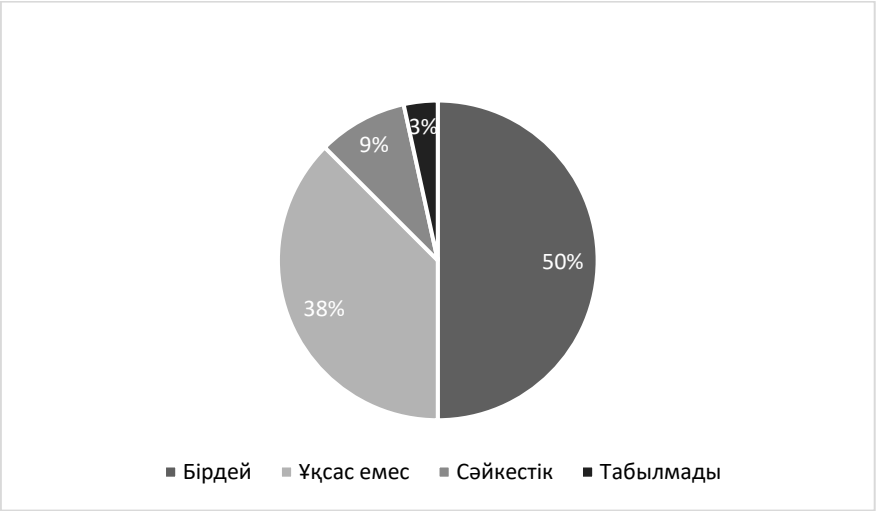
1. Күйкелек
2. Кеден
3. Кебіру
4. Жібеуші
5. Жирен
6. Ақсақал
7. Тоқал
8. Бәйбіше
9. Отағасы.

Ым тіліне аударманың түрлерін қарастырмас бұрын, ымдау тілі феноменіне және оның әлеуметтік-мәдени ерекшеліктеріне тағы да қысқаша тоқталу қажет. Қазақ ымдау тілін құрайтын калькалық таңбалық сөйлеу мен ауызекі сөйлеу жүйелерінің арасында түбегейлі айырмашылықтар болғандықтан, аудармашылық пен аударманың мәселелерімен, пікірлердің қайшылықтарына қарамастан, екі тілдік және әлеуметтік мәдени тамыры екеніне көз жеткізуге болады.

**Қорытынды**

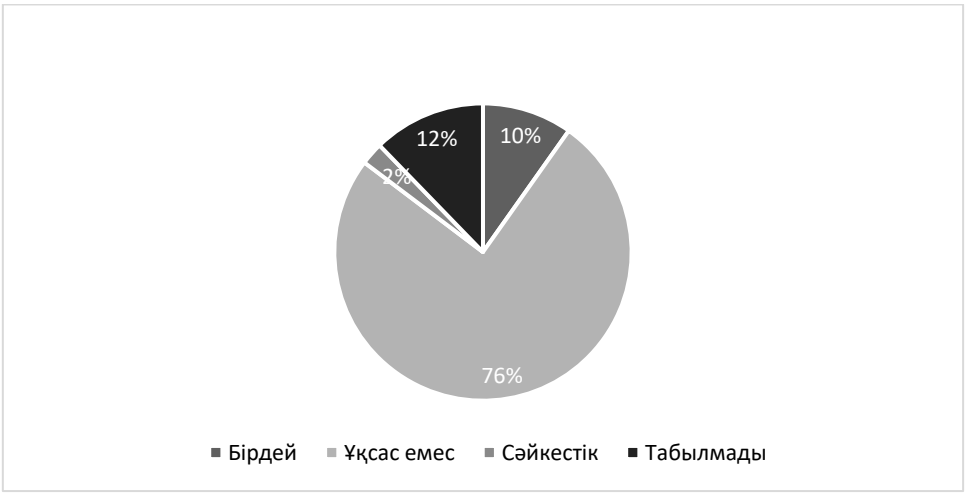
Қорыта келе зерттелген 3 тілмен қазақ ым тіліндегі сөздердің ұқсастығы диграммамен беруге болады. Диаграмманы көріп тұрғандарыңыздай, қарастырылған сөздірдің 50% орыс ым тілімен бірдей болғанын көріп тұрмыз. 38% тек қазақ тіліне тиісті басқа ым тілдеріне ұқсас емес екенін көріп отырмыз. Ерекшеліктерде болды, соларды біз сәйкестік деп алдық, ол 9% құрады. Қазақ орыс

аудармалары ұқсамай, іздеу барысында табылмаған сөздерде болды. Ендігі қазақ ым тілін келесі ым тілдерімен салыстырайық (сурет 8).



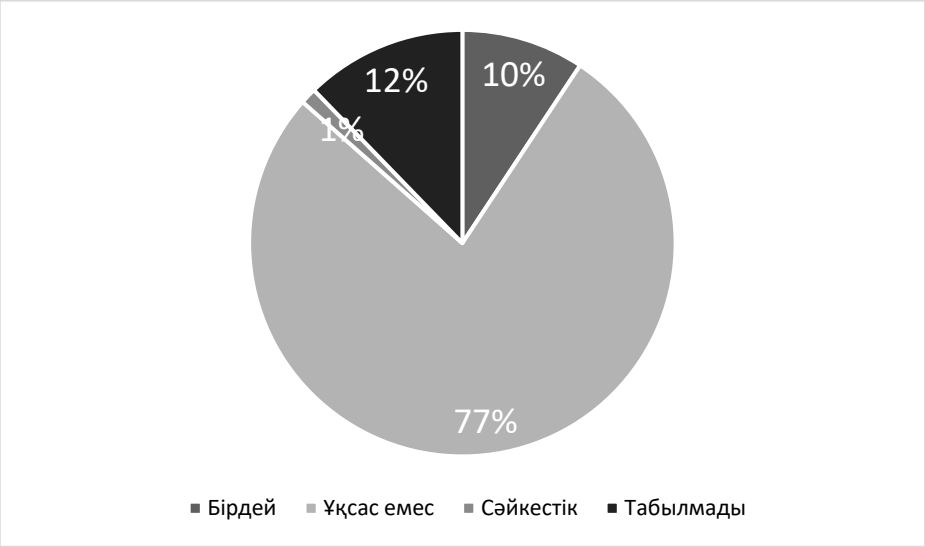

Сурет 4 Орыс ым тілі

Келесі диаграммаға кезек беретін болсақ, байқағаныздай ағылшын ым тілімен қазақ ым тілінің бірдейлігі 10 пайыздарды құрайды. 76% көрсеткіш ұқсас емес дегенді білдіреді, осыдан келе қазақ ым тілінің қалыптасуына ағылшын ым тілінің мүлдем қатысы жоқ екенін көреміз. Орыс ым тілі секілді ізденіс барасында, аудармалар дұрыс болмай 12%-дық көрсеткішке тиісілі сөздер табылмады (сурет 9).




Сурет 5 Ағылшын ым тілі

Түрік тілдес халықпыз ұқсас сөздеріміз көп болады деген мақсатпен жүргізілген талдау нәтижесі түрік-қазақ ым тілдерінің ұқсастығы өте аз көрсеткішті көрсетеді.




Сурет 6 Түрік ым тілі

Осы жасалған салыстырмалы талдау нәтижесінде Орыс ым тілімен 50 пайыздық ұқсастығына қарамастан Қазақ ым тілі жеке тіл деп айтуға болады, себебі өз кезегінде орыс ым тілінің, зерттеу барысындағы 1050 сөздің, 30-пайызға жуығы ағылшын тілінен алынған. Ым тілдер сөздік қоры табиғи тілдер сөздік қорынан көптеген есе аз болғандықтан және есту қабілеті бойынша мүгедектер өзара сөйлескенде ортасына орай, әңгімеге бейімдеп жаңа сөздер шығара беретін болғандықтан Қазақ ым тілі сөз қоры бойынша, өзіндік ерекшелігі бар жеке ым тілі деген қорытынды шығаруға болады.

**Әдебиеттер тізімі**

1. Димскис Л.С. Изучаем жестовый язык. М.: Изд. центр «Академия», 2002. 128 с.

2. Н. Амангелді, С.Ә. Кудубаева Қазақ ым тілін тану есебінің пән облысына шолу. Есептің қойылуы, ҚазҰТУ хабаршысы. № 5, 2020.

3. Зайцева Г.Л. Жестовая речь. Дактилология. М.: Владос, 2000. 192 с.

4. Kudubayeva S, Amangeldy N. The use of correlation analysis in the algorithm of dynamic gestures recognition in video sequence. ICEMIS 2019 (The International Conference on Engineering & MIS 2019). L.N.Gumilyov Eurasian National University Astana, Kazakhstan.

5. https://surdo-online.kz/

6. www.surdo.kz. Режим доступа 01.09.2019.

7. Н.Амангелді, С.Ә. Кудубаева. Қазақ ым тіліндегі сөз тіркесін танудың байланысқан облыстарды белгілеу және корреляциялық әдістері. ҚазҰТУ хабаршысы. № 5,2020

8. «Үміт» саңырау мүгедектерді қолдау орталығы. Қол қимыл әлемі. Қазақстандағы есту қабілеті шектеулі азаматтарға арналған сөздік. Алматы, «Үміт», 2007 жыл – 404 б.

9. Н. Амангельды, Ю.В. Крак, С. А. Кудубаева. Классификация форм демонстрации жестов на основе онтологической модели предметной области. 2020 IEEE International Conference on Advanced Trends in Information Theory ATIT / Kyiv / Ukraine

# МӘТІНДЕРДІ МОРФОЛОГИЯЛЫҚ, СИНТАКСИСТІК ЖӘНЕ СЕМАНТИКАЛЫҚ ӨҢДЕУ ТЕХНОЛОГИЯЛАРЫ

# ТЕХНОЛОГИИ МОРФОЛОГИЧЕСКОЙ, СИНТАКСИЧЕСКОЙ И СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

# TECHNOLOGIES FOR MORPHOLOGICAL, SYNTACTIC AND SEMANTIC TEXT PROCESSING

*ӘОК 004.912*

***[1]Сайранбекова А. Д., [2]Бекманова Г.Т.***
*Л. Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]sairanbekova98@gmail.com, [2]gulmira-r@yandex.kz*

## МАШИНАЛЫҚ АУДАРМА НЕГІЗІНДЕ МӘТІНДЕГІ ЖАҒЫМСЫЗ СЕНТИМЕНТТІ АНЫҚТАУ

**Аңдатпа.** Сентимент талдау, мәтіндегі оң немесе теріс көзқарастарды автоматтандырылған түрде анықтау соңғы онжылдықта зерттеушілердің назарын аударды. Сонымен қатар, Интернеттегі шолу сайттарын, әлеуметтік желілерді және жеке блогтарды өз пікірлерін білдіру үшін белсенді қолданатын интернет қолданушыларының танымалдығы жаңа технологиялармен қатар тез өсуде. Осы пайдаланушылар жазған пікірлер белгілі бір адамдарға, ұйымдарға, орындарға, оқиғалар мен идеяларға оң және теріс көзқарас қалыптастыруға ықпал етеді. Табиғи тілдерді өңдеу және машиналық оқыту құралдары, сондай-ақ мәтіннің үлкен көлемімен жұмыс істеудің басқа тәсілдері әлеуметтік желілердегі әртүрлі көңіл-күйлерді анықтауға мүмкіндік береді. Берілген мақалада сентимент талдаудың маңызын, оны жүзеге асыратын программалар түрлерін, түркі тілдеріндегі сентимент талдау ахуалын және машиналық аударма негізінде жасалған мультитілді «Negative Mood» сентимент талдау программасының жұмыс алгоритмімен танысатын боламыз.

**Кілттік сөздер:** сентимент талдау, әлеуметтік желі, машиналық аударма, мультитілді.

*УДК 004.912*
***[1]Сайранбекова А. Д., [2]Бекманова Г.Т.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]sairanbekova98@gmail.com, [2]gulmira-r@yandex.kz*

# ОПРЕДЕЛЕНИЕ НЕГАТИВНОГО СЕНТИМЕНТА В ТЕКСТЕ НА ОСНОВЕ МАШИННОГО ПЕРЕВОДА

**Аннотация.** Сентиментальный анализ, автоматизированное выявление положительных или отрицательных точек зрения в тексте привлекло внимание исследователей в последнее десятилетие. Кроме того, популярность интернет-пользователей, которые активно используют сайты онлайн-обзора, социальные сети и личные блоги для выражения своего мнения, быстро растет наряду с новыми технологиями. Отзывы, написанные этими пользователями, способ-ствуют формированию позитивного и негативного отношения к определенным людям, организациям, местам, событиям и идеям. Средства обработки естественных языков и машинного обучения, а также другие способы работы с большим объемом текста позволяют выявить различные настроения в социальных сетях. В данной статье мы познакомимся с понятием сентиментального анализа, видами реализуемых программ, состоянием сентиментального анализа на тюркских языках и алгоритмом работы мультиязычной программы сентиментального анализа «Negative Mood», созданной на основе машинного перевода.

**Ключевые слова:** сентиментальный анализ, социальная сеть, машинный перевод, мультиязычный.

*UDC 004.912*
***[1]Sairanbekova A., [2]Bekmanova G.***
*L.N.Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]sairanbekova98@gmail.com, [2]gulmira-r@yandex.kz*

# DETERMINATION OF NEGATIVE SENTIMENT IN A TEXT BASED ON MACHINE TRANSLATION

**Abstract**. Sentimental analysis, automated identification of positive or negative points of view in the text has attracted the attention of researchers in

the last decade. In addition, the popularity of Internet users who actively use online review sites, social networks and personal blogs to express their opinions is growing rapidly along with new technologies. The reviews written by these users contribute to the formation of positive and negative attitudes towards certain people, organizations, places, events and ideas. Natural language processing and machine learning tools, as well as other ways of working with a large volume of text, allow you to identify different moods in social networks. In this article we will get acquainted with the concept of sentimental analysis, the types of programs implemented, the state of sentimental analysis in Turkic languages and the algorithm of the multilingual sentimental analysis program «Negative Mood», built on the basis of machine translation.

**Keywords:** sentimental analysis, social network, machine translation, multilingual.

## 1. Кіріспе

Маңызды зерттеулер жүргізгенде немесе күнделікті шешімдер қабылдағанда, мүмкін өзіміз оны байқамай-ақ, жиі басқа адамдардың пікіріне жүгінетініміз рас. Саяси дауыс беру кезінде саяси пікірталас алаңдарынан кеңес аламыз, тұрмыстық техниканы сатып алғанда тұтынушылардың есептерін оқимыз, достарымыздан кешке қандай мейрамханаға баруға болады деп сұраймыз.

Ал ендігі уақытта Интернет соңғы трендте қандай гаджет екенінен бастап саяси пікірлерге дейін миллиондаған адамдардың пікірін білуге мүмкіндік береді. Pew компаниясының Интернет және азаматтық белсенділік бойынша соңғы зерттеуінде «интернет пайдалану-шыларының бестен бір бөлігі ғана (19%) саяси немесе әлеуметтік мәселе туралы мазмұнды пост жариялады немесе азаматтық, саяси қатысудың қандай да бір түрі үшін әлеуметтік желі сайтын пайдаланды» делінген.

Саяси мәселеде интернетті тиімді пайдалану туралы Мемлекет басшысы Қасым-Жомарт Тоқаев 2022 жылы 16 наурызда Қазақстан халқына «Жаңа Қазақстан: жаңару мен жаңғыру жолы» атты жолдауында былай дейді: «Байланыс технологиялары қарқынды дамып жатқан қазіргі заманда кандидаттар мен партиялардың әлеуметтік желідегі белсенділігінің маңызы зор. Бірақ, әлеуметтік желідегі үгіт-насихат қолданыстағы заңнама арқылы реттелмеген. Соған қарамастан сайлау науқаны кезінде онда үгіт-насихат жұмыстары бәрібір жүргізіледі. Осы олқылықтың орнын толтыру үшін тиісті регламент пен ережені бекіте отырып, әлеуметтік желіде үгіт-насихат жүргізуге заң бойынша рұқсат беруді ұсынамын». Келешеке осы бағыттарғы

жұмыстар артатыны рас болса, онда әлеуметтік желілердегі мәтіндерді анализдеу қажеттілігі де жоғарылай түсетіні кәміл.

Тағы бір зерттеулерде интернет қолданушылардың үштен бірі (33%) блогтарды оқитынын, ал 11%-ы ондай блогтартарды күнделікті оқитынын көрсетеді. Интернет барған сайын адамдар үшін пікірталас алаңына және ақпарат көзіне айналуда. Пікір – мәтінді талдаудың жаңа өрісін құрды. Соңғы онжылдықта мәтіннен көңіл-күйді анықтау өнеркәсіпте де, ғалымдар арасында да үлкен назар аударуды талап етті. Көптеген компаниялар интернетті пайдаланушылардың өз өнімдері мен қызметтері туралы пікірлерінің маңыздылығын түсінді.

Сентимент-талдау (sentiment analysis) — мәтінмайнинг (text mining) бөлімі, мәтіннен субъективті пікірлерді автоматты түрде алу жүйесі, ақпаратты іздеу және есептеу лингвистикасы тоғысындағы пән, ол мәтіннің мазмұндылығын емес оның тоналдылығын зерттейді [6].

## 2. Сентимент анықтауға арналған программалық шешімдер

Бүгінгі таңда ғылыми қол жетімді программалық инструменттер жиырмадан астам алгоритмдерді қамтиды. Олардың кейбіреулері бірнеше тілдерге арналған дайын программалық өнімдер ретінде танылған. Мұндай программалық пакеттің мысалы Британдық SentiStrength [7], ол сентимент пен оның әсерін анықтау үшін қолданылады және бүгінде ағылшын, испан және басқа да еуропалық тілдермен жұмыс істейді. Оның аналогы SocialMention пакеті.

Қазіргі уақытта көңіл күйді анықтау түркі тілдері үшін тың тақырыптардың бірі, соның ішінде қазақ тілі де бар. Оған мысал ретінде төмендегі кестені қарастырсақ болады.

Кесте 1. Сентимент талдауды жүзеге асыратын құралдардың қолдайтын тілдерінің салыстырмалы кестесі

| № | Құрал | Тілдер | Түркі тілдері |
|---|---|---|---|
| 1 | Brand 24 | Ағылшын, араб, хорват, чех, дат, голланд, фин, француз, неміс, венгр, индонезия, итальян, корей, норвег, поляк, португал, румын, орыс, словак, испан, швед, тай, түрік | түрік |
| 2 | Repustate | Ағылшын, араб, португал, неміс, голланд, дат, итальян, швед, фин, норвег, поляк, орыс, француз, | түрік |

| | | | |
|---|---|---|---|
| | | тай, корей, испан, урду, қытай, түрік, иврит, малайзия, жапон және индонезия тілдері. . | |
| 3 | Magellan Text Mining languages | Араб, қытай, голланд, ағылшын, француз, неміс, иврит, итальян, жапон, португал және испан, болгар, каталан, хорват, чех, дат, эстон, фин, грек, венгр, ирланд, исланд, латыш, литва, норвег, Парсы, поляк, румын, орыс, словак, словен, швед, түрік, украин, вьетнам | түрік |
| 4 | Social Mention | Араб, армян, белорус, болгар, каталан, қытай (жеңілдетілген), қытай (дәстүрлі), хорват, чех, дат, голланд, ағылшын, эсперанто, эстон, филиппин, фин, француз, неміс, грек<br>Еврей, венгр, исланд, индонезия, итальян, жапон, корей, латыш, литва, норвег, парсы, поляк, португал, румын, орыс, серб, словак, словен, испан, швед, тай, түрік, украин, вьетнам | түрік |
| 5 | SentiStrength | Фин, неміс, голланд, испан, орыс, португал, француз, араб, поляк, парсы, швед, грек, уэльс, итальян, түрік. | түрік |

Жоғары келтірілген құралдар әлімдегі танымал 20 құралға кіретін үздік инструменттер ретінде бағаланған. Соның ішінде түркі тілдерінен тек түрік тілі ғана бар екендігіне көз жеткізе аламыз. Қазақ тілі үшін бұл сала әлі де тың екенін көреміз.

Қазақ тілі бойынша табиғи тілді өңдеу мәселелерімен Қазақстанда А.Ә. Шәріпбай, У.А. Тукеев, Г.Т. Бекманова, Б.Ш. Разахова, Д.Р. Рахимова, Ө.Ж. Мамырбаев, Ж. А. Есенбаев, М.Х. Карабалаева, А.С. Муканова, А.К. Бурибаева, А.Ө., Макажанов, Ж.М. Жуманов және т.б. сынды ғалымдар айналысады [19].

### 3. Машиналық аударма негізіндегі «Negative Mood» мультитілді сентимент талдау программасын құру

Қазіргі уақытта әлемдегі ең танымал сентимент талдау программалары мен зерттеу жұмыстары негізінен ағылшын тілінде екенін байқауға болады. Жоғарыда келтірілген зерттеу нәтижелерінен түркі тілдес тілдердің ішінде тек түрік тілі ғана қамтылғанын көре аламыз. Қазақ тілі үшін мұндай программаларды жасау енді қолға алынып, сентимент сөздіктері жасалып жатыр. Ал мұнда ұсынылған жұмыс негізінен қазақ тіліндегі мәтіндердің (басқа да тілдер кіріктірілген) көңіл-күйін талдау құралын осы тілдің лингвистикалық ресурстарын барынша аз пайдалану арқылы жасау қажеттілігімен негізделген. Мұндай программа құру үшін толығымен қалыптасқан машиналық аударма жүйесі қолданылады, яғни программа мәтінді танып болғанан кейін оны ағылшын тіліне аударады, сосын ондағы жағымсыз сөздерді анықтап оның теріс бағасын беру үшін берілген формуламен есептейтін болады.

Жағымсыз көңіл-күйді анықтау үшін ең алдымен жағымсыз сөздер жинақталған және реңктелген деректер базасын құрамыз. Оның құрылымын төмендегі кестеден көруге болады.

Кесте 2. Жағымсыз сөздердің категориялары мен реңк көрсеткіштері

| № | Жағымсыз сөздер категориялары | Реңктік көрсеткіші | Қасиеті |
|---|---|---|---|
| 1 | Жағдайды сипаттау (қорқынышты оқиға, жантүршігерлік жағдайлар және т.б.) | 1 балл | Іс әрекетке қабілетсіз |
| 2 | Адамның көңіл-күйіне әсер ету (сені жек көремін, ашуыма тиесің және т.б.) | 2 балл | |
| 3 | Қорқыту шаралары (өлтіремін, тұншықтырамын және т.б.) | 3 балл | |
| 4 | Болған оқиға (тонап кетті, жаралады және т.б.) | 4 балл | Іс-әрекетке қабілетті |
| 5 | Адам өмірі мен мемлекет қауіпсіздігі (адам өлтіру, мемлекеттік төңкеріс жасау және т.б.) | 5 балл | |

Деректер базасы құрылып болғанан кейін, программа негізделетін формуланы шығарамыз. Мәтіндік негативті есептеу формуласы:

$$(0{,}7 \times N + 0{,}3 \times NC) \times 100\% = NS \quad (1)$$

NS – Теріс баға

TNT(Total Negative Threshold) – Жалпы теріс шек (Общий Отрицательный Порог) - 70%. Жалпы теріс шек 70 %-ға тең немесе одан жоғары болған жағдай да ғана мәтін жағымсыз ретінде танылады.

$$N = \frac{\Sigma NWV}{const\ NT} \quad (2)$$

N – Теріс шек. 1-ге дейін жуықталған, жағымсыз сөздердің реңк көрсеткіштері суммасының жағымсыз сөздер реңк көрсеткіштерінің шегіне қатынасы.

NWV (Negative Word Values) – жағымсыз сөздердің реңк көрсеткіштері.

NT (Negative Threshold) Жағымсыз сөздер реңк көрсеткіштерінің шегі (Порог для отрицательных значений) – 5.

$$NC = \frac{\Sigma NW}{const\ NTC} \quad (3)$$

NC – Теріс санау шегі. 1-ге дейін жуықталған, жағымсыз сөздердің жалпы санының жағымсыз сөздер санының шегіне қатынасы.

NW (Negative Word) – жағымсыз сөздер.

NTC (Negative Threshold Count) Жағымсыз сөздер санының шегі (Порог для количества отрицательных слов) – 3. Бұл жалпы бағалауға әсер етуі мүмкін сөздердің максималды саны. NC-нің бағалауға толық әсер етуі үшін сөздердің саны NTC-ге тең немесе одан көп болуы керек.

Осы алгоритм бойынша мәтінге сентимент талдауды қарастырайық. Мәселен, елімізде болған соңғы жағымсыз жаңалықты сипаттайтын төмендегі мәтінді алайық:

«*2022 жылғы Қазақстандағы наразылық шаралары сұйытылған газдың кенеттен қымбаттауына байланысты басталды. 2 қаңтарда бастау алған бұл оқиғалар тәуелсіз Қазақстанның 30-жылдық тарихындағы ең қарқынды әрі қатал қақтығысқа айналды. Бастапқыда бейбіт басталған наразылық шаралары артынан қарулы қақтығыстар мен тонаушылыққа ұласты. 5 қаңтарда басталған тәртіпсіздіктер ресми деңгейде мемлекеттік төңкеріс деп аталды*»

Кесте 3. Мәтіндегі жағымсыз сөздер тізімі

| № | Жағымсыз сөздер | Реңктік көрсеткіштері | Ағылшынша аудармасы |
|---|---|---|---|
| 1 | наразылық шаралары | 4 балл | protests |
| 2 | қатал | 1 балл | violent |
| 3 | қақтығыс | 4 балл | conflict |
| 4 | қарулы | 4 балл | armed |
| 5 | тонаушылық | 4 балл | looting |
| 6 | тәртіпсіздік | 4 балл | riots |
| 7 | мемлекеттік төңкеріс | 5 балл | coup |

$$N = \frac{4+1+4+4+4+4+5}{5} = \frac{26}{5} = 5{,}2$$

Теріс шектің шыққан мәні 1-ден артық болуына байланысты, біз оны $N \approx 1$ деп аламыз.

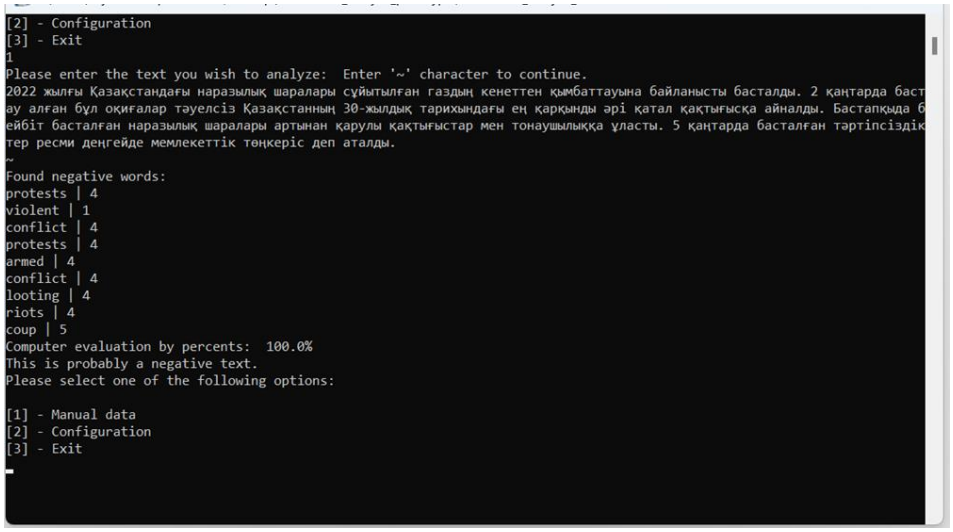
$$NC = \frac{1+1+1+1+1+1+1}{3} = \frac{7}{3} \approx 2{,}33$$

Теріс санау шегінің де шыққан мәні 1-ден артық болуына байланысты, біз оны да $NC \approx 1$ деп аламыз.

$$NS = (0{,}7 \times N + 0{,}3 \times NC) \times 100\% = (0{,}7 \times 1 + 0{,}3 \times 1) \times 100\% = 100\%$$

Есептеу нәтижелерінен біз берілген мәтіннің толық 100% негативті екенін көре аламыз, яғни бұл мәтінде жағымсыз хабар беріліп тұрғаны рас екеніне көз жеткіздік.

Осындай алгоритммен жұмыс істейтін құралды құру үшін Python жоғары программалау тілі қолданылып, деректер базасы толтырылды. Соны нәтижесінде әлеуметтік желідегі постты талдауға мүмкіндігі бар, мәтіндердегі жағымсыз көңіл-күйді анықтайтайтын «Negative Mood» программасы жасалды. Программаның жұмысын төмендегі суреттен көруге болады.




Сурет 1. Программа жұмысының нәтижесі

## 4. Қорытынды

Жасалып жатқан ғылыми жұмыс табиғи тілді өңдеудің кең етек алып келе жатқан сентимент талдау саласы бойынша зерттеу жұмыстарын жүргізе отыра, қазақ тілді мәтіндердің жағымсыз көңіл-күйін анықтауды көздеген болатын. Бұл есепті шешу үшін жағымсыз сөздер қоршаған ортаға әсер етуіне қарай категорияларға бөлініп, реңк көрсеткіштері тағайындалған деректер базасы құрылып, жағымсыз көңіл-күйді анықтау алгоритмі мен машиналық аударма негізінде жұмыс істейтін сентимент анықтауға арналған «Negative Mood» мультитілді (Google аудармашы қолдайтын барлық тілдер) программасы жасалды. Келешекте программа жұмысын жетілдіре отыра тек қазақ тілді мәтіндерді ғана емес әлеуметтік желідегі посттар мен пікірлерді де талдау көзделіп отыр.

**Әдебиеттер тізімі**

1. Bing Liu. Sentiment Analysis and Subjectivity // Handbook of Natural Language Processing (англ.) / под ред. N. Indurkhya и F. J. Damerau. — 2010.

2. Bo Pang, Lillian Lee. Opinion Mining and Sentiment Analysis (англ.) // Foundations and Trends in Information Retrieval : журнал. — 2008. — No. 2. — P. 1-135.

3. National Research Tomsk State University & E-Learning Development Fund// Онлайн курс Coursera: Введение в искусственный интеллект. URL: https://www.coursera.org/learn/vvedenie-v-iskusstvennyi intellekt/home/welcome

4. Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews (англ.) // Proceedings of the Association for Computational Linguistics. — 2002. — P. 417–424.

5. RCO Fact Extractor SDK [Электронный ресурс]: RCO. – Режим доступа: http://www.rco.ru/product.asp?ob_no=5047

6. Sentiment Analysis: A Definitive Guide — URL: https://monkeylearn.com/sentiment-analysis/

7. SentiStrength [Электронный ресурс]: SentiStrength – sentiment strength detection in short texts. – Режим доступа: http://sentistrength.wlv.ac.uk/#About 28.11.2012

8. Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A. Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. 2010.

9. Washington, Erin. Human Sentiment Analysis (англ.). Growing Social Media (14-11-2013).

10. Yergesh, B., Bekmanova, G., Sharipbay, A. Sentiment analysis of Kazakh text and their polarity // Web Intelligence, 2019, 17(1), p. 9–15

11. Yergesh, B., Bekmanova, G., Sharipbay, A. Sentiment analysis on the hotel reviews in the Kazakh language // 2nd International Conference on Computer Science and Engineering, UBMK 2017, 2017, p. 790–794, 8093531

12. Анализ текста (Microsoft Azure) — URL: https://azure.microsoft.com/ru-ru/services/cognitive-services/text-analytics/#features

13. Анализ тональности в социальных сетях — всё что вам нужно об этом знать. URL: https://youscan.io/ru/blog/social-media-sentiment-analysis-all-the-ins-and-outs/

14. Анализ тональности текста. — URL: https://ru.wikipedia.org/wiki/

15. Анна Пазельская и Алексей Соловьев, Метод определения эмоций в текстах на русском языке. Компьютерная лингвистика и интеллектуальные технологии. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Сб. научных статей / Вып. 11 (18).- М.: Изд-во РГГУ, 2011.– С.510-523.

16. Базенков Н. и др. Обзор информационных систем анализа социальных сетей //Управление большими системами: сборник трудов. – 2013. – №. 41

17. Душкин Р.Искусственный интеллект. –2019

18. Меньшиков, И. Л. Обзор систем анализа тональности текста на русском языке / И. Л. Меньшиков, А. Г. Кудрявцев. — Текст : непосредственный // Молодой ученый. — 2012. — № 12 (47). — С. 140-143. — URL: https://moluch.ru/archive/47/5951/

19. Ергеш Б. Ж. Қазақ тіліндегі арнайы мәтіндерді семантикалық талдау моделдері мен әдістері [Текст]: (PhD) философия докторы ғылыми дәрежесін алу үшін дайындалған диссертация / Ергеш Бану Жантуғанқызы.- 2020. 106 б.

20. Сентимент анализ текста. Блог компании PalitrumLab. — URL: https://habr.com/ru/company/palitrumlab/blog/262595/

*УДК 004.82, 004.93*
***[1]Кудубаева С.А., [2]Жусупова Б.Т.***
*[1]Евразийский национальный университет имени Л.Н. Гумилева*
*Нур-Султан, Казахстан*
*[2]Костанайский региональный университет имени А. Байтурсынова*
*Костанай, Казахстан*
*[1]saule.kudubayeva@gmail.com, [2]botashazhus@gmail.com*

# О ВОЗМОЖНОСТИ УЧЕТА СЕМАНТИЧЕСКОЙ СОСТАВЛЯЮЩЕЙ В СИСТЕМЕ СУРДОПЕРЕВОДА С КАЗАХСКОГО ЯЗЫКА НА КАЗАХСКИЙ ЯЗЫК ЖЕСТОВ

**Аннотация:** В данной статье проводится обзор существующих систем сурдоперевода, выявлены преимущества и недостатки существующих на сегодняшний день систем перевода на жестовые языки. Рассматривается вопрос разработки семантического словаря казахского языка для системы компьютерного перевода с казахского языка на казахский язык жестов, в которой будет учитываться семантика казахского языка и казахского жестового языка (КЖЯ). В качестве основы технологии компьютерного перевода казахского языка на казахский жестовый язык служит семантический словарь казахского языка. В дальнейшем он позволит проводить семантический анализ исходного текста. Авторами статьи проведен анализ и подбор имеющихся словарей казахского языка, используемых при разработке базы данных семантического словаря. Словари казахского языка дают возможность для осуществления компьютерного сурдоперевода на КЖЯ. Семантический словарь казахского языка содержит в себе несколько словарей: грамматический словарь, словарь фразеологизмов, словарь предлогов, словарь синонимов, словарь многозначных слов и омонимов для определения лексических значений слов казахского языка и другие. Так как семантический словарь для компьютерного сурдоперевода представляет собой базу данных взаимосвязанных таблиц. В статье также представлена возможность использования нотации Л.С.Димскис для разработки словаря структуры жестов казахского жестового языка. Раскрыта перспектива его включения в базу данных семантического словаря. А также раскрыта необходимость словаря жестов при разработке системы автоматизированного сурдоперевода в целом с учетом его эффективности и возможности полноценного практического использования.

**Ключевые слова:** *казахский язык, жест, жестовый язык, семантический словарь, компьютерный сурдоперевод.*

*ӘОК 004.82, 004.93*
***[1]Кудубаева С.А., [2]Жусупова Б.Т.***
*[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[2]А.Байтурсынов атындағы Қостанай өңірлік университеті*
*Қостанай, Қазақстан*
*[1]saule.kudubayeva@gmail.com, [2]botashazhus@gmail.com*

# ҚАЗАҚ ТІЛІНЕН ҚАЗАҚ ЫМДАУ ТІЛІНЕ СУРДОАУДАРМА ЖҮЙЕСІНДЕГІ СЕМАНТИКАЛЫҚ ҚҰРАМДАС БӨЛІКТІ ЕСКЕРУ МҮМКІНДІГІ ТУРАЛЫ

**Аңдатпа:** Бұл мақалада әлемдегі сурдоаударма жүйелеріне шолу жасалды, бүгінгі күні қолданыстағы ымдау тілдеріне аудару жүйелерінің артықшылықтары мен кемшіліктері анықталды. Қазақ тілі мен қазақ ымдау тілінің семантикасы ескерілетін қазақ тілінен қазақ ымдау тіліне компьютерлік аударма жүйесі үшін қазақ тілінің семантикалық сөздігін әзірлеу мәселесі қарастырылуда. Мақалада сондай-ақ қазақ ымдау тілінің ымдау құрылымының сөздігін әзірлеу үшін Л.С. Димскистің нотациясын пайдалану мүмкіндігі ұсынылған. Оны семантикалық сөздіктің деректер базасына енгізу перспективасы ашылды. Сондай-ақ, автоматтандырылған сурдоаударма жүйесін әзірлеу кезінде оның тиімділігі мен толыққанды практикалық пайдалану мүмкіндігін ескере отырып, ымдау сөздігінің қажеттілігі ашылды.

**Түйін сөздер:** қазақ тілі, ым-ишара, ымдау тілі, семантикалық сөздік, компьютерлік сурдоаударма.

*UDC 004.82, 004.93*
***[1]Kudubayeva S., [2]Zhussupova B.***
*[1]L.N. Gumilyov Eurasian national University*
*Nur-Sultan, Kazakhstan,*
*[2]A. Baitursynov Kostanay regional university*
*Kostanay, Kazakhstan*
*[1]saule.kudubayeva@gmail.com, [2]botashazhus@gmail.com*

# ABOUT THE POSSIBILITY OF TAKING INTO ACCOUNT THE SEMANTIC COMPONENT IN THE SURDO TRANSLATION SYSTEM FROM KAZAKH LANGUAGE INTO KAZAKH SIGN LANGUAGE

**Abstract:** This article reviews of the existing sign language translation

systems, reveals the advantages and disadvantages of the existing sign language translation systems. This article discusses the development of a semantic dictionary of the Kazakh language for a computer translation system from Kazakh to Kazakh sign language, which will take into account the semantics of the Kazakh language and the Kazakh sign language (KSL). The semantic dictionary of the Kazakh language serves as the basis of computer translation technology from the Kazakh language to the Kazakh sign language. In the future, it will allow semantic analysis of the source text. The authors of the article analyzed and selected the available dictionaries of the Kazakh language used in the development of the semantic dictionary database. Dictionaries of the Kazakh language provide an opportunity for computer-based sign language translation of the KSL. The semantic dictionary of the Kazakh language contains several dictionaries: a grammatical dictionary, a dictionary of phraseological units, a dictionary of prepositions, a dictionary of synonyms, a dictionary of polysemous words and homonyms to determine the lexical meanings of words of the Kazakh language and others. Since the semantic dictionary for computer sign language translation is a database of interconnected tables. The article also presents the possibility of using L. S. Dimskis notation to develop a dictionary of the structure of gestures of the Kazakh sign language. The prospect of its inclusion in the database of semantic dictionary is revealed. And also revealed the need for a dictionary of gestures in the development of automated sign language translation system as a whole, taking into account its effectiveness and the possibility of full practical use.

***Keywords:*** *kazakh language, gesture, sign language, semantic dictionary, computer sign language translation.*

## 1. Введение

Жестовый язык – это основное средство межличностной коммуникации большинства глухих и части слабослышащих людей. По данным Всемирной организации здравоохранения (ВОЗ) официально в мире порядка 360 миллионов человек страдают глухотой или имеют проблемы со слухом, из которых 328 миллионов взрослых людей и 32 миллиона детей. В Казахстане насчитывается порядка 200 тысяч людей с инвалидностью по слуху.

По своим коммуникативным функциям не уступая звучащим языкам, жестовый язык является полноценным самостоятельным языком. Аналогично звучащим языкам, имеющим слова с различными значениями, жестовый язык содержит в себе однозначные, многозначные, разнозначные жесты. Полностью понять правильное значение жеста можно из контекста.

Одной из важных задач любого современного государства является создание безбарьерной среды и необходимых условий для обучения и коммуникации людей с ограничениями по слуху.

**2. Обзор существующих систем сурдоперевода**

В мире были созданы различные системы сурдоперевода: Zardoz, TEAM, ViSiCAST, система машинного сурдоперевода на базе Microsoft Kinect, система SISI и др. Анализируя особенности каждой системы сурдоперевода, определены их преимущества и недостатки, которые необходимо учесть при разработке системы сурдоперевода с казахского языка на КЖЯ.

Система Zardoz была предложена в качестве системы перевода с английского языка на язык жестов, в которой язык-посредник (интерлингва) в качестве элемента перевода. Текущие исследования сосредоточены на разработке всеобъемлющей грамматики, морфологии и лексики для ирландского языка жестов [1].

Система TEAM (TranslationfromEnglishtoASLbyMachine) – это система машинного перевода с английского языка на американский жестовый язык. Перевод в системе TEAM состоит из двух этапов: первый - перевод введенного предложения с английского языка на промежуточное представление с учетом синтаксической, грамматической и морфологической информации, второй - отображение промежуточного представления в виде движения с небольшим набором параметров, которые в дальнейшем преобразуются в большее число параметров, которые управляют моделью человека, воспроизводящей жесты. Гибкость системы позволяет адаптировать ее к другим жестовым языкам [2].

Система ViSiCAST (VirtualSigning: Capture, Animation, StorageandTransmission) - это система машинного перевода с английского языка на американский жестовый язык. Основная цель проекта ViSiCAST – это улучшение качества доступа к различной информации, развлечениям, образованию и общественным услугам для глухих граждан Европы. Проект ViSiCAST является упрощенной системой, которая фиксирует движения и жесты человека-сурдопереводчика, а затем эти координаты рук переводчика передаются для последующего анализа для получения реалистичного аватара [3].

Система машинного сурдоперевода, разработанная на базе технологии Kinect от Microsoft, способна считывать движения рук и всего тела. Список функций системы включает в себя помимо распознавания движений также сурдоперевод, как часть нового исследовательского проекта, призванного помочь людям с отсутствием

слуха. Созданная технология не только переводит язык жестов в слова, проговариваемые компьютером, но и осуществляет обратный процесс: пользователь без недостатков слуха говорит или впечатывает слова в переводчик Kinect, а система затем воспроизводит слова на языке жестов с помощью виртуального аватара на экране.

Система «Say It Sign It», разработанная в исследовательском центре IBM Hursley в Великобритании, позволяет переводить устную речь в язык жестов. Система «Say It Sign It» (SiSi) объединяет несколько компьютерных технологий. Сначала специальный модуль распознавания речи преобразуют произнесённые одним из пользователей в микрофон слова в текст. Затем специальная программа «прогоняет» текст через программу-переводчика, которая анализирует сказанное и переводит текст в английский язык жестов, в то время как виртуальный аватар изображает переведённый фрагмент. Жестовые аватары и технология для анимации языка жестов из специальной системы обозначений жестов были разработаны Университетом Восточной Англии, а база данных жестов была разработана RNID (Royal National Institute for Deaf People).

В будущем SiSi сможет стать одним из сайтов в Интернете, при этом перевод слов в язык жестов осуществлялся бы в центральном сервере, а анимированный результат выводился бы на экран компьютера зарегистрировавшегося пользователя. Кроме того, система сможет работать в виде отдельной программы, которую адресат и адресант смогут установить на их собственные компьютеры. Также рассматривается вариант с функционированием системы через телевизионный преобразователь [4].

Для разработки системы компьютерного сурдоперевода сотрудниками Новосибирского государственного технического университета под руководством профессора Гриф М.Г. был предложен новый способ построения семантического блока системы. В ходе построения данного блока семантического анализа для установления соответствия «слово-жест» были определены лексические значения слов, среди множества альтернатив на основе алгоритма семантического анализа за каждым словом закреплялось единственное лексическое значение. Для простых предложений были разработаны и реализованы алгоритмы семантического анализа, предложен способ перевода русского текста на русский жестовый язык на основе сопоставления синтаксических конструкций, для определения которых была разработана библиотека [5].

Ученые из Института проблем управления им. В.А. Трапезникова РАН (ИПУ РАН) разрабатывают систему, которая с помощью

искусственного интеллекта сможет в режиме реального времени через видеокамеру переводить жестовый язык в слова, фразы и буквы. После того как группа исследователей под руководством заведующего лабораторией «Автоматизированных систем массового обслуживания и обработки сигналов» ИПУ РАН Маиса Фархадова соберет необходимые статистические данные, они будут предъявлены нейросети, которую с применением методов машинного обучения «натренируют» распознавать, какой звук или буква соответствует определенному жесту. В данной системе для распознавания жестов применяется искусственный нейронный кортекс (группа нейронов, ответственная за принятие решений), способный распознавать статические жесты. Дактильную азбуку он уже распознает и в будущем эта разработка будет доведена до автоматического сурдопереводчика.

Разработанный интернет-портал «Сурдосервер», призванный облегчить людям с проблемами слуха и их родственникам изучение жестового языка, содержит сотни обучающих видеороликов, речевой тренажер, который помогает глухим пользователям узнать, насколько правильно они произносят тот или иной звук, глоссарий с дактильными азбуками разных стран, словарь «диалектов жестового языка». Также ученые из ИПУ РАН разрабатывают мобильное приложение «Сурдосервис» и сурдооблако, в котором люди с проблемами слуха смогут мгновенно обмениваться информацией. Предполагается, что, после того как разработчики обучат нейросеть с максимальной точностью переводить жестовый язык в слова и буквы, все сурдосервисы, созданные в ИПУ РАН, будут интегрированы в единый пакет программ, который можно будет распространять среди слабослышащих людей [6].

### 3. Анализ преимуществ и недостатков систем сурдоперевода

Анализируя существующие на сегодняшний день системы сурдоперевода, можно отметить, что большинство зарубежных систем, кроме системы SISI, не может обрабатывать входную информацию, поступающую в виде голоса. Для систем перевода, способных обеспечить устный перевод, этот недостаток является существенным.

Использование пространственной информации вокруг говорящего, является спецификой жестового языка, которая учитывается только в системе Team. В системе Zardoz делаются попытки учета семантической составляющей жестового языка, помимо морфологической и синтаксической информации, необходимой для более качественного перевода. Недостатком технология перевода в системе ViSiCAST является привлечение человека в процесс перевода.

В системе ViSiCAST достигнута максимальная реалистичность аватара. Но по мнению носителей жестового языка виртуальные сурдо-переводчики (аватары), несмотря на свою уникальность, не являются совершенными, так как не могут передать выразительность и точность жестовой речи [6].

Основным недостатком рассмотренных выше систем является отсутствие учета семантической составляющей как звучащего, так и жестового языка. Системы, в которых в процессе перевода предусмотрен учет особенностей семантики исходного языка и язык перевода, имеют большое преимущество и обладают высоким качеством перевода. Только в системе семантического анализа, разработанной в Новосибирском ГТУ, ведется учет морфологической и синтаксической, а также семантической составляющих.

В ходе работы над разработкой систем сурдоперевода также столкиваются с такой проблемой, как недостаточность объёма словарей жестовых языков, не всегда предоставляющих корректную и современную информацию.

## 4. Способ реализации системы сурдоперевода для КЖЯ

Основой технологии компьютерного сурдоперевода с казахского языка на КЖЯ служит семантический словарь казахского языка, учитывающий особенностей семантики исходного языка. В дальнейшем он позволит проводить семантический анализ исходного текста.

Семантический словарь казахского языка содержит в себе несколько словарей: грамматический словарь, словарь фразеологизмов, словарь предлогов, словарь синонимов, словарь многозначных слов и омонимов для определения лексических значений слов казахского языка и другие. Так как семантический словарь для компьютерного сурдоперевода представляет собой базу данных взаимосвязанных таблиц «Словарные статьи», «Фразеологизмы», «Омомнимы», «Синонимы», «Предлоги» и др., то в ней необходимо также наличие таблицы «Жесты» и таблицы-связки «Жест - слово» для дальнейшего определения соответствия «слово-жест».

Морфологический модуль системы сурдоперевода должен содержать следующие морфологические словари: грамматический словарь, словарь имен собственных, словарь географических мест. Для компьютерной обработки казахского языка важно, прежде всего проведение морфологического анализа на уровне, требуемом системой обработки текста.

Основой создания морфологического словаря казахского языка служат следующие труды: «Қазақ тілінің функционалды грамматикасы»

(Алматы, 2012), «Қазіргі қазақ тілінің морфологиясы» (Алматы, 2007) автора Оралбай Н. и «Қазіргі қазақ тілінің морфемалар жүйесі» (Алматы, 2001) авторов Оралбаева Н., Қалыбаева А.

Для формирования словаря имен собственных и словаря географических мест используются словари, разработанные Институтом языкознания имени А.Байтурсынова:

- Жанұзақов Т. «Қазақ тіліндегі жалқы есімдер» (Алматы, 1965);
- Жанұзақов Т., Есбаева К. «Қазақ есімдері» (Алма-Ата, 1988);
- Т.Жанузаков, К.К.Рысбергенова, Н.Б.Онгарбаева «Қазақстан Республикасының топонимдері» (Алматы, 2001);
- «Қазақ жер-су аттары» Энциклопедиялық анықтағыш (Алматы, 2009);
- «Қазақ кісі аттары» Энциклопедиялық анықтағыш (Алматы, 2009).

На их основе создаются соответствующие таблицы в базе данных семантического словаря.

Результат морфологического анализа является входной информацией для синтаксического и первичного семантического анализа, в ходе которых определяются семантические отношения.

Словарь синонимов, содержащий 8246 слов и доступный на сайте lugat.kz, и словарь «Синонимдер сөздігі» автора Бизақова С. (Алматы, 2007) являются основой таблицы синонимов казахского языка, являющего частью разрабатываемого семантического словаря [7].

Для разрешения проблемы лексической многозначности необходимо обработать омонимы и фразеологизмы в предложении. Омонимия – это совпадение по звучанию и написанию различных слов. Фразеологизмы отличаются от обычных сочетаний слов тем, что общее значение фразеологического оборота не равно сумме отдельных значений слов. В семантическом словаре казахского языка словари многозначных слов или омонимов и фразеологизмов основываются на следующих трудах:

- 15-томный толковый словарь казахского литературного языка, содержащий 92 300 слов и 57 856 словосочетаний [8].
- «Қазақ тілінің омонимдер сөздігі» М.Белбаева (1988);
- «Мағыналас фразеологизмдер сөздігі» Г. Смағұлова (Алматы, 2010);
- «Қазақ тілінің фразеологиялық сөздігі» І.Кеңесбаев (Алматы, 2007).

При обработке омонимов и многозначных слов важно, чтобы система верно выбрала нужное лексическое значение слова.

Разработка базы данных семантического словаря, включающего в себя взаимосвязанные таблицы перечисленных словарей казахского

языка, является основной задачей при создании семантического модуля системы компьютероного сурдоперевода. Таким образом, словари казахского языка являются основой, дают возможность для осуществления компьютерного сурдоперевода КЖЯ.

Учитывая то, что при разработке системы компьютерного сурдоперевода жестового языка необходимо переводить текст на язык жестов, распознавать жесты, для этого требуется наличие в базе данных семантического словаря таблицы по жестам КЖЯ. Словарь жестового языка - основа любого компьютерного сурдоперевода.

Казахстанские ученые совместно с учеными ИПУ РАН создали первый электронный словарь казахского жестового языка на сайте www.surdo.kz, а также электронный учебник по казахскому буквенному жестовому языку.

Так как в жестовых языках присутствует большой процент заимствования жестов, которых насчитывается порядка 1500 жестов, они были включены в язык gestuno, используемый в рамках Всемирной федерации глухих (ВФГ). Из существующих на сегодняшний день разработанных словарей жестов только некоторые являются удобными в практическом использовании для изучения и распространения жестовых языков.

Однако следует отметить, что не существует такого словаря КЖЯ, который позволял бы по форме жеста находить его значение, хотя в мире создание подобных словарей жестовых языков – достаточно распространенная практика. Учитывая возможность применимости системы Л.С. Димскис к КЖЯ, нами было проверено достаточно ли компонентов (знаков) жестовой нотации Л.С.Димскис для точного описания жестов казахского ЖЯ, необходимы ли дополнительные знаки для записи жестов [9].

Разработанный словарь структуры жестов КЖЯ может быть использован для осуществления поиска слов в словаре, представления их в виде нотации и отображения с помощью аватара. Также словарь жестов позволит осуществлять распознавание КЖЯ. При этом подходе, представляя жесты в виде компьютерных нотаций, проводя поиск слов в словаре, лингвистическую обработку, результат распознавания представляется в виде текста или голосового сообщения.

## 5. Заключение

Проводя обзор исследований в области жестовых языков, современных подходов к разработке систем сурдоперевода, анализируя практическую применимость существующих словарей жестов, можно говорить о том, что научный мир все же проявляет достаточно большой

интерес к исследованию структуры жестовых языков, к распознаванию, анализу семантики жестов, построению систем перевода жестов.

Проводимые исследования по анализу и подбору словарей казахского языка, для создания семантического словаря системы компьютерного сурдоперевода казахского жестового языка обоснованы необходимостью разработать такую систему сурдоперевода, которая выполняет качественный перевод и способна обеспечить комфортную и доступную среду для безбарьерного общения людей с нарушением слуха со слышащими, в том числе при получении образования, медицинских и государственных услуг и прочее.

**Список литературы**

1. Veale, T., Conway, A. Cross modal comprehension in ZARDOZ: an English to sign-language translation system / T. Veale, A. Conway // Proceedings of the Seventh International Workshop on Natural Language Generation INLG'94, Kennebunkport, Maine. – 1994. – P. 249–252.

2. Zhao, L., Kipper, K., Schuler. W. A machine translation system from English to American sign language / L. Zhao, K. Kipper, W. Schuler // Lecture Notes in Computer Science. – 2000. – Vol. 1934. – P. 54–67.

3. Wakefield, M. VisiCAST Milestone: final report no. IST-1999-10500 / M. Wakefield // Information Societies Technology. – 10 December 2002. – 97 p.

4. http://www.sys-consulting.co.uk/web/ProjectSISI.html

5. Гриф М.Г., Мануева Ю.С. Разработка и тестирование алгоритма семантического анализа речи (текста) для перевода на русский жестовый язык / М. Г. Гриф, Ю. С. Мануева // Вестник. Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. – 2017. – Т. 15 – № 2. – С. 70–80. – ISSN 1818-7935

6. Недюк М. Голос жестов: искусственный интеллект применят для сурдоперевода. Программа поможет людям с проблемами слуха и их родственникам обучиться жестовому языку и чтению по губам / М.Недюк // Известия. – 22 ноября 2018 года.

7. Бизақов С. Синонимдер сөздігі. Алматы: «Арыс» баспасы, 2007. – 640 б.

8. Қазақ әдеби тілінің сөздігі. Он бес томдық. / Құраст. Т.Жанұзақ, С.Омарбеков, Ә.Жүнісбек және т.б. – Алматы, 2011.

9. Димскис, Л. С. Мы изучаем жестовый язык. Москва: Академия. – 2002. – 128 с.

*УДК 81.322*
***Азат Абдысадыр уулу***
*Канцелярия*
*Администрации Президента КР*
*Бишкек, Кыргызстан*
*aralkg@mail.ru*

# МЕТОД БИНАРНЫХ СВЯЗЕЙ ДЛЯ ВИЗУАЛИЗАЦИИ СМЫСЛА ПРЕДЛОЖЕНИЯ

**Аннотация.** В статье ставится задача визуализации для смысла предложения, исходя из линейного вектора расположения лексем, т.е. метода визуализации смысла предложения. Задача решается на базе бинарных связей и их процессов, парадигмы бинарных связей, а также визуализации структурированной парадигмы. Линейный вектор расположения лексем предполагает наличие бинарных связей, которые создаются благодаря морфологическим характеристикам лексем, где бинарные связи и процессы при бинарных связях определяются в рамках конкретного языка. В рамках определенного предложения формируется парадигма бинарных связей. На базе парадигмы бинарной связи создается визуализация структурированной парадигмы. Данный метод позволяет решить вопросы визуализации структурированной парадигмы предложения и визуальной типологии структуры смыла предложения.

**Ключевые слова:** визуализация, линейный вектор лексем, бинарные связи, парадигма бинарных связей, структурированная парадигма.

*UDC 81.322*
***Azat Abdysadyr uulu***
*Office of the Administration of the*
*President of the Kyrgyz Republic*
*Bishkek, Kyrgyzstan*
*aralkg@mail.ru*

# THE METHOD OF BINARY LINKS FOR VISUALIZATION OF THE MEANING OF A SENTENCE

**Abstract.** The article poses the task of visualizing the meaning of a sentence, based on the linear vector of the location of lexemes, i.e. method of visualizing the meaning of a sentence. The problem is solved on the basis of

binary links and their processes, the binary link paradigm, as well as the visualization of the structured paradigm. The linear vector of the location of lexemes implies the presence of binary links that are created due to the morphological characteristics of lexemes, binary relations and processes at binary relations are defined within the framework of a particular language. Within the framework of a certain proposal, a paradigm of binary relations is formed. On the basis of the binary connection paradigm, a visualization of the structured paradigm is created. This method allows solving the issues of visualization of the structured sentence paradigm and visual typology of the structure of the meaning of the sentence.

**Keywords:** visualization, linear vector of lexemes, binary links, binary links paradigm, structured paradigm.

## I. ВВЕДЕНИЕ

Предложение (K) является синтагматической единицей языка [Энциклопедия, 1997, с. 470]. Данная особенность предполагает наличие линейного вектора расположения лексем ($\vec{v}$).

$$K \ni (\vec{v}) \quad (1)$$

При этом смысл предложения не имеет синтагматическую характеристику. Грамматическая основа предложения и второстепенные члены предложения [Азыркы кыргыз тили, 2015, с. 379] создают структурированную парадигму (Y).

$$K_s \ni (Y) \quad (2)$$

Визуализация смысла конкретного предложения предполагает визуализацию (Y).

Таким образом возникает задача визуализации (Y) для смысла предложения, исходя из линейного вектора расположения лексем.

Если дано

$$K \ni (L_1 \dots L_n) \quad (3)$$

где

$L_n$ – лексемы,

тогда необходимо

$$П\ (Y) \quad (4)$$

где

П – оператор «визуализация».

Тем самым цель статьи – создание метода визуализации смысла предложения.

## II. РЕШЕНИЕ ЗАДАЧИ

### 1. Бинарные связи и процессы при бинарных связях

Линейный вектор расположения лексем предполагает наличие бинарных связей (В), которые создаются благодаря морфологическим характеристикам лексем. Следовательно, лексема, имеющая морфологическую характеристику, является элементом бинарной связи.

$$L_n \ni (I) \cup R_i \quad (5)$$

где

I – морфологическая характеристика;

$R_i$ – элемент бинарных связей;

∪ – оператор «является».

Далее

$$K \ni (R_{i1} \Rightarrow R_{i2} \Rightarrow R_{i3} \ldots \Rightarrow R_{in}) \quad (6)$$

где

⇒ – оператор бинарной связи.

Бинарные связи определяются в рамках конкретного языка. К примеру, на кыргызском языке имеются следующие бинарные связи:

1) бинарые связи для глагола:

$$R_a \Rightarrow R_v;\ R_t \Rightarrow R_v;\ R_j \Rightarrow R_v;\ R_r \Rightarrow R_v;\ R_p \Rightarrow R_v; \quad (7)$$

2) бинарные связи для номинатива:

$$R_q \Rightarrow R_a;\ R_s \Rightarrow R_a;\ R_p \Rightarrow R_a;\ R_w \Rightarrow R_a;\ R_a \Rightarrow R_a; \quad (8)$$

3) бинарные связи для конверба:

$$R_t \Rightarrow R_u;\ R_q \Rightarrow R_u;\ R_s \Rightarrow R_u;\ R_p \Rightarrow R_u;\ R_r \Rightarrow R_u;\ R_j \Rightarrow R_u; \quad (9)$$
$$R_a \Rightarrow R_u;\ R_h \Rightarrow R_u;$$

4) бинарные связи для падежных форм (без именительного падежа):

$$R_q \Rightarrow R_j;\ R_t \Rightarrow R_j;\ R_r \Rightarrow R_j;\ R_p \Rightarrow R_j;\ R_j \Rightarrow R_j;\ R_w \Rightarrow R_j; \quad (10)$$

5) бинарные связи для категории принадлежности:

$$R_b \Rightarrow R_r;\ R_s \Rightarrow R_r;\ R_a \Rightarrow R_r; \quad (11)$$

6) бинарные связи для наречия:

$$R_j \Rightarrow R_t; R_q \Rightarrow R_t; \quad (12)$$

7) бинарные связи для личных форм:

$$R_b \Rightarrow R_m;\ R_a \Rightarrow R_m. \quad (13)$$

где

$R_a$ – элемент бинарных связей с морфологической характеристикой номинатива;

$R_t$ – элемент бинарных связей с морфологической характеристикой наречия;

$R_j$ – элемент бинарных связей с морфологической характеристикой падежных форм;

$R_r$ – элемент бинарных связей с морфологической характеристикой категории принадлежности;

$R_v$ – элемент бинарных связей с морфологической характеристикой глагола;

$R_p$ – элемент бинарных связей с морфологической характеристикой имени прилагательного;

$R_q$ – элемент бинарных связей с морфологической характеристикой причастия;

$R_s$ – элемент бинарных связей с морфологической характеристикой имени числительного;

$R_w$ – элемент бинарных связей с морфологической характеристикой номинатива местоимения;

$R_u$ – элемент бинарных связей с морфологической характеристикой конверба;

$R_h$ – элемент бинарных связей с морфологической характеристикой деепричастия;

$R_b$ – элемент бинарных связей с морфологической характеристикой родительного падежа;

$R_c$ – элемент бинарных связей с морфологической характеристикой дательного падежа;

$R_d$ – элемент бинарных связей с морфологической характеристикой винительного падежа;

$R_e$ – элемент бинарных связей с морфологической характеристикой местного падежа;

$R_f$ – элемент бинарных связей с морфологической характеристикой исходного падежа;

$R_m$ – элемент бинарных связей с морфологической характеристикой личной формы;

$R_g$ – элемент бинарных связей с морфологической характеристикой служебных слов.

Также при бинарных связях происходят определенные процессы ($B_z$), присущие к конкретному языку.

$$B \ni (B_z) \quad (14)$$

Процессы при бинарных связях определяются в рамках конкретного языка. К примеру, на кыргызском языке имеются следующие процессы:

1) процесс конвергенции:

$$R_q \Rightarrow R_h = R_u;$$
$$R_h \Rightarrow R_q = R_u;$$
$$R_u \Rightarrow R_v = R_v;$$
$$R_a \Rightarrow R_a = R_a;$$
$$R_a \Rightarrow R_q = R_p;$$
$$R_p \Rightarrow R_v = R_v;$$
$$R_i \Rightarrow R_g = R_i;$$
$$R_{h-q} = R_v;$$
$$R_{h-h} = R_h;$$
$$R_{post} = R_v$$
$$R_x = R_0;$$
$$R_{i1} \nRightarrow R_{i2} = R_{i1} \Rightarrow R_0$$

где

$R_{post}$ – элемент бинарных связей в последней позиции линейном векторе расположения лексем;

$R_0$ – элемент бинарных связей с нулевой характеристикой;

$R_x$ – пунктуационные знаки в качестве элемента бинарных связей.

2) процесс комбинации:

$$R_b \nRightarrow R_i \Rightarrow R_e = R_b \Rightarrow R_e;$$
$$R_i \Rightarrow R_0 = R_i \Rightarrow R_v;$$

3) процесс дифференциации:

$$R_a \Rightarrow R_g \Rightarrow R_a = R_a;$$
$$R_a \Rightarrow R_a \Rightarrow R_a = R_a.$$

## 2. Парадигма бинарных связей

В рамках определенного предложения формируется парадигма бинарных связей ($B_{st}$).

$B_{st}$ формируется согласно формальному определению наличия или отсутствия бинарных связей в конкретном языке.

$$(R_{i1} \Rightarrow R_{i2}) \vDash B \quad (15)$$

где

$\vDash$ – оператор «имеется в»

Тогда при

$$K \ni (R_{i1} \Rightarrow R_{i2} \Rightarrow R_{i3} \ldots \Rightarrow R_{in}) \quad (16)$$

формируется

$$(R_{i1} \Rightarrow R_{i2}) \vDash B = (R^1_{i1} R^1_{i2}); \quad (17)$$
$$(R_{i2} \Rightarrow R_{i3}) \vDash B = (R^2_{i1} R^2_{i2}); \quad (18)$$
$$(R_{i3} \Rightarrow R_{in}) \vDash B = (R^3_{i1} R^3_{i2}) \quad (19)$$

где

$R^{n}_{i1;i2}$ – элемент парадигмы бинарных связей, при этом n – порядковый номер строки парадигмы бинарной связи, i1 – первый элемент строки парадигмы бинарной связи, i2 – второй элемент строки парадигмы бинарной связи.

Таким образом имеется парадигма бинарной связи

$$B_{st}\begin{vmatrix} R^{1}_{i1}R^{1}_{i2} \\ R^{2}_{i1}R^{2}_{i2} \\ R^{3}_{i1}R^{3}_{i2} \\ R^{n}_{i1}R^{n}_{i2} \end{vmatrix} (20)$$

## 3. Визуализация структурированной парадигмы (Y)

На базе парадигмы бинарной связи создается визуализация структурированной парадигмы (Y)

$$T \uparrow [R^{n}_{v} \sim R^{n}_{f} \rightarrow R^{1}_{i1}] (21)$$

где

T – оператор «создаем схему»;

↑ – оператор порядка «внизу сверх»;

~ – оператор порядка

$R^{n}_{f}$ – элемент парадигмы бинарных связей, имеющее позицию после $R^{n}_{v}$ в парадигме бинарных связей.

В случае

$$B_{st} \ni (R^{n}_{0}) (11)$$

$$T \uparrow \begin{bmatrix} R^{n}_{v} \sim R^{n}_{f} \rightarrow R^{n}_{0}; \\ R^{n}_{v} \sim R^{n}_{\phi} \rightarrow R^{1}_{i1} \end{bmatrix} (22)$$

где

$R^{n}_{\phi}$ – элемент парадигмы бинарных связей, имеющее позицию после $R^{n}_{0}$ в парадигме бинарных связей.

В случае

$$B_{st} \ni (R^{n}_{v}) (23)$$

когда v количественно больше двух, т.е. $v > 1$

$$T \uparrow \begin{bmatrix} R^{n}_{v1} \sim R^{n}_{f1} \rightarrow R^{n}_{1v}; \\ R^{n}_{v2} \sim R^{n}_{f2} \rightarrow R^{1}_{2v} \\ R^{n}_{vn} \sim R^{n}_{fn} \rightarrow R^{1}_{nv} \end{bmatrix} (24)$$

где

$R^{n}_{v1}$ – элемент парадигмы бинарных связей с морфологической характеристикой глагола 1;

$R^{n}_{f1}$ – элемент парадигмы бинарных связей, имеющее позицию после $R^{n}_{v1}$ в парадигме бинарных связей;

$R^{n}_{1v}$ – элемент парадигмы бинарных связей, имеющее позицию перед $R^{n}_{v2}$ в парадигме бинарных связей;

$R^{n}_{v2}$ – элемент парадигмы бинарных связей с морфологической характеристикой глагола 2;

$R^{n}_{f2}$ – элемент парадигмы бинарных связей, имеющее позицию после $R^{n}_{v2}$ в парадигме бинарных связей;

$R^{1}_{2v}$ – элемент парадигмы бинарных связей, имеющее позицию перед $R^{n}_{v3}$ в парадигме бинарных связей;

$R^{n}_{vn}$ – элемент парадигмы бинарных связей с морфологической характеристикой глагола n;

$R^{n}_{fn}$ – элемент парадигмы бинарных связей, имеющее позицию после $R^{n}_{vn}$ в парадигме бинарных связей;

$R^{1}_{nv}$ – элемент парадигмы бинарных связей, имеющее позицию перед $R^{n}_{vn+1}$ в парадигме бинарных связей.

Таким образом была осуществлена визуализация структурированной парадигмы (Y).

## 4. Пример визуализации структурированной парадигмы (Y)

В качестве примера предлагается визуализация предложения из повести Ч. Айтматова «Гүлсарат»:

«Андайда Танабай арабадан ыргып түшүп, кийимин ыйыгына арта салып, шыпылдай басып өр талаша жөө жөнөчү» [Айтматов Ч., 1982. 1 том, с. 30].

Исходя из (5)-(20), имеем следующую $B_{st}$

$$B_{st}\begin{vmatrix} R^{1}_{t}R^{1}_{0} \\ R^{2}_{a}R^{2}_{0} \\ R^{3}_{f}R^{3}_{h} \\ R^{4}_{h}R^{4}_{h} \\ R^{5}_{h}R^{5}_{0} \end{vmatrix}\begin{vmatrix} R^{6}_{d}R^{6}_{e} \\ R^{7}_{e}R^{7}_{h} \\ R^{8}_{h}R^{8}_{h} \\ R^{9}_{h}R^{9}_{0} \\ R^{10}_{h}R^{10}_{h} \end{vmatrix}\begin{vmatrix} R^{11}_{h}R^{11}_{0} \\ R^{12}_{a}R^{12}_{h} \\ R^{13}_{h}R^{13}_{0} \\ R^{14}_{p}R^{14}_{v} \end{vmatrix}$$

Далее исходя из (21)-(24), имеем визуализацию структурированной парадигмы (Рис.1).

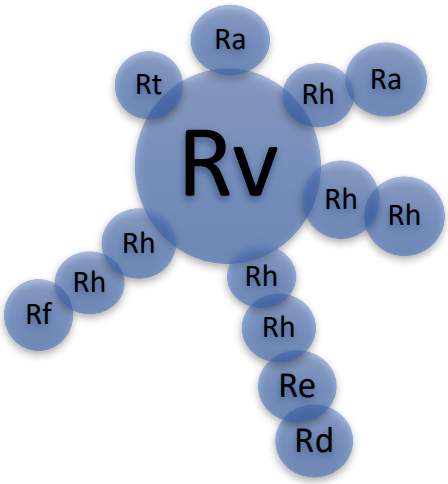



Рисунок 1. Визуализация структурированной парадигмы.
Figure 1. Visualization of the structured paradigm.

## III. ВЫВОДЫ

Таким образом в данной статье разработан метод бинарных связей для визуализации смысла предложения, т.е. решена задача визуализации (Y) для смысла предложения, исходя из линейного вектора расположения лексем.

Данный метод позволяет решить вопросы: 1) визуализации структурированной парадигмы предложения; 2) визуальной типологии структуры смыла предложения.

**Список литературы**

1. Азыркы кыргыз тили: фонетика, лексикология, морфология, синтаксис. – Б.: 2015. – 518. – [1. Azyrky kyrgyz tili: fonetika, leksikologiya, morfologiya, sintaksis. – B.: 2015. – 518.].

2. Айтматов Ч. 3 томдон турган чыгармалар. – Ф.: 1-том., 1982. – 448. – [2. Aytmatov Ch. 3 tomdon turgan chygarmalar. – F.: 1-tom., 1982. – 448.].

3. Энциклопедия “Русский язык”. М.: 1997. – [3. Entsiklopediya “Russkiy yazyk”. M.: 1997.].

*UDC 83.00*
***Khamroeva Shahlo Mirdjanovna***
*Tashkent State university of Uzbek language and literature named Alisher Navai*
*Tashkent, Uzbekistan*
*hamroyeva81@mail.ru*

# FINITE STATE MACHINE MODEL OF NOUNS FOR UZBEK LANGUAGE MORPHOLOGICAL ANALYZER

**Abstract.** This article discusses the Finite State Transduser (FST) models of creating a morphological analyzer of the Uzbek language. The FST method of morphological analysis, the model of the analysis of nouns in the Uzbek language are discussed.

**Keywords:** FST, model, Uzbek language morphological analyzer, noun, wordform, affixe

*УДК 83.00*
***Хамроева Шахло Мирджановна***
*Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои*
*hamroyeva81@mail.ru*

# КОНЕЧНАЯ АВТОМАТНАЯ МОДЕЛЬ СУЩЕСТВИТЕЛЬНЫХ ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА УЗБЕКСКОГО ЯЗЫКА

**Аннотация.** В данной статье рассматриваются модели Finite State Transduser (FST) для создания морфологического анализатора узбекского языка. Обсуждаются метод FST морфологического анализа, модель анализа существительных в узбекском языке.

**Ключевые слова:** FST, модель, морфологический анализатор узбекского языка, существительное, словоформа, аффикс.

For Turkic languages, which are included in agglutinative languages, the automatic approach of creating a morphological analyzer is more appropriate. The automated approach-based morphoanalyzer has a processing system FST (finite state transducer) and WFST (weighted finite state transducer), which can perform input and output (analysis / synthesis) analysis. The essence of such analyzers is that they follow the rule of "grammatical sequence": the

process is based on the rule of morphological unity of the word. They differ in what units are used:

1) show the sequence of morphemes and the required allomorphic sequence;

2) rules of sequence of allomorphs.

In constructing a morphological analyzer of natural language, many scientists say [1, 2, 3, 4, 5, 6] that it is an advantage to use the Finish State Machine method. In this article, we discuss the development of a model of the Uzbek language FST (finite state transducer) in the creation of a morphological analyzer of the Uzbek language.

There are key factors for automatic morphological analysis, such as stem, base, prefix, suffix, spelling rules.

To do this, you need to create a database of word-formers in Uzbek (pre-/post-stem), lexical and syntactic suffixes, particles in the form of suffixes.

If all of this forms the database of the morphological analyzer, it is necessary to run the morphological analyzer and develop an analysis model to develop the program.

Hence, for the morphological analysis of the word *"daraxtlar"* at the initial stage, the following information is required to be in the database:

1) Base (information indicating to which category the basis belongs);

2) Suffix (information representing the type).

Based on this information, the word form of *"daraxtlar"* are analyzed as follows: daraxt - ; -s is the plural form.

When Ashref Adali wrote about the creation of a morphological analyzer of Turkish and English, he distinguished the following database for the analyzer, which can identify the series of stems, additional and grammatical interpretation of the word[2]:

1) dictionary;
2) a series of suffixes;
3) spelling rules.

To do this, it is determined what suffixes a particular group of words can take. In Table 1 we present the suffixes that can be associated with a particular word and their position.

**1.1. Finite state machine (FSM) models of the Uzbek language for morphological analysis for the Uzbek language.**

It is designed to show boundary and transition states in processes involving the Restricted Case Machine (**FSM**). **SDM** and its special forms are widely used in the construction of various linguistic analyzers. To understand the Limited Case Machine (**FSM**), we need to know its basic terms.

1. Initial state: Indicates the initial state of the limited state machine. It is indicated by an arrow coming from an unknown place.

2. Acceptance status: A status indicating that the **FSM** has successfully completed its mission.

3. Recipient and Recognizers: Indicates whether the application has been accepted or not.

4. Transition motion: the transition from one state to another.

5. Converter: forms the output state using the input and motion used. There are two types of converters: (a) **FSM** uses only input motions; the output depends on the situation; (b) (**FSM**) uses only input actions: output depends on input and status.

**1.2. Spelling rules.** Spelling rules also play an important role in creating a morphological analyzer. Although Uzbek is an agglutinative language among Turkic languages, there are some cases of inflection. Flexion cases occur as sound changes. For the analysis of word forms on the basis of the morphological analyzer should reflect the typical cases of sound changes, all sound changes in the Uzbek language. There are three main types of sound changes in the Uzbek language:

**1. Sound changes:**

The vowel at the end of a word changes with the addition of an suffix:

1) When the suffix -v, -q, -qi is added to verbs ending in **a** vowel, the vowel **a** is pronounced as **o** and written as: *sayla – saylov, sina – sinov, aya – ayovsiz; so‘ra – so‘roq, bo‘ya – bo‘yoq; o‘yna – o‘ynoqi, saura – sayroqi*;

2) When most verbs ending in the vowel **i** are followed by the suffix **-v, -q,** the vowel is pronounced and written as **u**: *o‘qi – o‘quvchi, qazi – qazuvchi, sovi – sovuq*. However, in some verbs ending in the vowel **i**, when the suffix **-q** is added, the vowel **i** is pronounced and written as **i**: *og‘ri – og‘riq, qavi – qaviq*.

When the possessive suffix is added to multi-syllable words ending in **k, q,** as well as to certain syllables such as **bek, yo’q,** the consonant **k** becomes the consonant **g**, the consonant **q** becomes the consonant **g**’, and is written as follows: *tilak – tilaging, yurak – yuragim, kubok – kubogi, bek – begi; tayoq – tayog‘i, qoshiq – qoshig‘i, yaxshiroq – yaxshirog‘i, yo‘q – yo‘g‘i*. But in multi-syllable mastery words, when the possessive suffix is added to single-syllable plural words, the sound **k, q** is actually pronounced and written: *ishtirok – ishtiroki, ocherk – ocherki, erk – erki, huquq – huquqim, ravnaq – ravnaqi, yuq – yuqi.*

**2. Sound drop:**

With the addition of the following suffixes, the sound in the word structure drops:

1) When the possessive suffix is added to some words, such as o‘rin, qorin, burun, o‘g‘il, bo‘uin, ko‘ngil, and the suffix **-il**, which forms the relative form to verbs such as qayir, ayir, is added to ikki, olti, yetti words **-ov**, The vowel in the second syllable is not pronounced or written when the suffix **-ala** is added: *o‘rin – o‘rnim, qorin – qorni, burun – burning, o‘g‘il – o‘g‘ling, ko‘ngil – ko‘ngli, yarim – yarmi; qayir – qayril, ulug‘ – ulg‘ay, sariq – sarg‘ay, ikki – ikkov, ikki – ikkala, yetti – yettov;*

**3. Sound increase:**

With the addition of the following suffixes, the sound content of the word increases:

2) It is pronounced with the sound **n** when the suffixes **-da, -dan, -day, -dagi, -ga, -gacha, -cha** are added to those pronouns **u, bu, shu, o‘sha**, and it is written as follows: unda, bunday, shunda, o‘shancha; the possessive suffixes to these pronouns are added as follows; buningiz, o‘shanisi;

3) Possessive suffixes are added to words ending in **o, o, u, e** vowels as follows:

a) A lot of possessive suffixes **-m, -ng, -si; -miz, -ngiz, -si** (or **-lari**) are added without sound: bobom, bobong, bobosi, bobomiz, bobongiz, bobosi (yoki bobolari); orzum, orzung, orzusi; orzumiz, orzungiz, orzusi;

b) When the first and second person possessive suffixes are added to the words **parvo, obro, mavqe, mavzu, avzo**, a vowel is added and written as follows: parvoyim, parvoying; parvoyimiz, parvoyingiz; obro‘yim, obro‘ying; obro‘yimiz, obro‘yingiz; The possessive suffix of the third person is added to the words **parvo, avzo, obro’, mavqe** in the form **-yi**, and to the words xudo, mavzu in the form of **-si**: as avzoyi, mavzusi(as dohiy, **-si** is added to the word ending a “**-y”** consonant in the third person: like a dohiysi);

4) When the suffixes **-ni, -ning, -niki** are added to the pronouns men, sen, the sound n in the suffix is not pronounced or written: as meni, mening, meniki; seni, sening, seniki.

**1.3. Database of morphemes.** Another component required to run a morphological analyzer is a database of morphemes. Additions in the Uzbek language are divided into three groups in terms of function:

1. Word formers.
2. Syntactic formers.
3. Vocabulary formers.

In order to distinguish word-forming suffixes in the process of morphological analysis, a complete list of them should be provided. Below is a list of word-formative suffixes that are actively used in the Uzbek language.

**Table 2.**

## Noun form suffixes

| Noun - noun form suffixes (ad yapım ekleri) | | | | |
|---|---|---|---|---|
| | ek | teg | açıklama | yasalma (gövde) örneğ |
| 1 | -bin: | Ŷ 1 | noun - noun | Folbin |
| 2 | -bon: | Ŷ 1 | noun - noun | darvozabon, soyabon, tarozibon, xazinabon |
| 3 | -boz: | Ŷ 3 | noun - noun | masxaraboz, qimorboz, dorboz |
| 4 | -voy: | Ŷ 4 | noun - noun | novvoy (nonvoy) |
| 5 | -gar//-kar: | Ŷ 5 | noun - noun | zargar, savdogar, da'vogar, miskar |
| 6 | -garchilik: | Ŷ 6 | noun - noun | yog'ingarchilik, odamgarchilik |
| 7 | -gin | Ŷ 7 | noun - noun | jahongir, fazogir |
| 8 | -goh: | Ŷ 8 | noun - noun | oromgoh, saylgoh, sayrgoh, qarorgoh, ziyoratgoh, bazmgoh |
| 9 | -go’y: | Ŷ 9 | noun - noun | kalimago'y, maslahatgo'y |
| 10 | -diq//-dik | Ŷ 10 | noun - noun | o'rindiq |
| 11 | -don: | Ŷ 11 | noun - noun | guldon, kuldon, qalamdon |
| 12 | -don | Ŷ 12 | noun - noun | muhrdor, chorvador |
| 13 | -dosh: | Ŷ 13 | noun - noun | sinfdosh, kursdosh, maslakdosh |
| 14 | -do'z: | Ŷ 14 | noun - noun | etikdo'z, mahsido'z, kashtado'z |
| 15 | -zor: | Ŷ 15 | noun - noun | olmazor, gulzor, olchazor |
| 16 | -iston: | Ŷ 16 | noun - noun | guliston, go'riston, O’zbekiston |
| 17 | -kash: | Ŷ 17 | noun - noun | aravakash, qalamkash, suratkash |
| 18 | -kor: | Ŷ 18 | noun - noun | ganchkor, paxtakor, sholikor, san'atkor |
| 19 | -kov: | Ŷ 19 | noun - noun | go'rkov |
| 20 | -lik/liq: | Ŷ 20 | noun - noun | bolalik, vaqtichog'lik, do'stlik, boshliq |
| 21 | -loq: | Ŷ 22 | noun - noun | O’tloq, qumloq, toshloq |
| 22 | -noma: | Ŷ 23 | noun - noun | taklifnoma, tabriknoma, pandnoma |
| 23 | -navis: | Ŷ 24 | noun - noun | tarixnavis, voqeanavis, romannavis |
| 24 | -paz | Ŷ 25 | noun - noun | oshpaz, kabobpaz, somapaz |
| 25 | -soz | Ŷ 26 | noun - noun | soatsoz, kemasoz |
| 26 | -fuwsh: | Ŷ 27 | noun - noun | baliqfurush, nosfurush |
| 27 | -xon: | Ŷ 28 | noun - noun | kitobxon, she'rxon |
| 28 | -xona: | Ŷ 29 | noun - noun | darsxona, mehmonxona, ishxona, |

| | | | | |
|---|---|---|---|---|
| 29 | -xo'r | Ŷ 30 | noun - noun | merosxo'r |
| 30 | -cha | Ŷ 31 | noun - noun | qalamcha |
| 31 | -chak//-choq: | Ŷ 32 | noun - noun | o’yinchoq |
| 32 | -chi: | Ŷ 33 | noun - noun | ishchi, temirchi, terimchi, gulchi, bosqinchi |
| 33 | -chilik | Ŷ 34 | noun - noun | hunarmandchilik, o'zbekchilik, dehqonchilik |
| 34 | Ham- | Ŷ 35 | noun - noun | hamqishloq, hamshahar, hamyurt |
| Noun formers from the base of another word category | | | | |
| 35 | -a (l): | Ŷ 36 | imitation | qahqah, sharshar, g'arg'ar, jizz, |
| 36 | | Ŷ 37 | adjectıve - noun | bo'z, quyuq |
| 37 | -a | Ŷ 38 | adjectıve - noun | xarob, vayron |
| 38 | -ak: | Ŷ 39 | imitation | bizbiz, pirpir, guldir, var, qar, xur |
| 39 | -archilik: | Ŷ 40 | adjectıve - noun | Och |
| 40 | -at: | Ŷ 41 | verb - noun | ko'chat, o'lat |
| 41 | | Ŷ 42 | adjectıve - noun | ko'kat |
| 42 | -vchi//-uvchi: | Ŷ 43 | verb - noun | o'quvchi, yozuvchi, uchuvchi, aniqlovchi, to'kjiruvchi |
| 43 | -garchilik | Ŷ 44 | adjectıve - noun | namgarchilik, xafagarchiiik, xunobgarchilik, sharmandagarchilik |
| 44 | -gi//-ki// -qi//-g'i//-g'u: | Ŷ 45 | verb - noun | sezgi, sevgi, supurgi, kukju, turtki sanchqi, tomizg'i, tuyg'u. |
| 45 | -gach//-kich// -qich//-g'ich: | Ŷ 46 | verb - noun | kulgich, o'tkazgich, ko'rsatkich, yoritqich, tutqich, ochqich, to'g'nag'ich, chizg'ich, o'chirg'ich |
| 46 | -gin//-qin// -kin//-gun// -qun: | Ŷ 47 | verb - noun | tizgin, surgun, tolqin, to'sqin, quvg'in, yong'in, uchqun |
| 47 | -dak//-doq: | Ŷ 48 | verb - noun | yugurdak, kekirdak, qovurdoq, qo'ndoq |
| 48 | -diq | Ŷ 49 | verb - noun | qoldiq, topildiq, hordiq |
| 49 | -ik: | Ŷ 50 (1) | verb - noun | ko'rik, teshik (noun - noun va adjectıve - noun), kekirik |

| | | | | |
|---|---|---|---|---|
| 50 | -ik: | Ŷ 50(2) | verb - noun | bilik, bitik |
| 51 | -ik: | Ŷ 50(3) | adverb - noun | ko'pik |
| 52 | -ildoq: | Ŷ 51 | imitation | hiqildoq, chirildoq |
| 53 | -imlik: | Ŷ 52 | verb - noun | o'simlik, ichimlik |
| 54 | -in//-un: | Ŷ 53 | verb - noun | yig'in, yog'in, ekin, tiqin, tugun, tutun |
| 55 | -indi// undi//ndi: | Ŷ 54 | verb - noun | chiqindi, yuvundi, chirindi, cho'kindi, kuyundi, yig'indi |
| 56 | -it: | Ŷ 55 | verb - noun | chiqit |
| 57 | -ich: | Ŷ 56 | verb - noun | cho'mich, cho'kich, o'pich, bog'ich |
| 58 | -ish: | Ŷ 57 | verb - noun | Qarg'ish |
| 59 | -iq//-uq: | Ŷ 58 | verb - noun | chaqiriq, kesatiq, yutqiziq, chopiq, yutuq |
| 60 | -k: | Ŷ 59 | verb - noun | ko'rik, elak, tilak, kurak, bezak, to'shak |
| 61 | -kilik//-gilik: | Ŷ 60 | verb - noun | ichkilik, ko'rgilik |
| 62 | -lik | Ŷ 61(1) | numeratıv-noun | birlik, to‘rtlik |
| 63 | | Ŷ 61(2) | Pronoun - noun | o'zlik |
| 64 | | Ŷ 61(3) | adverb - noun | tezlik, sekinlik, birgalik |
| 65 | | Ŷ 61(4) | Modal - noun | borliq, yo'qlik |
| 66 | -m//-im//-um: | Ŷ 62(1) | verb - noun | to'plam, ho'plam, chidam, tishlam, kechirim, qo'nim, terim, chiqim, bitim, bosim, unum, tuzum |
| 67 | | Ŷ 62(2) | imitation | qultum |
| 68 | -ma: | Ŷ 63 | verb - noun | surma, o'sma, tortma |
| 69 | -mak//-moq: | Ŷ 64 | verb - noun | yemak, ilmoq, chaqmoq, topishmoq |
| 70 | -machoq: | Ŷ 65 | verb - noun | bekinmachoq, quvlashmachoq |
| 71 | -mish: | Ŷ 66 | verb - noun | noun - nounmish, kechmish, qilmish. |
| 72 | -movchilik: | Ŷ 67 | verb - noun | anglashilmovchiiik, kelishmovchilik, yetishmovchilik |
| 73 | -on: | Ŷ 68 | verb - noun | qiron, to'zon |
| 74 | -os: | Ŷ 69 | imitation | uvvos, chuvvos |
| 75 | -ot | Ŷ 70 | Adjective-noun | ma'lumot, mushkulot xarobot |
| 76 | -at: | Ŷ 71 | adjectıve - noun | she'riyat, madaniyat, majburiyat |

| | | | | |
|---|---|---|---|---|
| 77 | -ch//-j//-inch: | Ŷ 72 | verb - noun | sevinch, quvonch, yupanch, ilinj, qo'rqinch |
| 78 | -cha: | Ŷ 73(1) | verb - noun | tushuncha |
| 79 | -cha: | Ŷ 73 (2) | adjectıve - noun | qizilcha, olacha |
| 80 | -chak | Ŷ 74 | verb - noun | belanchak, ovunchoq, taqinchoq |
| 81 | -chi | Ŷ 75(1) | adjectıve - noun | qiziqchi |
| 82 | -chi | Ŷ 75(2) | verb - noun | suyunchi, tilanchi, tomchi |
| 83 | -chilik: | Ŷ 76(1) | undov | haybarakallachi |
| 84 | -chilik: | Ŷ 76(2) | adjectıve - noun | pishiqchiiik, arzonchilik |
| 85 | -chiq: | Ŷ 77 | adverb - noun | ko'pchilik, ozchilik |
| 86 | -shunos: | Ŷ 78 | verb - noun | suyanchiq, yopinchiq |

These suffixes come in handy in distinguishing word-forming suffixes from other types of suffixes in the process of morphological analysis.

Various lexical and syntactic suffixes are also actively used in the Uzbek language. We present these additions in Table 3 below.

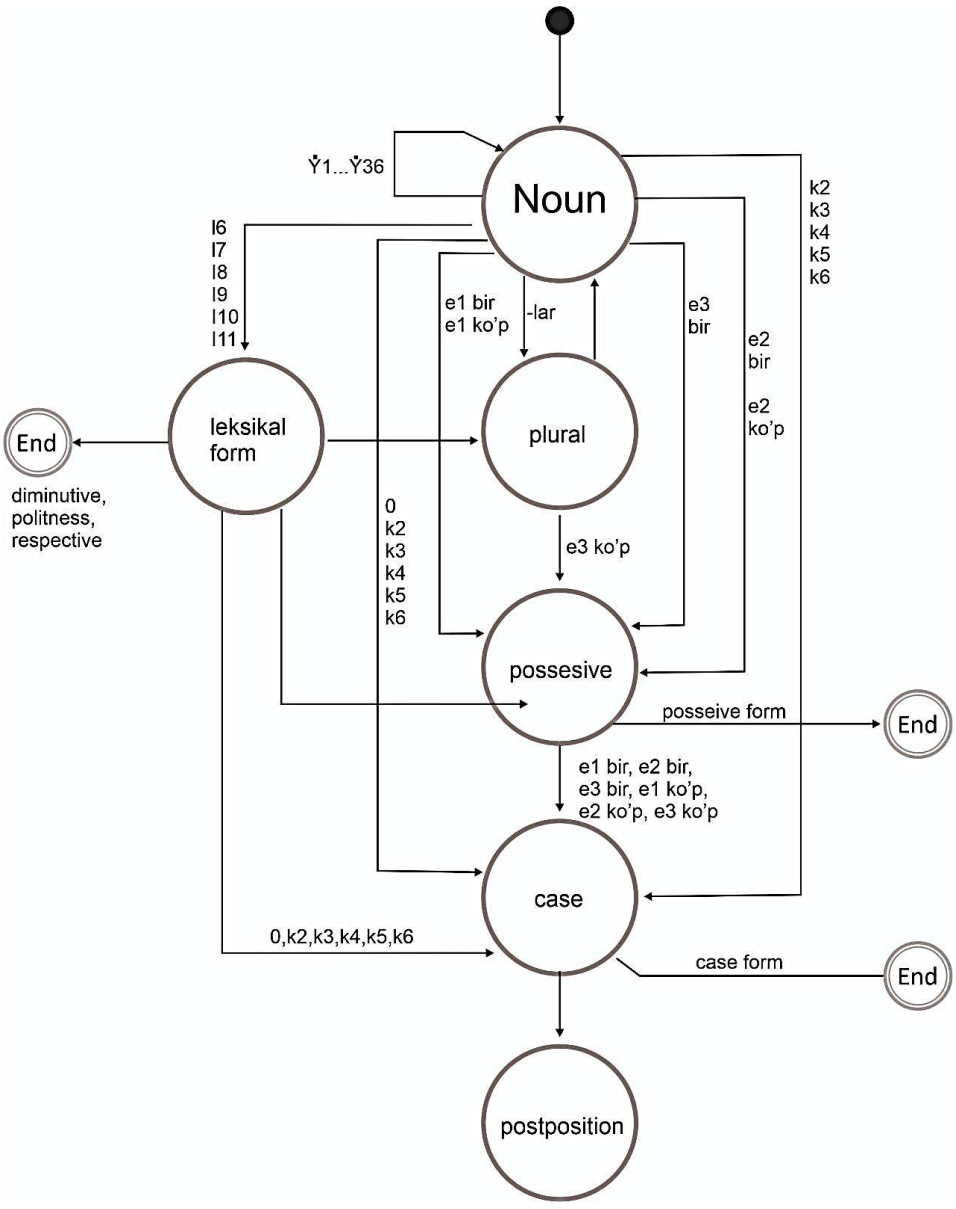
**Table 3.**

| **I** | **VOCABULARY FORMERS** | | | |
|---|---|---|---|---|
| **I.5** | **Number form** | | | |
| 1) | singular | -Ø | s1 | Indicates that it is singular |
| 2) | plural | -lar | s2 | Indicates that there are many |
| **I.6** | **Decreasing** | -choq// chak | kich1 | It means to diminution and to caress |
| | | -kay | kich2 | It means to diminution and to caress |
| | | -ak | kich3 | It means to diminution and to caress |
| **I.7** | **Personal attitude** | -jon | shm1 | Positive attitude |
| | | -xon | shm2 | Positive attitude |
| | | -oy | shm3 | Positive attitude |
| | | -loq/ aloq | shm4 | Positive attitude, caress |
| | | -gina/ kina/ qina | shm5 | Positive attitude, caress |
| | | -boy | shm6 | Positive attitude, respect |
| | | -bek | shm7 | Positive attitude, respect |
| **I.8** | **Dependence** | -niki | q | Dependence |
| **I.9** | **Place and** | -dagi | o‘-p | Means place and time |

| | | | | |
|---|---|---|---|---|
| | **time** | | | |
| **I.10** | **Limit** | -gacha/ kacha/ qacha | ch | Time and place and sometimes mean boundaries between different objects |
| **I.11** | **Assimilation** | -dek | o‘x1 | means assimilation |
| | | -day | o‘x2 | means assimilation |
| **II.4** | **Kelishik** | | | |
| 1) | Case 1 (Bosh kelishik) | -Ø | k1 | Connects words |
| 2) | Case 2 (tushum kelishigi) | -ning | k2 | Indicates that an object or event belongs to another object or person |
| 3) | Case 3 (qaratqich kelishigi) | -ni | k3 | Represents the ability to completely take an action of an object, percived from the added base |
| 4) | Case 4 (jo`nalish kelishigi) | -ga/ ka/ qa | k4 | Represents the direction of motion, object, place, time, and so on |
| 5) | Case 5 (o`rin-payt kelishigi) | -da | k5 | Action represents the point where an object have to be, the object, the place, the time |
| 6) | Case 6 (chiqish kelishigi) | -dan | k6 | Harakatning chiqish nuqtasi, obyekt, o‘rin, payt, sabab ma’nolarini anglatadi |
| **II.5** | **Possession** | | | |
| 1) | I Person singular | -m/ im | e1bir | Means I Person singular and possession |
| 2) | II Person singular | -ng/ ing | e2bir | Means II Person singular and possession |
| 3) | III Person singular | -i/ si | e3bir | Means III Person singular and possession |
| 4) | I Person plural | -miz/ imiz | e1ko‘p | Means I Person plural and possession |
| 5) | II Person plural | -ngiz/ ingiz | e2ko‘p | Means II Person plural and possession |
| 6) | III Person plural | -i(lari)/ si(lari) | e3ko‘p | Means III Person plural and possession |

## 2. About Uzbek FSM models

**2.1. Morphological analysis of word group noun.** Word-formation, lexical and syntactic suffixes are added to the nouns in the Uzbek language. In the Uzbek language, noun-forming suffixes from different stems are shown in Table 2. Possession and case suffixes to the noun are set out in Table 3. In the Uzbek language, the forms of words in the noun category are usually arranged in the following sequence: Base + word-former + number + possession + case. We have given these suffixes in the tables above, which are necessary in the analysis of the Uzbek language horse phrase based on FSM models. The rules of spelling formed as a result of the addition of a horse are given in paragraph 1.2. The FSM model of the analysis based on these appendices is given in Figure 1 below.




The first introductory part of the analysis is marked with a black dot. To make the model clear, the tags shown in Table 2-3 of the appendices are

written between actions. When suffixes are involved in the transition from one grammatical form to another, they are indicated by an arrow, cases of non-indicative change of grammatical meaning are indicated by an empty arrow. We explain the diagram below.

**Noun is the base.** The condition indicated by a black dot indicates a word coming from outside for analysis. This provides access to the FSM. It is considered an artificial noun with Ŷ1 ... Ŷ35 suffixes in the base. Even if a noun-former comes after the base, the noun is made. For example: *fol+bin, darvoza+bon, soya+bon, zar+gar, savdo+gar, nam+garchilik, xafa+garchilik, sez+gi, sev+gi.*. The drawing shows that a new noun was made from the noun's core. The noun can be in a state of transition from the base to several forms. These can be noun cases in the plural, lexical form, possessive, and accusative forms.

**Plural:** The form base+lar makes the plural forms of the noun: *bola+lar, olma+lar*. The word can end in this form or take the forms of possession and agreement, so there are three exits from this form:

1) plural noun: *qizlar, gullar*;

2) a plural noun that takes the form of a case: *qizlarlarning, qizlarni, qizlarga, qizlarda, qizlardan*;

3) noun in possession and plural: *qizlarim, qizlaring, qizlari, qizlarimiz, qizlaringiz, qizlari*.

**Possession:** status e1bir, e2bir, e3bir, e1kop, e2kop, e3kop take suffixes have many symbols and have two outputs:

1) A noun in the form of possession: *kitobim, kitobing, kitobi, kitobimiz, kitobingiz, kitoblari*;

2) A noun in the form of possession and case: *kitobim, kitobimning, kitobimni, kitobimga, kitobimda, kitobimdan*.

**The state of the lexical form of the noun:** the lexical form of the noun is formed by suffixes marked with tags I6, I7, I8, I9, I10, I11. For example: *odamniki, tog‘agi, hovligacha, bo‘taloq, onaxon, ukajon*. There will be 3 exit points from this case:

1) The lexical form of the noun: *uydagi, opajon*;

2) The lexical form of the possessive noun: *uyimizdagi, onaxonimiz*;

3) The lexical form of the plural noun: *uydagini, onajonimni*.

The state of the noun in the **form of a case** is formed in the following conditions.

*The first condition*. A direct case suffix is added to the base: base + 0 / k2 / k3 / k4 / k5 / k6. For example: *xona, xonaning, xonani, xonaga, xonada, xonadan*.

*The second condition*. To the base that takes the lexical form, an case suffix is added: base + I6 / I7 / I8 / I9 / I10 / I11 + 0 / k2 / k3 / k4 / k5 / k6.

For example: *bolajon, bolajonning, bolajonni, bolajonga, bolajonda, bolajondan*.

*The third condition*. A case suffix is added to the base that takes the form of possession: base + e1bir / e2bir / e3bir / e1kop / e2kop / e3kop + 0 / k2 / k3 / k4 / k5 / k6. For example: *kitobim, kitobimning, kitobimni, kitobimga, kitobimda, kitobimdan*.

There can be two output states from the contract form. A noun in the form of a case and a noun that receives an auxiliary after the contract.

**References**

1. Тузов В.А. Морфологический анализатор русского языка //Вестник СПбГУ, сер. 1. 1996. Вып. 1 (N15). − С. 41-45.

2. Формальные модели и програмные инструменты компьютерной обработки татарского языка / Р.Р.Гатауллин, А.Р.Гатиатуллин, О.А.Неврозова, Д.Р.Мухамедшин, Д.Ш.Сулейманов, Б.Э.Хакимов, А.Ф.Хусаинов. – Академия наук РТ, Институт приклалной семиотики АН РТ. – Казань: Академии наук, 2019. – 260 с. – С.39-40.

3. Eşref Adalı. Turkçe Doğal dil İşlemi. – Ankara, 2020. – 754 s.

4. Gelbukh A. F. Effectively realizable morphologic model of inflective language [Эффективно реализуемая модель морфологии и эктивного языка]. Научно-техническаяИнформация [Scientific and Technical Information], series 2, №1, 1992. – P. 24-31.

5. Kemal Oflazer. Two-level Description of Turkish Morphology. Literary and Linguistic Computing, – Vol. 9, No 2, – 1994.

6. Çağrı Çöltekin (2014) A Set of Open Source Tools for Turkish Natural Language Processing In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) Ed. by N. Calzolari et al. 1079-1086.

*UDC 004.89*
***[1]Nasrullayeva A., [2]Mukanova A.***
*Astana International University,*
*Nur-Sultan, Kazakhstan*
*[1]nasrullayevaik@gmail.com, [2]asiserikovna@gmail.com*

## RESEARCH AND ANALYSIS OF MECHANISMS FOR DETECTING PROHIBITED CONTENT ON THE INTERNET

**Abstract:** It is no secret that the internet is a source of various information that is not always safe for users. In addition, the network often meets and communicates with a large number of people, among whom there may be those who try to use the digital space and trust of the interlocutor to achieve criminal goals. Thus, there was a need to process a large amount of information to determine the attitude of users to a particular object.
**Keywords:** Internet, tonality, negative, positive, neutral, microblog.

*УДК 004.89*
***[1]Насруллаева А. Б., [2]Муканова А. С.***
*Международный университет Астана*
*Нур-Султан, Казахстан*
*[1]nasrullayevaik@gmail.com, [2]asiserikovna@gmail.com*

## ИССЛЕДОВАНИЕ И АНАЛИЗ МЕХАНИЗМОВ ОБНАРУЖЕНИЯ ЗАПРЕЩЕННОГО КОНТЕНТА В СЕТИ ИНТЕРНЕТ

**Аннотация:** Ни для кого не секрет, что интернет является источником разнообразной информации, которая не всегда безопасна для пользователей. Кроме того, в сети часто встречается и общается большое количество людей, среди которых могут быть те, кто пытается использовать цифровое пространство и доверие собеседника для достижения преступных целей. Таким образом, возникла необходимость обработки большого объема информации для определения отношения пользователей к тому или иному объекту.
**Ключевые слова:** Интернет, тональность, негатив, позитив, нейтрал, микроблог.

Since 2017, many countries have adopted laws to introduce additional mechanisms and measures to systematize content on the internet. These measures, first of all, plan to prevent the dissemination of offensive

information in the digital environment in accordance with estimates that are considered contrary to national legislation. It should be noted that due to the different approaches of each country, the same concept and terminology of "offensive (destructive) content" has not yet been developed. Depending on the size of the development and growth of large multinational internet companies, the number of violations of citizens ' rights and freedoms in the digital environment increases proportionally, and one of the "foundations" of digital freedom of states is the introduction of a mechanism to protect the internal virtual environment from negative or harmful information. As a result of the release of information products (News Services, film industry, social networks, entertainment industry, etc.), a large number of socio-political, historical, religious, cultural and moral views are absorbed. Taking into account such threats, many states have developed and implemented a systematization base that corresponds to the internet in order to protect their information security, national economic views, and the rights of their citizens. The most pressing issues of legal regulation of relations in the global digital space include the Prevention of the dissemination of prohibited information, the fight against cybercrime, the protection of intellectual rights in the digital environment, freedom of speech and access to information, the protection of personal information and the rights of users on the Internet, taxation and the protection of competition between electronic organizations.

It is no secret that microblogging is the most popular tool among internet users. On internet platforms such as Twitter, futubra, and facebook, users leave thousands of messages. They write about their lives, discuss news, and express their opinions about various types of goods and services. In recent years, the number of microblog drivers among internet users has increased and activity has increased. In this regard, we will highlight the specific properties of the microblog:

- The tools needed for microblogging are very accessible, and working with it allows the public to;
- There is no specific template for leaving a message, that is, you can write freely, without limiting yourself.
- The use of symbols to write a message is limited (you don't have to give your thoughts too long here);
- The more microblogging is used by users ' friends, the wider the range of users (the snowball method).

In microblogging, users often express their views on the products they use, talk about the services provided to them, and increase their readers in microblogging by raising issues related to politics and religion that are relevant for social and marketing research. Microblogging messages are used to divide text messages into positive, negative, and neutral classes. For

research, the microblogging platform Twitter was chosen and the reason for this was the following:

- The Twitter platform is one of the most popular platforms today. The number of posts and their users is increasing every day. There is enough information for appropriate research.
- Microblogs are used to express their subjective opinions on different things, so the microblogging record template is ideal for conducting tone analysis;
- The socio-demographic range of the audience on this platform is diverse: from students and stars of show business to politicians and presidents, this platform is another proof of its popularity. For this reason, twitter also allows you to collect information related to the interests of people from different groups.
- Users of the Twitter platform also cover different countries and continents: this will allow them to expand the scope of work, including different languages in research in the future.

With the twitter API, the case was collected from 31,000 twitter posts. The Twitter API has a limited range of capabilities, for example, it can generate up to 1,000 twitter posts based on a geographical link during each search. The geographical link was used to collect twitter posts exclusively in Russian.

As a result, a collection of texts containing about 15 million short messages was collected:

- positive (114 991 messages);
- negative ( 111 923 messages);
- neutral (107 990 messages)

The case was divided into three parts: positive, negative and neutral. Since users are not limited to either the format or the form of writing messages in microblogs, it is not enough to create a generalized solution or create a single dictionary to determine the emotionality of any abstract messages. Therefore, the method [1] was used to identify positive or negative messages. To summarize the case, a search was conducted for queries to express an emotional attitude to something. It is possible to identify emotions with high accuracy if in the user's texts the author indicates signs (emoticons) that express emotions in the message:

- Positive : “:-)”, “:)”, “-)”, “=)”, “:0”, “))”, “haha” (in different versions);
- Negative : “:-(”, “:(”, “=(”, “;(” , “:’-(” and so on.

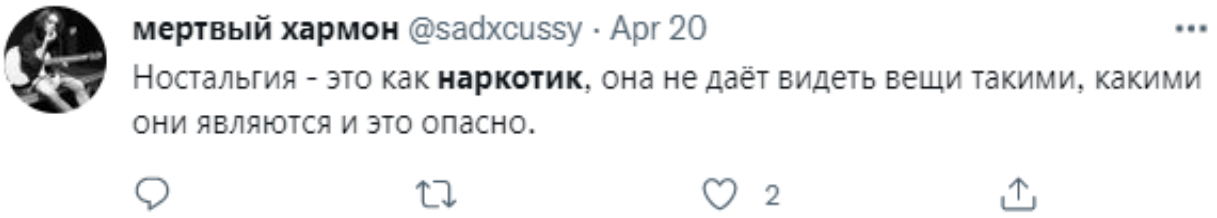
In accordance with the written sign of emotions, measures were taken to search for positive and negative messages, and two samples were selected accordingly. The two selected options will be used to analyze positive and

negative tweets, identify general trends, and create a structure of positive, negative, and neutral messages. The Test selection was sorted by the following conditions:

- All tweets of positive and negative meaning have been disabled.
- Such texts negatively affect the results of the analysis and do not allow us to determine the emotionality of messages;
- All retweets are also disabled. As a rule, retweets are accompanied by the abbreviation RT. These texts may add additional meaning to parts of speech during analysis.
- As a result of research, the twitter API gives results when getting a copy of twitter posts. Records of the same value were excluded from the test sample at the time of the analysis, so as not to add additional meaning in parts of speech.

To study the influence of linguistic features of the text on the classification of short texts with tonal coloring, morphological features of collections were developed. The purpose of the designation is to determine the patterns of distribution of speech parts between "positive" and "negative" collections.

For example,




Picture 1-tweet posted on the social network Twitter

A tweet posted on the social network Twitter in Picture 1, let's look at it in detail into individual words :

| | |
|---|---|
| Ностальгия | 1 |
| это | 2 |
| как | 3 |
| наркотик | 4 |
| она | 5 |
| не | 6 |
| дает | 7 |
| видеть | 8 |
| вещи | 9 |

| | |
|---|---|
| такими | 10 |
| какими | 11 |
| является | 12 |
| и | 13 |
| опасно | 14 |

Table 1. numbering of publications

Each word is assigned an individual number. This is how the dictionary is created. Words in the dictionary have a two-tone designation. They are negative and positive.

Thus, after testing the words, we can determine the tonality of the words in the Real life testing.

For example,

```
tweets_for_testing = [
    "Они планирует организовать теракт "
]
for tweet in tweets_for_testing:
    test_tweet(tweet)
    print("---------")
```

Picture 2- Result from Real life testing

Result: Unknown token: планирует
Original tweet: Они планирует организовать теракт
P(positive) = 0.00285. Result: Negative

```
tweets_for_testing = [
    "я счастлива"
]
for tweet in tweets_for_testing:
    test_tweet(tweet)
    print("---------")
```

Picture 3- Result from Real life testing

Result: Original tweet: я счастлива
P(positive) = 0.98378. Result: Positive

But there is an error here, because the expression of emotions by the author depends on his tendency to repeatedly use the same thing in speech.

For example, an author with emotions often used classification pronouns such as "I, you, he, they" in his writing. And in tweets with a neutral value, it uses names. In emotional notes, authors often write about themselves and their experiences, so pronouns on the first and second sides are most noticeable in them. If pronouns are used in messages with a neutral value, they are reflected in the third person.

In messages with a neutral meaning, nouns with a noun and a participle are actively used. Numbers are also often used.

While the adjective amplification is reflected in emotional messages, the relative radiance is found in messages with a neutral meaning.

Adverbs are often used in emotional messages to increase the meaning of the verb, and in other cases in messages with a neutral meaning.

With the help of such studies, it was possible to establish its own regularity of determining the meaning of messages, and samples were developed that helped create a classification algorithm for using parts of speech.

At the same time, mutual comparisons were made between emotional notes. As a result, in tweets with a negative meaning, the form of the Past Tense is often found, expressing people's regret and dissatisfaction with the past, and at the same time expressing their joyful (positive) state in the present tense. Verbs are more common in negative messages than in positive messages.

In the course of the work, the collection of texts was filtered according to the filters proposed by the authors to reduce possible noise. As a result of studying the body of the received texts, it was found that depending on the tone of the message, the authors of the message often use a certain part and form of speech.

**References**

1. Jonathon Read. 2005 Using emoticons to reduce dependency in machine learning techniques for sentiment classification. InACL. The Association for Computer Linguistics.

2. Hu M., Liu B. Mining and Summarizing Customer Reviews. KDD, Seattle, 2004.

3. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышински Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных// Изд-во НИУ ВШЭ, 2017.

4. Четверкин И. И. , Лукашевич Н. В. Тестирование систем анализа тональности на семинаре РОМИП-2012 // Т. 2: Доклады специальных секций РОМИП — М.: Изд-во РГГУ, 2013.

5. Erik Boiy, Marie-Francine Moens. A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts// Information Retrieval, Volume 12, Number 5 (2009), (pp. 526-558).

6. Milos Radovanovic, Mirjana Ivanovic. Text mining: approaches and applications// Novi Sad Journal of Mathematics 38(3), 2008, (pp. 229-233).

7. Adnan Duric, Fei Song. Feature Selection for Sentiment Analysis Based on Content and Syntax Models// ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2011.

*ӘОК 004.8*
***[1]Оралбекова І.Т., [2]Ергеш Б.Ж.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]iinkar9822@gmail.com, [2]b.yergesh@gmail.com*

## ҚАЗАҚ ТІЛІНДЕГІ ҚОНАҚҮЙЛЕР ТУРАЛЫ ПІКІРЛЕРДІҢ РЕҢКІН АСПЕКТІЛЕРГЕ ТАЛДАУ

**Аңдатпа**. Интернетте пайдаланушыларға өз пікірлерімен алмасуға және тауарлар мен қызметтердің барлық түрлері туралы пікірлер қалдыруға мүмкіндік беретін көптеген платформалар бар. Бұл пікірлер басқа пайдаланушылар үшін ғана емес, сонымен қатар өздерінің беделін қадағалап, өнімдері мен қызметтері туралы дер кезінде кері байланыс алғысы келетін компаниялар үшін де пайдалы болуы мүмкін. Бұл саладағы мәселенің ең егжей-тегжейлі тұжырымы пайдаланушының жалпы объектіге ғана емес, сонымен қатар оның жеке аспектілеріне деген қатынасын анықтайтын аспектілі-бағытталған сентимент талдауда қойылады. Мақалада машиналық оқыту негізінде қонақүй туралы пікірлерді аспектілерге бөлу мәселесін шешу қарастырылады.

**Түйін сөздер:** машиналық оқыту, сентимент талдау, қонақүй туралы пікілер, Naive Bayes, SVM.

*УДК 004.8*
***[1]Оралбекова І.Т., [2]Ергеш Б.Ж.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]iinkar9822@gmail.com, [2]b.yergesh@gmail.com*

## АНАЛИЗ ТОНАЛЬНОСТИ КОМЕНТАРИЕВ ОБ ОТЕЛЯХ НА КАЗАХСКОМ ЯЗЫКЕ

**Аннотация**. В Интернете существует множество площадок, которые позволяют пользователям делиться своим мнением и оставлять комментарии обо всех видах товаров и услуг. Эти комментарии могут быть полезны не только другим пользователям, но и компаниям, которые хотят следить за своей репутацией и своевременно получать отзывы о своих продуктах и услугах. Наиболее детальная постановка проблемы в этой области производится при анализе аспектно-ориентированных настроений, определяющих отношение пользователя не только к общему объекту, но и к его отдельным аспектам. В статье

рассматривается решение задачи разделения мнений об отеле на аспекты на основе машинного обучения.

Для определения эффективности программы было проведено экспериментальное исследование и проанализированы результаты. В ходе эксперимента использовались методы обучения учителей машинному обучению: метод наивного Байеса, линейный классификатор SVM и классификаторы логистической регрессии.

Результаты могут быть использованы для мониторинга общественного мнения, проведения маркетинговых кампаний, оценки новостных событий, прогнозирования мнений на основе проанализированных текстов, выявления эмоционального насилия. Анализ настроений позволяет компаниям или любому предпринимателю изменить комплекс маркетинговых мероприятий для улучшения положения продукта на рынке, выявить сильные и слабые стороны своих продуктов и услуг конкурентов и фирм. Анализ настроений на уровне аспектов обычно представляет собой подробный (конкретный) уровень, необходимый для практического применения. На этом основаны многие промышленные системы. Несмотря на большую работу в исследовательском сообществе и создание множества систем, проблема до сих пор решается. Каждая внутренняя задача остается очень сложной задачей.

**Ключевые слова:** машинное обучение, сентимент анализ, коментарии о гостиницах, Naive Bayes, SVM.

*UDC 004.8*

***[1]Oralbekova I., [2]Yergesh B.***
*L. N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]iinkar9822@gmail.com , [2]b.yergesh@gmail.com*

## ANALYSIS OF THE TONALITY OF COMMENTS ABOUT HOTELS IN THE KAZAKH LANGUAGE

**Abstract**. There are many platforms on the Internet that allow users to share their opinions and leave comments about all kinds of goods and services. These comments can be useful not only to other users, but also to companies that want to monitor their reputation and receive feedback on their products and services in a timely manner. The most detailed statement of the problem in this area is made when analyzing aspect-oriented moods that determine the user's attitude not only to the general object, but also to its

individual aspects. The article deals with the solution of the problem of dividing opinions about the hotel into aspects based on machine learning.

To determine the effectiveness of the program, a pilot study was conducted and the results analyzed. During the experiment, methods for teaching machine learning to teachers were used: the naive Bayes method, the SVM linear classifier, and logistic regression classifiers.

The results can be used to monitor public opinion, conduct marketing campaigns, evaluate news events, predict opinions based on analyzed texts, and identify emotional abuse. Sentiment analysis allows companies or any entrepreneur to change the marketing mix to improve the position of the product in the market, to identify the strengths and weaknesses of their products and services of competitors and firms. Aspect-level sentiment analysis is usually the detailed (specific) level required for practical application. Many industrial systems are based on this. Despite a lot of work in the research community and the creation of many systems, the problem is still being solved. Every internal task remains a very difficult task.

**Keywords**: Machine learning, sentiment analysis, hotel reviews, Naive Bayes, SVM.

**Кіріспе**

Интернет-ресурстар – қарым-қатынасқа, пікірталасқа және жаңа ақпаратты іздеуге ыңғайлы алаң. Белгілі бір реңкті қамтитын пайдаланушы пікірлері өздерінің беделін қадағалап, өз өнімдері мен қызметтері туралы дер кезінде кері байланыс алғысы келетін компаниялар үшін өте маңызды. Бұл саладағы мәселенің ең егжей-тегжейлі тұжырымы – пайдаланушының жалпы объектіге ғана емес, сонымен қатар оның жеке аспектілеріне деген қатынасын анықтайтын аспектіге бағытталған сентимент талдау (АБСТ). Мысалы, мейрамхананың шолуында «Менің қызмет көрсетуге ешқандай шағымым жоқ және маған мұндай интерьер ұнайды, бірақ француздық сиыр еті дәмсіз болды», үш аспектіні ажыратуға болады - қызмет көрсету, интерьер және тағам. Аспекттердің реңктері әртүрлі болуы мүмкін. Әрбір аспект әртүрлі сөздер немесе сөз тіркестері арқылы көрсетіыледі, олар аспектілік терминдер (АТ) деп аталады. Келтірілген мысалда АТ «қызмет», «интерьер» сөздері және «француздық сиыр еті» тіркесі болып табылады [1].

**АБСТ саласындағы зерттеулерге шолу**

Жұмыс [1] АБСТ-дың барлық ішкі міндеттерін шешуге арналған тәсілдер мен әдістердің егжей-тегжейлі шолуын ұсынады. Осы зерттеу нысаны болып табылатын АТ алудың қосалқы міндетін шешу үшін келесі тәсілдер қолданылады:

– бақыланатын машиналық оқыту әдістерін пайдаланып аспектілерді алу [3], [4];

– жиі кездесетін зат есімдер мен есімді тіркестерді (requent nouns and noun phrases) іздеу [2], [7], [10];

– пікір мен объектінің арақатынасына негізделген аспектілерді шығару (relation-based methods) [2];

– тақырыптық модельдеу (topic modeling) арқылы аспектілерді шығару [6].

**Классификаторлар**

Белгілі болғандай, мәтіннің тоналдылығын талдау үшін көбінесе Naive Bayes классификаторы және SVM әдісі қолданылады. Бұл ретте мәселенің типтік тұжырымы – фильмдер немесе тауарлар туралы, саяси тұлғалар немесе оқиға туралы (блогта, твиттерде және т.б.) пікірлерді оң және теріс болып жіктеу қажет. Naive Bayes моделі әдетте негізгі, ең қарапайым үлгі ретінде пайдаланылады, ал SVM әдісі күрделірек, өйткені ол тиімдірек деп саналады. Дегенмен, Бермингем мен Смитон [1] және Ванг пен Мэннинг [9] Naive Bayes классификаторы твиттер сияқты қысқа мәтіндерде SVM-ге қарағанда жақсырақ жұмыс істейтінін көрсетті. Сонымен қатар, кейбір жұмыстарда сөздердің векторлық көрінісі бар нейрондық желілерді белгілер ретінде қолдануға негізделген тәсіл ұсынылады. Бұл тәсіл тілге тәуелді емес және семантикалық тезаурусты қажет етпейді. Жұмыстың бөлігі ретінде Naive Bayes классификаторы, логистикалық регрессия және Linear SVM сияқты классификаторларымен эксперименттер жүргізілді.

**Пікірлер корпусы**

Пікірлер (www.TripAdvisor.com) және (www.booking.com) сайттарынан алынды.

Пікірледің ұзақтығы 1-ден 10 сөйлемге дейін өзгереді, орташа есеппен 5 сөйлемді құрайды. Корпустың шағын бөлігі кейінгі машиналық оқыту үшін белгіленеді. Бұл корпусқа Нұр-Сұлтан қаласындағы 5 қонақүйдің 100 пікірі кіреді.

**Қонақүй аспектілерін анықтау**

Бұл жұмыста объектілердің аспектілерінің тізімі пікірден автоматты түрде алынбайды, бірақ көңіл-күйді талдау модулі болатын жүйені пайдаланушылардың қажеттіліктері негізінде қолмен құрастырылатын тәсіл қолданылады. Пікірлерден алынған аспектілер тізімінде мекеме түрі мен асханасы, тағам мен қызмет көрсету сапасы, жайлылықтың болуы, романтикалық атмосфера, бардың, би алаңының, балалар бөлмесінің болуы, жақын жерде сауда үйінің болуы, автотұрақ және т.б кіреді. Осы жұмыста қарастырылатын қонақүйдің аспектілері кестеде көрсетілген (Кесте 1). Әр аспекті үшін кестеде осы аспект

қабылдай алатын мәндер жиынтығы көрсетілген. Қонақүйдің объективті аспектілері де (мысалы, би алаңының, бардың, балалар бөлмесінің және т.б. болуы) және субъективтік аспектілері бар . Объективті аспектілер үшін аспектілерді жіктеу тапсырмасы ақпаратты алу тапсырмасы болып табылады, ал субъективті аспектілер үшін бұл сезімді талдау тапсырмасы.

Бұл жұмыста 3 баллдық шкаламен бағаланатын { -1; 0; 1 } мәндерімен белгіленетін аспектілер қарастырылады және реңкті талдау мәселесі шешіледі.

Кесте 1

Қонақүй аспектілері (Аспекты гостиниц / Aspects of hotels)

| ***Аспект*** | **Мәндер жиыны** |
|---|---|
| *Food* | {-1; 0; 1;} |
| *Location* | {-1; 0; 1;} |
| *Room* | {-1; 0; 1;} |
| *Service* | {-1; 0; 1;} |
| *General* | {-1; 0; 1;} |

– Әрбір пікірге сәйкес реңк беріледі: Оң,теріс және бейтарап;
– Сөйлем немесе пікір бірнеше аспектілерден тұруы мүмкін;
– Тапсырма: пікірлерді аспектіге бағытталған сентимент талдау.

**Бағалау өлшемдері**

*Accuracy (Дәлдік)* – кең таралған және түсінуге оңай метрика. Бұл барлық дұрыс болжамдардың барлық болжанған үлгілердің жалпы санына қатынасы. Бірқатар тапсырмаларда дәлдік ақпаратсыз болуы мүмкін.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

*Дәлдік (precision)* – бұл барлық оң болжамды объектілер үшін шын мәнінде оң нәтиже болып табылатын болжамды оң нәтижелердің үлесі. *Precision* формула (2) арқылы есептелінеді.

$$precision = \frac{TP}{TP+FP} \quad (2)$$

*Толықтық (recall)* – барлық шынайы-оң болжанған объектілердің шын мәнінде оңды объектілердің жалпы санына пропорциясы. Яғни, толықтық барлық оң мысалдардың қанша үлгі дұрыс жіктелгенін көрсетеді. Ол формула (3) арқылы есептелінеді.

$$recall = \frac{TP}{TP+FN} \quad (3)$$

*F-өлшем* – толықтық пен дәлдіктің өлшенген гармоникалық ортасы. Бұл көрсеткіш модель қанша жағдайды дұрыс болжайтынын және үлгінің қанша шынайы дананы өткізіп жібермейтінін көрсетеді.

Precision және recall, неғұрлым жоғары болса, соғұрлым жақсы екені анық.

Бірақ нақты өмірде максималды дәлдік пен толықтыққа бір уақытта қол жеткізу мүмкін емес, белгілі бір тепе-теңдікті іздеу керек.

Сондықтан, біз алгоритмнің дәлдігі мен толықтығы туралы ақпаратты біріктіретін белгілі бір метрикаға ие болғымыз келеді. Бұл жағдайда бізге қандай жүзеге асыруды өндіріске енгізу туралы шешім қабылдау оңайырақ болады. F-өлшемі дәл осындай метрика [10].

F-өлшем пайдаланылатын модельдің толықтығы мен дәлдігі туралы ақпаратты біріктіреді. F-өлшем формула (4) арқылы есептелінеді.

$$F = 2\frac{Precision * Recall}{Precision + Recall} \quad (4)$$

**Тәжірибе нәтижесі**

Бағалау шаралары precision, recall, f1-score және accuracy қамтиды. Нәтижелер төменде кесте түрінде (Кесте 2) берілген.

Кесте 2

Тәжірибе нәтижесі (Результат эксперимента / Experiment result)

| ***Classifier*** | | **Naive bayes** | **Logistic regression** | **Linear SVM** |
|---|---|---|---|---|
| *Accuracy* | | 0.7 | 0.7 | **0.75** |
| *Service* | Precision | 0.75 | 0.60 | 0.60 |
| | Recall | 1.00 | 1.00 | 1.00 |
| | F1- score | **0.86** | 0.75 | 0.75 |
| *General* | Precision | 0.67 | 0.67 | 0.67 |
| | Recall | 0.80 | 0.80 | 0.80 |
| | F1- score | 0.73 | 0.73 | **0.75** |
| *Food* | Precision | 0.67 | 0.67 | 1.00 |
| | Recall | 0.67 | 0.67 | 0.67 |
| | F1- score | 0.67 | 0.67 | **0.80** |
| *Room* | Precision | 0.67 | 0.75 | 0.75 |
| | Recall | 1.00 | 0.75 | 0.75 |
| | F1- score | **0.80** | 0.75 | 0.75 |
| *Location* | Precision | 1.00 | 1.00 | 1.00 |
| | Recall | 0.20 | 0.40 | 0.60 |
| | F1- score | 0.33 | 0.57 | **0.75** |

Linear SVM классификаторын қолдану әдісі ең тиімді екені анықталды. Оның accuracy мәні 0,75-ке тең болды. Яғни, ол 75 % дәлдікпен жұмыс жасайды. Naive bayes әдісі мен Logistic regression әдісі де аса төмен нәтиже көрсеткен жоқ. Олардың accuracy мәндері 0,7-ге тең. Linear SVM әдісі ‘General’ , ‘Food’ және ‘Location’ аспектілері үшін жақсы мән көрсетті. General – ‘0,75’-ке , Food – ‘0.8’-ге , Location – ‘0.75’-ке тең. Ал Naive bayes әдісі ‘Service’ және ‘Room’ аспектілерінде жоғары нәтиже берді. Service – ‘0.86’-ға, Room – ‘0,80’-ге тең болды.

**Қорытынды**

Программаның тиімділігін анықтау үшін эксперименттік зерттеу жүргізілді және соған сәйкес нәтижелерге талдау жасалды. Тәжірибе барысында Naive bayes әдісі, Linear SVM классификаторы және Logistic Regression классификаторлары қолданылды.

Нәтижелер мәліметтің көлеміне байланысты өзгеруі мүмкін. Деректер жиыны неғұрлым көбірек болса, соғұрлым дәлірек жұмыс істейді. Аспектілі бағытталған сентимент талдауды іске асыру үшін осы Naive bayes әдісін және SVM классификаторын қолдану тиімді екені анықталды. Бұл әдістер қарапайым және ыңғайлы болып табылады.

Алынған нәтижерлерді қоғамдық пікірді бақылау, маркетингтік науқандар жүргізу, жаңалықтар оқиғаларын бағалау, талданған мәтіндер негізінде пікірлерді болжау, эмоционалды теріс қылықтарды анықтау үшін қолдануға болады. Сентимент талдау компанияларға немесе кез келген кәсіпкерлерге өзінің және бәсекелес фирмалардың тауарларының немесе қызметтерінің күшті және әлсіз жақтарын анықтау үшін, нарықтағы өнімнің позициясын жақсарту үшін маркетингтік іс-шаралар кешенін өзгертуге мүмкіндік береді. Аспект деңгейіндегі сентиментті талдау, әдетте, практикалық қолдану үшін қажет егжей-тегжейлі(нақты) деңгей болып табылады. Көптеген өнеркәсіптік жүйелер осыған негізделген. Зерттеу қоғамдастығында көп жұмыс жасалып, көптеген жүйелер құрылғанына қарамастан, мәселе әлі де шешіліп жатыр. Әрбір ішкі тапсырма өте қиын міндет болып қала береді.

**Әдебиеттер тізімі**

1. Bermingham A., Smeaton A. Classifying Sentiment in Microblogs: Is Brevity an Advantage? // Proceedings of the International Conference on Information and Knowledge Management (CIKM), 2010.

2. Hu M., Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, 2004, pp. 168–177.

3. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars.

Proceedings of the 21st International Conference on Computational Linguistics (Dialog-2015), 2015, pp. 46–57.

4. Jakob N., Gurevych I., Extracting opinion targets in a single-and cross-domain setting with conditional random fields, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1035-1045.

5. Liu B., Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 2012, pp. 1–167.

6. Mukherjee A, Liu B. Aspect extraction through semi-supervised modeling. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012, volume 1, pp. 339-348.

7. Popescu A. M., Nguyen B., Etzioni O. OPINE: Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP on interactive demonstrations, 2005, pp. 32–33.

8. Scaffidi C., Bierhoff K., Chang E., Felker M., Ng H., Jin C. Red Opal: product-feature scoring from reviews. Proceedings of the 8th ACM conference on Electronic commerce, 2007, pp. 182–191.

9. Wang S., Manning Ch. D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers – vol. 2, pp. 90–94, 2012.

10. Дудченко П. В. Метрики оценки классификаторов в задачах медицинской диагностики / П. В. Дудченко // Молодежь и современные информационные технологии : сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск. – Томск : Изд-во ТПУ, 2019. – [С. 164-165].

# ТҮРКІ ТІЛДЕРІНІҢ ЭЛЕКТРОНДЫ МӘТІНДІК ЖӘНЕ АУДИО КОРПУСТАРЫ

# ЭЛЕКТРОННЫЕ ТЕКСТОВЫЕ И АУДИО КОРПУСЫ ТЮРСКИХ ЯЗЫКОВ

# ELECTRONIC TEXT AND AUDIO CORPUS OF THE TURKIC LANGUAGES

***Мухамедшин Д.Р.***
*Институт прикладной семиотики Академии наук РТ,*
*Казань, Татарстан, Россия*
*damirmuh@gmail.com*

## НЕКОТОРЫЕ НОВЫЕ ПОИСКОВЫЕ И ИССЛЕДОВАТЕЛЬСКИЕ ВОЗМОЖНОСТИ СИСТЕМЫ УПРАВЛЕНИЯ КОРПУСНЫМИ ДАННЫМИ

**Аннотация.** Системы управления корпусными данными и поисковые системы, работающие с корпусными данными, являются одним из наиболее востребованных инструментов для исследования языков с использованием языковых электронных корпусов. Чем более обширный функционал предлагает система управления корпусными данными, тем более гибкими становятся возможности исследователей. В статье рассматриваются некоторые новые поисковые и исследовательские возможности системы управления корпусными данными.

**Ключевые слова:** система управления корпусом, корпусные данные, корпусная лингвистика, поисковая система.

***Mukhamedshin D.R.***
*Institute of Applied Semiotics of the AS of the RT,*
*Kazan, Tatarstan, Russia*
*damirmuh@gmail.com*

## SOME NEW SEARCH AND RESEARCH CAPABILITIES OF THE CORPUS DATA MANAGEMENT SYSTEM

**Abstract.** Corpus data management systems and search engines that work with corpus data are one of the most popular tools for language

research using language electronic corpora. The more extensive functionality the corpus data management system offers, the more flexible the possibilities of researchers become. The article discusses some new search and research capabilities of the corpus data management system.

**Keywords:** corpus management system, corpus data, corpus linguistics, search engine.

**Введение.** Система управления корпусными данными, разработанная автором, нацелена на работу с различными лингвистическими корпусами [Козлова, 2013]. Функционал, предлагаемый системой, включает в себя поиск лексических единиц, морфологический поиск, лексико-морфологический поиск, поиск синтаксических единиц, поиск n-грамм с учетом грамматики [Невзорова, 2015], поиск списков словоформ или лемм [Mukhamedshin, 2017], поиск с учетом метаданных документов, поиск с группировкой по документу, контексту, словоформе, лемме, морфологическим признакам. Также в системе реализован функционал формирования частотных списков на основе поискового функционала [Мухамедшин, 2021], доступен открытый API для быстрого обмена данными с другими системами. Поисковые технологии реализованы на базе современных общедоступных программных средств: система управления базой данных MariaDB [Bartholomew, 2014] и хранилище данных Redis [Carlson, 2013]. Благодаря концептуальной модели представления корпусных данных, поиск в корпусе производится менее, чем за 0,05 сек. в 98,71% случаев.

Разработанный корпус-менеджер ориентирован в первую очередь на поддержку электронных корпусов тюркских языков, что является весьма актуальным для активно развивающегося направления тюркской корпусной лингвистики. Функционал системы управления корпусными данными развивается и ежегодно пополняется новыми возможностями. В данной статье будут рассмотрены некоторые новые функциональные возможности системы.

**Просмотр коллекций по метаданным.** В поисковом движке, который предназначен для работы с основным корпусом татарского языка, реализована возможность поиска с учетом метаданных:

Название документа (поиск по части текста). Можно указать полное название документа или его часть. Результаты поиска будут включать только контексты, содержащиеся в найденных документах.

Автор документа (поиск по части текста). Можно указать полное имя автора или часть имени (например, если написание имени может меняться в зависимости от источника). Результаты поиска будут включать контексты, содержащиеся в документах искомого автора.

Год публикации. Можно указать год публикации документа. Расширенный синтаксис позволяет также искать документы по диапазонам или отдельным годам публикации документа, можно указывать диапазоны через дефис и перечислять года через запятую, например, «1991,1994-1996,2001-2008». Результаты поиска будут включать контексты, содержащиеся в документах, опубликованных в указанный период времени.

Источник (поиск по выбранному значению). Можно выбрать источник документа из предложенных системой.

Дополнительно (поиск по части текста в дополнительных метаданных). Можно указать дополнительные произвольные метаданные, например, категорию документа.

Пример поисковой выдачи при поиске с учетом года публикации представлен на Рисунке 1. В представленном случае год публикации указан с использованием расширенного синтаксиса: «2012,2014- 2016».

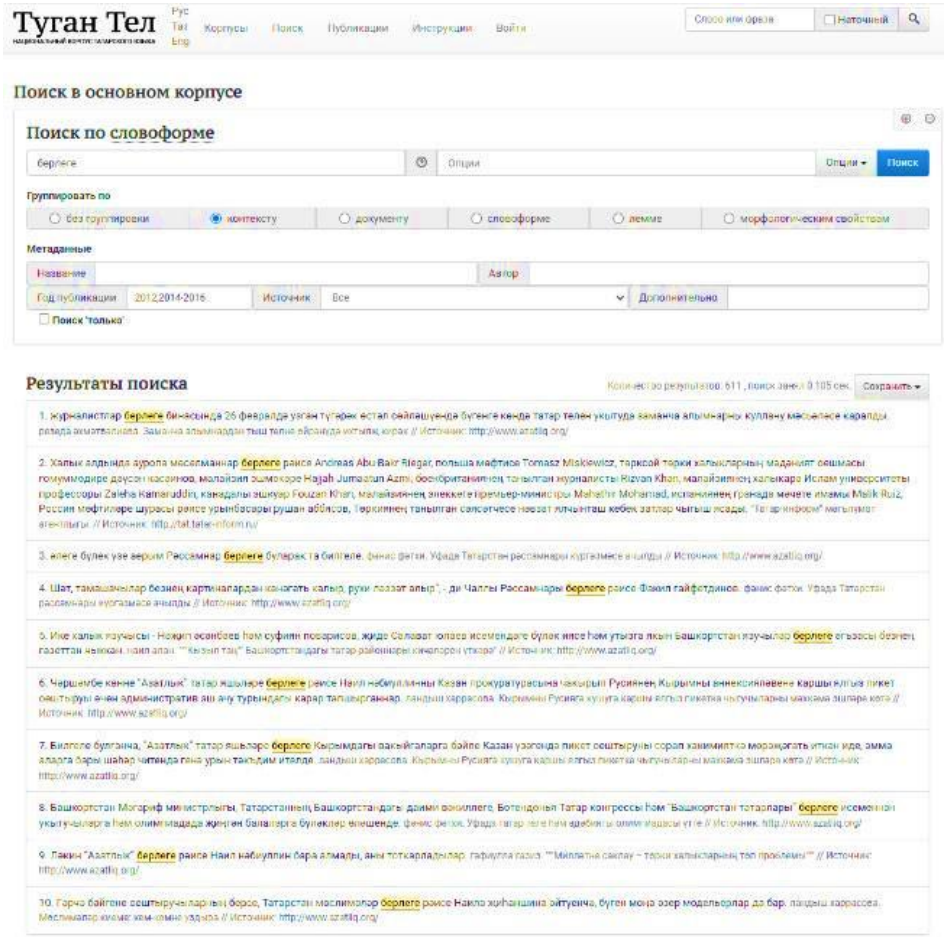



Рис. 1. Пример поиска с учетом года публикации
Fig. 1. Example of a search based on publication year

**Скачивание результатов поиска в виде файла.** В интерфейсе результатов поиска реализована возможность скачивания результатов в формате CSV [Shafranovich, 2005]. При нажатии кнопки «Сохранить» появляется возможность выбрать количество результатов для сохранения: 10, 20, 50, 100 результатов (Рисунок 2).

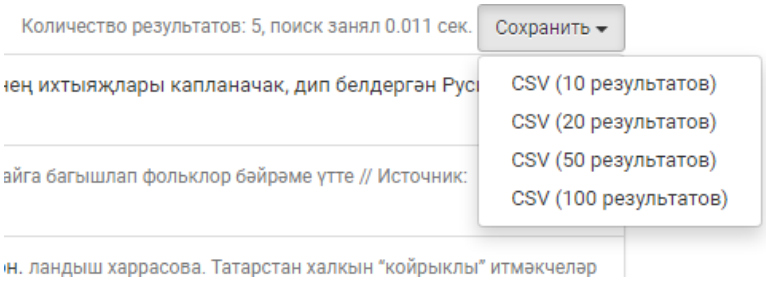



Рис. 2. Скачивание результатов поиска в виде файла
Fig. 2. Download search results as a file

Скачивание результатов поиска в формате CSV позволяет производить с результатами поиска произвольные операции группировки, сортировки, поиска, используя сторонние приложения. Например, с данным форматом могут работать такие приложения, как Microsoft Office Excel, LibreOffice Calc, Apache OpenOffice идругие.

**Поиск с группировками.** В рамках решения задач по улучшению функционала корпус-менеджера, реализован просмотр результатов поиска с группировкой по объектам базы данных электронного корпуса татарского языка «Туган Тел». В основном корпусе реализована возможность поиска с группировкой по:

Контексту (по умолчанию);

Документу;

Словоформе;

Лемме.

Также реализована возможность вывода результатов поиска без группировки.

При выборе группировки по контексту в интерфейсе поисковой выдачи выводятся уникальные контексты, содержащие словоформы, соответствующие поисковому запросу. Например, при запросе с использованием морфологической формулы [V], будет выведен список контекстов, в которых выделено первое вхождение словоформы, относящейся к части речи «глагол» (Рис. 3).

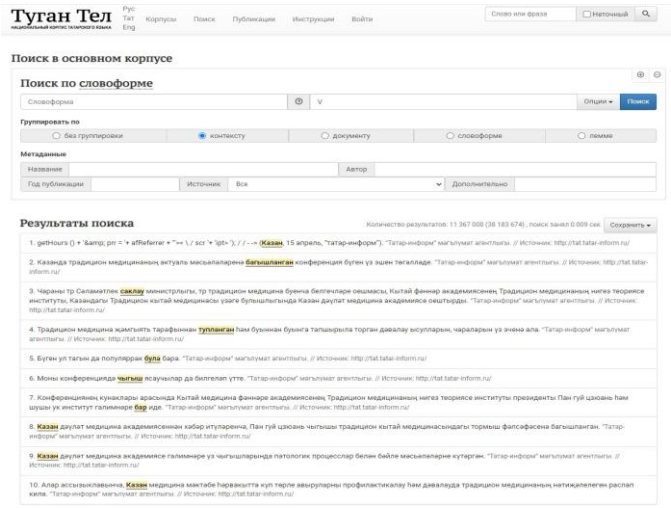



Рис. 3. Вывод результатов поиска с группировкой по контексту
Fig. 3. Displaying search results with grouping by context

Группировка по документу выводит все уникальные документы, в которых содержится хотя бы одно вхождение искомой словоформы. Например, при запросе малочастотных словоформ, можно получить список документов, в которых искомая словоформа встречается. Пример такого сценария использования поисковой системы корпус-менеджера показан на Рисунке 4 при поиске словоформы «ыжгырып» (тат. – «с хрипом»). При запросе выводится 46 уникальных документов с примерами контекстов, где встречается искомая словоформа.

Также одним из возможных сценариев использования группировки по документу является вывод списка документов при поиске по метаданным. Например, с появлением данного функционала, становится возможным получить список произведений автора Вакыйфа Нуриева, которые содержатся в электронном корпусе татарского языка «Туган Тел» (Рисунок 5).

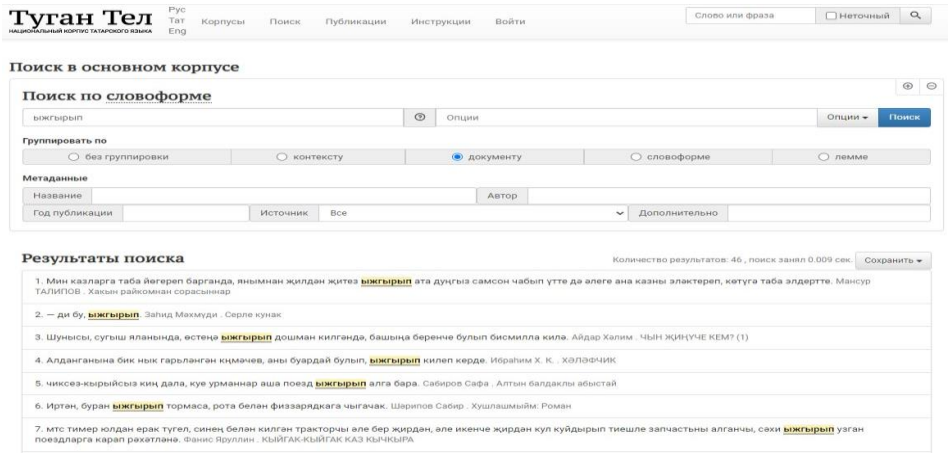



Рис. 4. Вывод результатов поиска с группировкой по документу
Fig. 4. Displaying search results with grouping by document

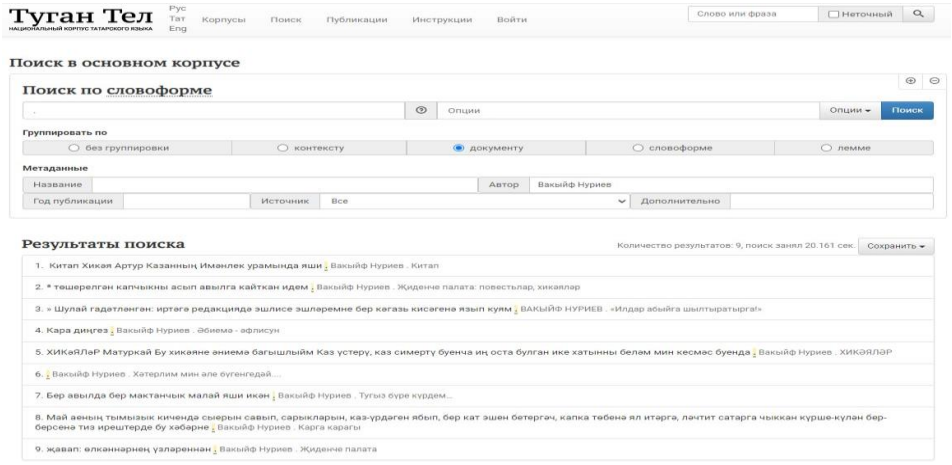



Рис. 5. Вывод результатов поиска по метаданным с группировкой по документу

Fig. 5. Displaying search results by metadata with grouping by document

Тип группировки «по словоформе» изменяет принцип вывода результатов в поисковой выдаче. Это происходит, потому что объекты типа «словоформа» являются частями контекстов. Таким образом, при группировке по словоформе в списке выводятся словоформы вместо контекстов. Наиболее простым сценарием использования группировки по словоформе является вывод всех словоформ, образованных от определенной леммы. Например, в результате поискового запроса по лемме «авыл» (тат. – «деревня») с группировкой по словоформе будет выведен список всех 147 словоформ, образованных от леммы «авыл», встречающихся в документах электронного корпуса татарского языка «Туган Тел» (Рисунок 6).

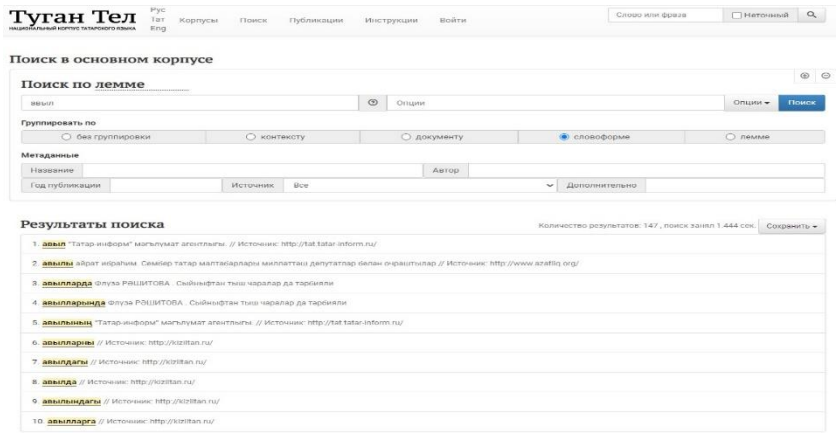



Рис. 6. Вывод результатов поиска по лемме с группировкой по словоформе

Fig. 6. Displaying search results by lemma with grouping by wordform

При поиске с группировкой по лемме ярким примером является сценарий поиска всех словообразовательных единиц определенной части речи. Так, например, при поисковом запросе [PN] с группировкой по лемме будут выведены контексты, содержащие вхождения словоформ части речи «местоимение», образованные от уникальных лемм. Пример на рисунке 7 показывает, что среди лемм, образующих местоимения, имеются «үз», «ал», «ул», «бар», «мо», «шушы», «алар», «һәрвакыт», «шу», «бу» и другие (всего 137 словообразовательных единиц).

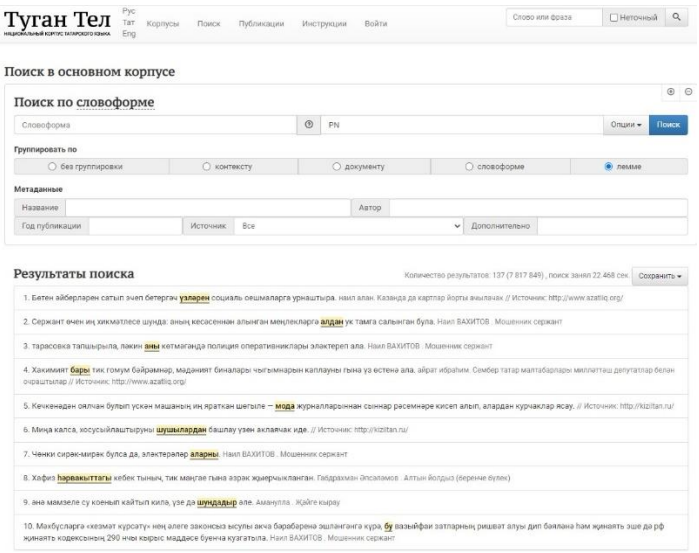



Рис. 7. Вывод результатов поиска с группировкой по лемме.

Fig. 7. Displaying search results with grouping by lemma.

**Заключение**

Представленные в данной статье возможности системы управления корпусными данными применяются в системе управления корпусом «Туган Тел» [Suleymanov, 2013]. Несмотря на расширение поискового функционала, время, необходимое для обработки и выполнения поискового запроса системой, не превышает 0,05 сек. в 98,71% случаев для лексического поиска, в 77,71% случаев для морфологического поиска и в 98,08% случаев для лексико-морфологического поиска. Во многом таких показателей удалось достичь благодаря использованию предложенных автором новых методов и технологий хранения и обработки корпусных данных, впервые примененных в системах управления корпусными данными. Новые поисковые и исследовательские возможности системы управления корпусными данными, реализованные в системе управления корпусом «Туган Тел», открывают исследователям языка более обширные возможности для исследования татарского языка. Общий

подход к решению задач в системе управления корпусными данными позволяет использовать разработанную систему не только для работы с электронным корпусом текстов на татарском языке, но и с корпусами других языков, без существенных изменений всистеме.

**Список литературы**

1. Козлова Н. В. Лингвистические корпуса: определение основных понятий и типология // Вестник Новосибирского государственного университета. Серия: Лингвистика имежкультурная коммуникация. – 2013. – Т. 11. – №. 1. – С. 79-89.

2. Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. Корпус- менеджер для тюркских языков: основная функциональность

3. // Труды международной конференции «Корпусная лингвистика - 2015». – СПб.: С.-Петербургский гос. Университет, филологический факультет, 2015. – С. 344-350.

4. Mukhamedshin D., Nevzorova O., Khusainov A. Complex Search Queries in the Corpus Management System // International Conference on Computational Collective Intelligence. – Springer, Cham, 2017. – С. 407-416.

5. Мухамедшин Д.Р. Система корпус-менеджер: реализация поискового функционала и частотных списков // Proceedings of the 9th International Conference on Turkic Languages Processing (TURKLANG-2021). (Tyva, September 21-23, 2021). - Tyva, 2021.

6. Bartholomew D. MariaDB cookbook. – Packt Publishing Ltd,

7. 2014.

8. Carlson J. Redis in action. – Simon and Schuster, 2013.

9. Shafranovich Y. Common format and MIME type for comma-separated values (CSV) files. – 2005.

10. Suleymanov D., Nevzorova, O., Gatiatullin, A., Gilmullin, R., Khakimov, B. National corpus of the Tatar language “Tugan Tel”: grammatical annotation and implementation. Procedia-Social and Behavioral Sciences, 95, 2013. С. 68-74.

*УДК 37:004*
***[1]Мадиева Г.Б., [2]Мансурова М.Е.***
*Казахский национальный университет им. Аль-Фараби*
*Алматы, Казахстан*
*[1]Gulmira.Madiyeva@kaznu.edu.kz, [2]Madina.Mansurova@kaznu.edu.kz*

# РАЗРАБОТКА УЧЕБНОГО МАТЕРИАЛА ПО КОРПУСНОЙ ЛИНГВИСТИКЕ: К ВОПРОСУ О ЯЗЫКОВОМ РЕСУРСЕ

**Аннотация.** Для подготовки специалистов по компьютерной и корпусной лингвистике необходимы специальные учебные материалы, которые могут обеспечить учебный процесс необходимой литературой и сократить временные затраты на ее поиск. Кроме того, этот материал позволяет систематизировать необходимые знания и предоставить актуальные практические задания, ориентированные на формирование как теоретических, так и практических компетенций. В статье анализируется предназначение учебного пособия по языковым ресурсам и терминологического словаря для формирования терминологического аппарата в целях его грамотного использования в последующей практике работы в сфере компьютерной и корпусной лингвистики.

**Ключевые слова:** корпус, национальный корпус, корпусная лингвистика, компьютерная лингвистика

*ӘОК 37:004*
***[1]Мәдиева Г.Б., [2]Мансұрова М.Е.***
*Әл-Фараби атындағы Қазақ ұлттық университеті*
*Алматы, Қазақстан*
*[1]Gulmira.Madiyeva@kaznu.edu.kz, [2]Madina.Mansurova@kaznu.edu.kz*

# КОРПУСТЫҚ ЛИНГВИСТИКА БОЙЫНША ОҚУ МАТЕРИАЛДАРЫН ӘЗІРЛЕУ: ТІЛДІК РЕСУРС МӘСЕЛЕ ТУРАЛЫ

**Аңдатпа.** Компьютерлік және корпустық лингвистика мамандарын даярлау үшін оқу үдерісін қажетті әдебиеттермен қамтамасыз ететін және оны іздеу уақытын қысқартатын арнайы оқу материалдары қажет. Сонымен қатар, бұл материал қажетті білімді жүйелеуге және теориялық және практикалық құзіреттіліктерді қалыптастыруға бағытталған нақты практикалық тапсырмаларды ұсынуға мүмкіндік береді. Мақалада компьютерлік және корпустық лингвистика саласындағы кейінгі тәжірибеде оны сауатты пайдалану мақсатында

терминологиялық аппаратты қалыптастыру үшін тіл ресурстары бойынша оқу құралы мен терминологиялық сөздіктің мақсаты талданады.

**Кілт сөздер:** оқу құралы, корпус, ұлттық корпус, корпустық лингвистика, компьютерлік лингвистика, тілдік ресурс.

*UDC 37:004*

***[1]Madiyeva G.B., [2]Mansurova M.Y.***
*Al-Farabi Kazakh National University*
*Almaty, Kazakhstan*
*[1]Gulmira.Madiyeva@kaznu.edu.kz, [2]Madina.Mansurova@kaznu.edu.kz*

## ON THE DEVELOPMENT OF EDUCATIONAL MATERIALS ON CORPUS LINGUISTICS

**Abstract.** To train specialists in computer and corpus linguistics, special educational materials are needed that can provide the educational process with the necessary literature and reduce the time spent on its search. In addition, this material allows you to systematize the necessary knowledge and provide relevant practical tasks focused on the formation of both theoretical and practical competencies. The article analyzes the purpose of a textbook on language resources and a terminological dictionary for the formation of a terminological apparatus for its competent use in subsequent practice in the field of computer and corpus linguistics.

**Keywords**: corpus, national corpus, corpus linguistics, computational linguistics

### Ведение

Современные знания и технологии требуют подготовки особой категории специалистов, владеющих компетенциями в информационной сфере. По этой причине создаются образовательные программы, включающие практико-ориентированные дисциплины на стыке информатики и лингвистики. Так, на филологическом факультете Казахского национального университета им. аль-Фараби с 2018 г. действует образоватлеьная программа «Компьютерная лингвистика», ориентированная на лингвистов, а на факультета информационных технологий работает такая же программа, но ориентированная на подготовку программистов. В бакалавриате и магистратуре успешно апробированы такие дисциплины, как Корпусная лингвистика и инновационные технологии в лингвистике, Корпусная лингвистика и компьютерные инструменты в обучении языку, Национальный корпус

казахского языка, Языковые ресурсы, Теория и практика корпусной лингвистики, Обработка речи, Методы поиска и извлечения информации, Статистические методы для обработки естественного языка, Анализ языка и др. В ходе изучения этих дисциплин обучающиеся анализируют проблемы компьютерной и корпусной лингвистики как активно развивающихся направлений современного языкознания и информационных технологий, необходимые в современный век цифровизации. Кардинальная цель этих дисциплин – сформировать у обучающихся знание основ компьютерной, в том числе корпусной, лингвистики. Им предлагается овладеть информационными технологиями в рамках подхода к изучению языка с использованием концепций и методов компьютерной лингвистики, корпусной лингвистики, в арсенал которой входит технология построения корпусов различных языков. В задачи дисциплин входит формирование различных компетенций: компетенций по формированию базовых основ компьютерных и корпусных технологий, а также компетенций по владению терминологическим аппаратом как компьютерной лингвистики, так и корпусной лингвистики и др. В результате обучающиеся приобретают навыки практической работы с различными типами корпусов (национальными, специализированными, специальными), использования языковых ресурсов, развиваются исследовательские умения определять теоретическое и практическое значение языковых корпусов и других языковых ресурсов в разных целях для лингвистических исследований и исследований в области компьютерной/корпусной лингвистики.

В рамках этих направлений важно понять, что представляет собой фактический материал, т.е. языковые ресурсы (ЯР) или Language Resources (LR), который используется и обрабатывается специалистами по компьютерной лингвистике. Функционально-прагматическая направленность современных исследований породила необходимость изучения языка/языков в различных коммуникативных ситуациях с целью презентации реального речевого действия, использования различных языковых единиц в определенном контексте. Эта довольно сложная проблема должна решаться при наличии объемного иллюстративного материала, который можно получить при помощи новых информационных технологий на основе Интернет-ресурсов, избегая рутинной работы.

В этом направлении предпринимают определенные шаги и казахстанские ученые. Данный факт обусловлен мощным развитием компьютерных технологий, их широким использованием практически

во всех сферах социума, в том числе в обучении и специальных исследованиях различных языков, в частности, казахского языка.

Информатизация различных областей жизнедеятельности человека обусловила массовую коммуникацию в пространстве Интернет (особая роль отводится пандемии Covid-19, возникшей в 2020 г., и активизации онлайн коммуникации), а также широкомасштабную цифровизацию всех сфер экономики, в том числе образования и науки. Этот факт обуславливает актульность специализированную подготовку кадров. В связи с этим возникает необходимость представить вниманию обучающихся системный характер программного обеспечения разных видов профессиональной деятельности, а также терминологического аппарата компьютерной лингвистики, в том числе терминологии корпусной лингвистики как составляющего компонента метаязыка этих направлений, что возможно на материале учебного пособия и учебного терминологического словаря.

Идея разработки учебного пособия возникла в ходе работы над проектом «Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities» («Разработка междисциплинарной магистерской программы по компьютерной лингвистике в университетах Центральной Азии», Эрасмус+ EAC-A03-2016 CLASS). Учебное пособие очень важно для подготовки не только специалистов по компьютерной или корпусной лингвистике, но и, что значимо, для подготовки лингвистов со знанием компьютерных технологий и их использованием для обработки языка и исследования лингвистических фактов в различных целях.

Фундаментальной теоретической основой для составления учебного пособия и словаря послужили многочисленные теоретические труды зарубежных и казахстанских авторов.

Разрабатываемое учебное пособие позволяет углубленно рассмотреть такие важные вопросы, как: основные понятия ЯР, обзор необходимых программных средств, корпус языка, национальный корпус, языковые модели, основные принципы инфраструктуры ЯР: документирование, разработка и доступность, обзор национальных корпусов текстов, программное обеспечение для управления корпусом, структура языкового корпуса, вопросы компьютерной лексикографии, структура и содержание одноязычных, двуязычных и многоязычных словарей, общий язык и Wordnet, базы данных как ЯР, онтология как ЯР, основные понятия онтологии. Помимо этого, особое внимание уделяется этическим и правовым аспектам использования языковых ресурсов.

Учебный материал в предлагаемом пособии размещен в соответствии с программным материалом учебной дисциплины «Языковые ресурсы». Содержание материала дифференцировано на 5 модулей:

1. Введение в языковые ресурсы (ЯР).

2. Языковые корпуса в качестве ЯР. Основные понятия корпусной лингвистики.

3. Лексикон как ЯР. Введение в лексикографию.

4. Базы данных как ЯР. Базы данных для письменных и устных языковых данных.

5. Онтология как ЯР. Основные понятия онтологии.

Все модули характеризуются логической последовательностью и взаимозависимостью, формируют необходимые компетенции для использования языковых ресурсов в практической деятельности.

Благодаря учебному пособию обучающиеся получают системные компетенции по знанию языковых ресурсов для дальнейшего практического их применения в работе над корпусами различных языков, знакомятся с их структурой, сравнивают и используют эти компетенции на практике.

Обучающиеся изучают такие параметры, как:

1. морфологический анализ, выполняемый на основе автоматического морфологического анализатора, способного обрабатывать объемные массивы, что невозможно сделать вручную. У ручной разметки, безусловно, имеются свои плюсы, но разметить 100 и более миллионов слов руками, рутинным способом, принципиально невозможно, нужна автоматическая морфология;

2. возможность неограниченного онлайн-поиска во всем корпусе, а не только в небольшом его фрагменте;

3. интерфейс на разных языках, в частности, казахском, английском, русском языках, что позволяет различным пользователям с разной языковой подготовкой работать с корпусами;

4. перевод всех лексем на русский или английский язык при морфологическом анализе, поскольку эти языки являются международными;

5. возможность поиска не только по лемме и частям речи, но и по другим грамматическим характеристикам (падеж, число, залог и т.п.);

6. статистический (показывается количество найденных результатов и документов);

7. возможность выбора подкорпуса (например, возможность целенаправленно искать какое-либо слово только в художественной литературе или периодической печати, или только в диалекте).

У обучающихся формируется понимание того, что морфологический разбор в корпусе нужен в первую очередь не для того, чтобы показать пользователю, какие характеристики имеет та или иная словоформа, а для поиска по этим характеристикам. Вероятностные подходы к морфологическому анализу (включая вероятностное снятие омонимии после морфологического анализа), безусловно, имеют свои положительные стороны, однако их точность даже в языках с бедной морфологией никогда не поднимается выше 97% (в морфологически богатом казахском языке это число, скорее всего, будет ниже). Если для практических целей такой результат вполне приемлем, то для многих видов лингвистического исследования этого недостаточно.

Помимо этого, обучающиеся должны понимать, что качественно разработанный корпус дает возможность поиска любой сложности с использованием словоформы или её части, леммы, грамматических характеристик или их комбинаций (с использованием логических функций), пунктуации и позиции словоформы в предложении. Следовательно, от качества их работы напрямую зависит качество корпусов. Так, например, в интерфейсе корпуса Алматинского корпуса казахского языка (АККЯ) возможен поиск не только слов, но и фраз, если ввести искомую фразу в поле в нижней правой части окна ("быстрый поиск"), например, см. рисунок 1 [Алматинский корпус казахского языка].

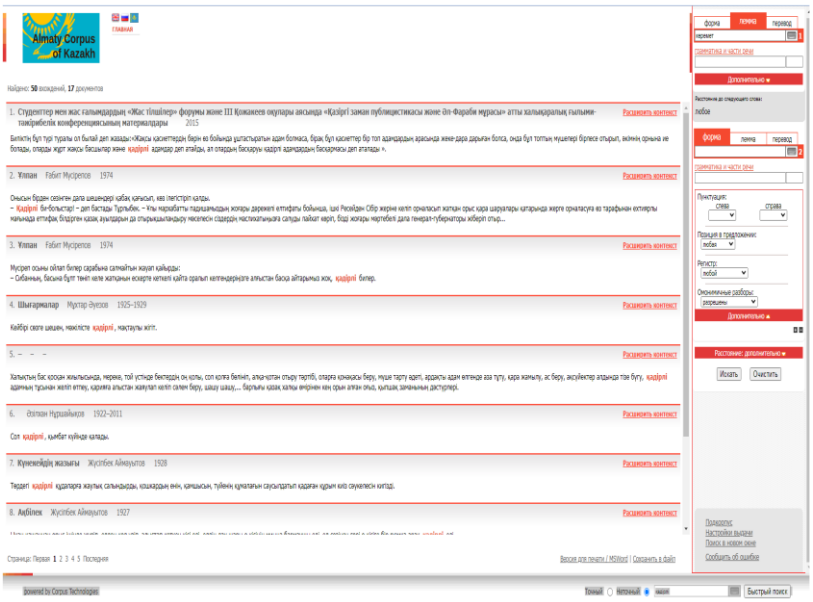



Рисунок 1. Быстрый поиск в Алматинском корпусе казахского языка
Figure 1. Quick search in the Almaty corpus of the Kazakh language

Кроме того, поисковый интерфейс позволяет искать одновременно любое количество слов, задавая расстояния между ними (в том числе диапазонами), что позволяет искать не только фиксированные фразы, но и сложные конструкции, некоторые части которых не фиксированы,

Система поиска в АККЯ позволяет искать не только фразы, но и целые контексты со ссылкой на источник, что во много раз повышает репрезентативность и достоверность конкретной реализации слова в речи, а это существенно для различного рода исследований.

Все эти навыки обучающиеся получают на основе изучения материала учебного пособия и апробируются практически на основе изучения корпусов как одного из видов ЯР и продукта корпусной лингвистики. Как отмечает В.М. Плунгян, корпус «это очень эффективный и полезный инструмент (которым могут пользоваться далеко не только узкие специалисты), особенно в том случае, когда корпус является большим по объему и полным по охвату материала, т.е. представляет собой так называемый Национальный корпус языка <…> это не просто собрание текстов наподобие электронной библиотеки, это текстовая база данных, которая дает возможность быстрого и эффективного представления различной информации о слове и его реализации в контексте на материале текстов различного жанра, стиля, хронологического периода, определенного автора <…>». [Плунгян, 2005]

Усвоение знаний и приобретение навыков в искомых направлениях активизирует необходимость системного изучения терминологии компьютерной/корпусной лингвистики. Поэтому к учебному пособию прилагается краткий словарь, который основан на принципе системной подачи базовых терминов корпусной лингвистики, необходимый для использования в исследовательской работе. Поскольку это учебный словарь, то термины в нем расположены с традиционными принципами лексикографии терминов в обучающих словарях. На основе этого словаря обучающиеся узнают его дефиницию, системно формируют терминологический аппарат.

**Список литературы**

1. Алматинский корпус казахского языка / http://web-corpora.net/KazakhCorpus/search/?interface_language=ru

2. Плунгян В.А. Зачем мы делаем Национальный корпус русского языка? // Отечественные записки. 2005. № 2, 296-308.// http://www.strana-oz.ru/2005/2/

3. Национальный корпус русского языкаhttp://www.ruscorpora.ru/corpora-intro. html

4. https://ru.wikipedia.org/wiki

5. http://til.gov.kz/wps/portal

6. Добрушина Н.Р. Как использовать Национальный корпус русского языка в образовании? // Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005, 308-329.

7. Рахилина Е.В. Корпус как творческий проект // Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 7-26.

*ӘОК 781.1*
***[1]Сакенова Ж.Ж., [2]Маткаримов Б.Т.***
*[1 2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]gioxkkzs@gmail.com , [2]matkarimov_bt@enu.kz*

# ТҮРКІ ТІЛДЕРІНДЕГІ ЖӘНЕ ТҮРКІ МУЗЫКАСЫНДАҒЫ ӘН ЖАЗБАЛАРЫНЫҢ МУЗЫКАЛЫҚ-КОРПУСЫ

**Аңдатпа.** Бұл мақалада түркі халықтарының музыкалық дәстүрлерінің байлығын зерттеу, соның ішінде музыкалық шығармаларды, әсіресе ән айтуды талдау үшін жаңа музыкалық корпустар мен есептеу технологияларын дамыту бойынша жүргізіліп жатқан жұмыстар сипатталған.

Бүкіл әлемде дыбыс технологияларын әзірлеуге үлкен қызығушылықтың бар екенін назарға ала отырып, біз түркі халықтарының музыкасын компьютерлік талдаудың өңірлік, сондай-ақ жаһандық мәнмәтінде жаңа технологияларын жинақтау және әзірлеу, сондай-ақ түркі музыкасының шығу тегі мен эволюциясы туралы қазіргі түсінігімізді кеңейту үшін музыкалық корпусты ұсынамыз. Бұл тақырып түркі халықтарының мәдени мұрасын сақтау және дамыту қажеттілігі тұрғысынан өзекті. Орталық Азияда тұратын түркі халықтарының (қазақтар, қырғыздар, өзбектер, түрікмендер, қарақалпақтар) музыкалық дәстүрлеріне баса назар аударылады.

**Түйінді сөздер:** мәдени мұра, корпус, музыкалық корпус, цифрлық технологиялар, түркі музыкасы.

*UDC 781.1*
***[1]Сакенова Ж.Ж.,[2]Маткаримов Б.Т.***
*[1 2]Евразийский национальный университет имени Л.Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]gioxkkzs@gmail.com, [2]matkarimov_bt@enu.kz*

# МУЗЫКАЛЬНЫЙ КОРПУС ТЮРКСКОЙ МУЗЫКИ И ПЕСЕННЫХ ЗАПИСЕЙ НА ТЮРКСКИХ ЯЗЫКАХ

**Аннотация.** В данной статье описывается ведущаяся работа по изучению богатства музыкальных традиций тюркских народов, включая разработку новых музыкальных корпусов и вычислительных технологий для анализа музыкальных произведений, в особенности пения.

Принимая во внимание наличие большого интереса к разработке звуковых технологий во всем мире, мы предлагаем музыкальный корпус для коллекции и разработки новых технологий компьютерного анализа музыки тюркских народов, как в региональном, так и в глобальном контексте, а также расширения существующего понимания происхождения и эволюции тюркской музыки. Данная тема актуальна с точки зрения необходимости сохранения и развития культурного наследия тюркских народов. Основное внимание уделяется музыкальным традициям тюркских народов, проживающих в Центральной Азии (казахи, киргизы, узбеки, туркмены, каракалпаки).

**Ключевые слова:** культурное наследие, тюркская музыка, музыкальный корпус, вычислительные методы анализа музыки и пения.

*UDC 781.1*

***[1]Sakenova Zh.Zh.,[2]Matkarimov B.T.***
*[1 2] L.N.Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]gioxkkzs@gmail.com , [2]matkarimov_bt@enu.kz*

## MUSICAL CORPORA OF SONG RECORDINGS IN TURKIC LANGUAGES AND TURKIC MUSIC

**Abstract.** This article describes the ongoing work to study the richness of Turkic people's musical traditions, including the development of new music corpora and computing technologies for the analysis of music, and especially singing.

Taking into account the great interest in the development of sound technologies around the world, we offer a musical corpus for the collection and development of new technologies for computational analysis of Turkic music, both in a regional and global context, as well as expanding the existing understanding of the origin and evolution of Turkic music. This topic is relevant from the point of view of the need to preserve and develop the cultural heritage of the Turkic peoples. The main attention is paid to the musical traditions of the Turkic peoples living in Central Asia (Kazakhs, Kirghiz, Uzbeks, Turkmens, Karakalpaks).

**Keywords:** cultural heritage, Turkic music, music corpora, computational methods for analyzing music and singing.

### 1. Кіріспе

Бүгінгі таңда әлемде цифрлы технологиялардың кең ауқымда дамуына байланысты, цифрлық ақпараттар көлемі еселеп ұлғаюда, яғни

желіде миллиондаған ақпараттың ішінде өзіңе қажетті ақпаратты табу өте оңай, алайда сол миллиондаған ақпараттың ішінде түркі халықтарының музыкасының дыбыстық әлеміне жүргізілген зерттеулер жоқтың қасы. Бұл өте өкінішті жағдай, себебі Азия құрлығының орта бөлігіндегі байтақ аумақты мекендеген түркі халықтарының музыкасы-ұзақ тарихы мен бай дәстүрі бар бірегей құбылыс. Түркілер музыкасының дыбыстық әлемі әр түрлі және ерекшелігі, ол дыбыстардың кең спектрімен ерекшеленеді: төмен, обертонды бай, кеуде қуысы сыңғырлаған, фальцетті үндері жоғары және кернеулі. Әр түрлі «тығыздық» дәрежесі, биіктіктегі қозғалғыштығы бар, олар «ұлттық бірегейлік» ұғымынан бөлінбейді. Олардың тембрлік ерекшелігі, сондай-ақ жергілікті музыканттардың сезімтал және тазартылған есту қабілетінің болуы түркілердің музыкасының дыбыстық әлемі сияқты, әзірге арнайы зерттеу объектісіне айналмаған түпнұсқа дыбыстық жүйенің болуын көрсетеді.

Дәстүрлі музыкалық мәдениеттердің дамуының қазіргі кезеңінде социализм дәуірінің мәдени саясатының идеологиялық көзқарастарымен құрылған теория мен практика арасындағы тұжырымдамалық сәйкессіздікті жоюдың маңыздылығы барған сайын айқын бола түсуде. Дәстүрді «ауызша» немесе «халықтық» деп анықтау музыкалық құрылымдардың қалыптасуы мен дамуы туралы өзіндік тұжырымдаманың жоқтығын білдірмейді. Дәстүрлі мәдениетті «кәсібилендіру» тәжірибесінің нәтижесі тек дәстүрлі музыкалық аспаптардың ғана емес, сонымен қатар дәстүрлі музыкалық және дыбыстық жүйелердің бүкіл эстетикалық және нормативтік негізінің мутациясы екенін түсіну керек. Ал дәстүрлі мәдениеттер түбегейлі ауызша бола отырып, қалыптасу және даму процесінде әлі күнге дейін зерттелмеген білім берудің өзіндік тетіктері мен әдістерін жасады. Қазіргі уақытта ауызша есту типіндегі дәстүрлі мәдениеттердің тасымалдаушыларының табиғи өмірден кетуіне байланысты сөзсіз сандық төмендеуі байқалады, сондықтан әсер етудің әлсіреуі және олардың мәдениеттегі рөлінің төмендеуі байқалады. Дәстүрлі құндылық бағдарларында тәрбиеленбеген музыкалық мектептер, училищелер, консерваториялар мен университеттер, мәдениет және өнер академиялары түлектерінің саны айтарлықтай өскені байқалады. Диплом алған мамандардың санының өсуі академиялық (еуропалық) терминологияны белсенді түрде енгізе бастайды және музыкалық практикаға дәстүрлі мәдениетке тән емес бағалау критерийлерін енгізеді. Бұл дәстүрлі музыкалық мәдениеттер бойынша отандық және шетелдік ғалымдардың елеулі ғылыми еңбектері жеткілікті болған жағдайда. Дегенмен, Ғылыми зерттеулердің нәтижелері әзірше білім

беру жүйесіне дәстүрлі музыканттарын оқыту үшін пайдаланылатын оқу бағдарламаларын, оқу-әдістемелік құралдарды әзірлеу кезінде тартылмаған [1, 2].

Бұл білім ұрпақтан-ұрпаққа беріліп отырды және жазбаша түрде көрсетілмеген нақты заңдылықтарды білу міндеті дәстүрдің өзі емес, тек музыкатану ғылымының проблемасы болып табылмайды, сонымен қатар жаһандану заманында, цифрлық технологиялар дамыған заманда, цифрлық ақпараттандыру ғылымадарыныңда басты мәселелерінің бірі болып табылуы қажет, себебі бұл ата-балаларымыздан қалған ерекшеліктерді біз мәңгілік өшпейтін ақпарат күйіне айналдыруымыз қажет. Өзіміздің дәстүрлі музыкалық мәдениетіміздің, басқа батыс елдерінен келген мәдениетпен араласып мутацияға ұшырамауына жол бермеу түркі халықтары аасындағы маңызды мәселелердің бірі.

Интернетте көптеген қазақ халық әндерінің ноталары жоқ, мысалы, «Елім-ай».

Халық музыкасының таралуы мен қолданылуы авторлық құқықпен шектелмейді, сонымен бірге жазылған әлемдік музыканың көп бөлігі авторлық құқық пен басқа да этикалық мәселелерге байланысты әлі күнге дейін қол жетімді емес.

Әр музыкалық мәдениеттің аудио-музыкалық жинақтары және ілеспе ақпарат цифрлық музыка жобасының негізін құрайды. Олар біздің зерттеу корпусымызды құрайды. Бірақ жақсы зерттеу корпусы дегеніміз не?

Зерттеу корпустарының дизайны зерттеу мәселесі болып табылады. Бұл корпустар барлық деректерге негізделген зерттеулерге сәйкес келуі керек, бірақ информатика тұрғысынан корпоративті дизайн принциптері туралы өте аз жазылған, сөйлеуді өңдеу және тіл білімі сияқты салаларда стандартты корпустарды қолдануға бірнеше сілтемелер бар, бірақ негізінен ғажайыптар туралы ештеңе жоқ .

Мәтіндік және дыбыстық корпустың дамуы сөйлеу технологиясындағы ең үлкен проблемалардың бірі болып табылады. Американдық Ағылшын сияқты тілдер үшін TIM IT, Wall Street Journal және Switchboard сияқты фонетикалық бай және үлкен корпустар қолданылады. Қазіргі таңда, түркі халықтарының зерттеушілерінің сөйлеуді және ән айтуды өңдеудегі ең үлкен мәселелерінің бірі фонетикалық бай мәліметтер базасының болмауы. Сол себепті жобаның мақсаты түркі халықтарының музыкаасына морфологиялық және синтаксистік аннотацияланған аудио корпусын жасау болып табылады [3].

Зерттеу корпустары идеясын сынақ корпустары идеясымен шатастырмауымыз керек (сонымен қатар тест жинақтары деп те аталады). Зерттеу корпустары - бұл білімді кеңейту мақсатында

эксперименттер жүргізу үшін қолданылатын сенімді мәліметтер жиынтығы. Сынақ корпустары-бұл негізгі шындықтар, эксперимент-терде қолданылатын құралдарды сынау, бағалау және калибрлеу үшін қолданылатын сенімді немесе жалған мәліметтер жиынтығы. Мұнда біз зерттеу корпусымен айналысамыз, осылайша зерттеу жұмысын орындау үшін белгілі бір музыкалық мәдениеттің мәнін көрсететін корпустар жасаймыз.

Жиналған деректер түрлері - бұл аудио жазбалар мен редакторлық метадеректер, олар бізде бар материалдар туралы сипаттамалық ақпаратпен, ал кейбір жағдайларда музыкалық ноталармен және ән мәтіндерімен толықтырылады. Корпустағы негізгі бөлім-аудио жазба және онымен бірге жүретін ақпараттық элементтер жиынтығы. Корпусты құру кезінде біз назарға алған негізгі критерийлер: мақсаты, қол жетімділігі, толықтығы, сапасы және қайта пайдалану мүмкіндігі.

**Мақсаты:** корпусты дамытудағы алғашқы қадам-шешілуі керек зерттеу мәселесін және қолданылатын зерттеу тәсілін анықтау. Біздің жағдайда біз негізінен әуен мен ырғаққа байланысты аудио жазбалардан музыкалық маңызды элементтерді алуға болатын әдістемелерді жасағымыз келеді. Тәсілдер сигналдарды өңдеу және машиналық оқыту әдістеріне негізделген; осылайша, корпус осы мақсатқа сәйкес келуі керек.

**Қамту:** Корпус зерттелетін барлық тұжырымдамалардың өкілі болып табылатын мәліметтерді қамтуы керек және біздің сандық көзқарасымызды ескере отырып, статистикалық маңызды болуы үшін әр дананың үлгілері жеткілікті болуы керек. Зерттеу үшін бізге аудио жазбалар, сонымен қатар әр музыкалық мәдениеттегі әр түрлі әуендер мен ырғақтарды қамтитын тиісті ілеспе ақпарат қажет.

**Толықтығы:** әр жағдайда әр аудио жазба деректер өрістерінің жиынтығымен толықтырылады және толықтық идеясы толтырылған өрістердің пайызымен, яғни корпустың қаншалықты толық екендігімен байланысты. Біздің корпустар үшін бұл негізінен редакторлық метадеректер мен әр аудио жазбамен бірге сипаттамалық ақпараттың толықтығына қатысты [4].

**Сапасы:** деректер сапалы болуы керек. Дыбыс жақсы жазылуы керек, ал ілеспе ақпарат дәл болуы керек. Мүмкін болған кезде біз жақсы дайындалған жазбаларды қолдандық, ал қосымша ақпарат сенімді көздерден алынды және сарапшылар тексерді.

Қайта пайдалану мүмкіндігі: зерттеу нәтижелері жаңғыртылуы керек, яғни корпусты зерттеу қоғамдастығы қолдана алады. Біздің жағдайда біз қазірдің өзінде жарамды немесе біздің қажеттіліктерімізге

бейімделуі мүмкін нақты ашық репозиторийлерді қолдануға назар аударды.

Бұл мақалада біз қазақ-түрік тіліндегі әндерді орындалуын зерттеуге арналған жаңа музыкалық корпус құралдарын жасау бойынша жұмысымызды ұсынамыз.

**2. Корпус дегеніміз не? Қандай түрлері бар?**

Корпус-бұл белгілі бір мақсатпен жасалған табиғи тілдің жиынтығы, мәтін және сөйлеудің немесе белгілердің транскрипциясы. Көптеген қол жетімді корпустар тек мәтіндік болса да, мультимодальдық корпустар, соның ішінде ымдау тілінің корпустары көбейіп келеді. Мультимодальдық корпус - бұл бірнеше сенсорлық модальділікті немесе бірнеше өндірістік модальділікті қолданатын тілдік және коммуникативті материалдардың компьютерлік жиынтығы онда сенсорлық модальділікке көру, есту, жанасу, иіс немесе дәм, сонымен қатар сөйлеу, белгілер, көрініс, дене пішіні сияқты өндірістік модальділік кіреді және қимылдар. Яғни, мультимодальды корпус - бұл адамдар арасындағы бейне және аудио жазбалардың жиынтығы. Бірақ кез - келген аудио және видео материалдар жиынтығы корпус емес. Біріншіден, аудиовизуалды материалдар мұқият таңдалуы керек, ал мазмұны метадеректер көмегімен сипатталуы керек. Екіншіден, материал стандартталған форматта транскрипциялар мен аннотация-лармен талданып, сипатталуы керек. Ең дұрысы, корпус-бұл кездейсоқ жиналған мәліметтер жиынтығын емес, мұқият іріктеу арқылы тілдің (немесе тілдің) өкілі болу үшін жасалған тілдік өнімдердің үлгілері. Корпустың қаншалықты өкілі, нақты зерттеу мәселесін ескере отырып, корпустың тепе-теңдігі мен үлгісімен анықталады. Біз өкілдікті сұраққа жауап ретінде қарастыра аламыз: бұл корпус тілдің тиісті жақтарын қаншалықты жақсы сипаттайды? Ортақ корпусты құру үшін барлық жастағы ерлер де, әйелдер де осы тілде сөйлейтін аймақтың әртүрлі бөліктерінен дайындалған тілдік үлгілерді және т. б. қосу керек [5, 6].

Өкілдік, тепе-теңдік және іріктеуге қатысты бірдей принциптер мәтіндік және мультимодальдық корпустарға да қатысты, және одан үйренуге болатын корпусты дамыту бойынша үлкен жұмыс бар. Деректерді таңдаудың әртүрлі тәсілдері бар. Мұның бір жолы-тілге «өнім» ретінде назар аудару және диалог немесе монолог немесе сценарий немесе риясыз сөйлеу сияқты тілдік материалдардың әртүрлі түрлерін сынап көру.

Тағы бір тәсілі - тілдің «өндірушісіне» назар аудару және жас, жыныс, әлеуметтік тап, ана /екінші тіл, білім деңгейі, мамандық және аймақтық шығу тегі сияқты спикердің сипаттамаларына негізделген

ақпарат берушілерді таңдау. Кейбір жағдайларда, мысалы, белгілі бір жұмыс орнындағы әріптестер арасындағы қарым-қатынасты жазу кезінде ақпарат берушілер таңдалады, өйткені олар спикердің сипаттамаларына емес, сол жерде жұмыс істейді. Мұндай корпустар жалпыға қарағанда мамандандырылған, бірақ деректерді талдауда динамиктердің сипаттамалары әлі де маңызды [6].

Музыкалық корпустың сөйлеу корпусынан ұқсастықтары мен айтарлықтай айырмашылықтары бар. Музыкалық және сөйлеу корпусының жалпы сипаттамалары:

- сандық аудиожазбалар ақпараттың негізгі тасымалдаушысы ретінде,
- бір мәтінге немесе музыкалық шығармаға арналған көптеген аудио жазбалардың болуы,
- аудиожазба мазмұнының аналитикалық және мағыналық сипаттамасы, көркемдік нюанстар мен екпіндер және т. б.,
- фразаларды, сөздерді, фонемаларды немесе ноталарды бөлектеу үшін уақыт белгілері бойынша аудио жазбаларды аннотациялау қажеттілігі,
- аудио жазбаларды талдау үшін сандық сигналдарды өңдеу әдістерін кеңінен қолдану.

Сөйлеу корпусынан айырмашылығы, музыкалық корпустың өзіндік ерекшелігі бар:

- дауысты дыбыстарды, соның ішінде дауысты дыбыстарды айту,
- бір аудио жазбада көптеген дауыстардың болуы, мысалы, әннің аспаптық сүйемелдеуі,
- музыкалық шығарманың көлемі, қарқыны, тоналдылығы, құрылымы, дауыс әуендері, үйлесімділік аккордтары және т.б. туралы деректерді қамтитын ноталық жазбаның болуы.

Мультимодальды корпустар жағдайында жазбалар студияда немесе нақты әлемде натуралистік жағдайда жасалған ба, жоқ па, маңызды аспект болып табылады. Сондай-ақ, кез-келген әрекетті байқамай қарау (мысалы, ата-ана мен бала үйде ойыншықтар жиынтығымен ойнайды) және нұсқауларға сәйкес тапсырманы орындайтын адамдардың жазбалары арасында айырмашылық бар (мысалы, екі ересек адам фильмді зертханалық жағдайда талқылайды).

## 3. Музыкалық корпус және аудио файлды спектрлік талдау

Түркі халықтарының музыкасына арналған жаңа музыкалық корпус технологияларын жасау дәл таңбаланған және аннотацияланған деректерді қажет етеді.

DS/ML саласындағы соңғы жылдардағы негізгі бағыт-бұл NLP, әсіресе трансформатор архитектурасына негізделген нейрондық

желілерді пайдалану перспективалары [7]. Олар дауыстық көмекші жүйелерде де қолданылады, ал дауыстық көмекшілер біздің өмірімізге берік енеді. Алайда, дауыстық көмекшілердің жетістігінің маңызды құрамдас бөлігі-бұл «дауыстық», яғни оларға жүгіну дауыс арқылы жүзеге асырылады, яғни дыбыс. Көбінесе дыбыстық сигналмен жұмыс дыбысты да, спектрограмманың бейнесін де талдау арқылы жасалады.

Дәстүрлі түрде, сандық дыбыстық жазбада Аудио трек дыбыстық толқынның пішінін (waveform), яғни дыбыс амплитудасының уақытқа тәуелділігін көрсететін осциллограмма түрінде ұсынылады. Бұл қойылым тәжірибелі дыбыс инженері үшін өте айқын: осциллограмма дыбыстағы негізгі оқиғаларды, мысалы, дыбыс деңгейінің өзгеруін, шығарманың бөліктері арасындағы үзілістерді және көбінесе аспаптың жеке жазбасындағы жеке жазбаларды көруге мүмкіндік береді. Бірақ осциллограммада бірнеше аспаптардың бір уақытта дыбысы «араласады» және сигналды визуалды талдау қиынға соғады. Алайда, біздің құлағымыз кішкентай ансамбльдегі жеке аспаптарды оңай ажыратады. Бұл қалай болады?

Аса күрделі дыбыстық тербеліс құлаққа тигенде, ол бірқатар есту жүйелері көмегімен ұлулар деп аталатын мүшеге беріледі. Ұлулар-спиральға бұралған серпімді түтік. Ұлулардың қалыңдығы мен қаттылығы шетінен спиральдың ортасына қарай біртіндеп өзгереді. Күрделі тербеліс ұлулардың шетіне түскенде, бұл ұлулардың әртүрлі бөліктерінің кері тербелістерін тудырады. Бұл жағдайда ұлулардың әр бөлігінің резонанстық жиілігі әртүрлі. Осылайша, ұлулар күрделі дыбыстық тербелісті жеке жиілік компоненттеріне бөледі. Есту нервтерінің жеке топтары ұлулардың әр бөлігіне сәйкес келеді, олар ұлулардың миға ауытқуы туралы ақпарат береді [8]. Нәтижесінде жиілікте ыдыраған дыбыс туралы ақпарат миға түседі және адам жоғары дыбыстарды төмен дыбыстардан оңай ажыратады. Сонымен қатар, біз көп ұзамай көретін болсақ, дыбыстың жиіліктерге бөлінуі полифониялық жазбадағы жеке құралдарды ажыратуға көмектеседі [9], бұл редакциялау қабілетін едәуір кеңейтеді.

Спектрлік талдау бұл музыкалық файлдағы деректерді көрсетудің көрнекі тәсілі. Әр музыкалық нотаға белгілі бір жиілік сәйкес келеді: төмен ноталар төмен жиіліктерге, ал жоғары ноталар жоғары жиіліктерге сәйкес келеді. Барлық жиіліктер музыкалық файлдың ұзақтығына қатысты барлық жиіліктердің графикалық көрінісі болып табылатын спектрлік диаграммада (спектрограмма) көрсетіледі. Жиіліктер герцпен (Гц) және килогерцпен (1000 Гц) өлшенеді. Адамның дыбыстық сезімталдығының диапазоны 20 герцтен 20 килогерцке дейін (20000 герц).

Спектрограммалар файлдың барлық деректерін көрсететіндіктен, әннің кодталғанын немесе жоқтығын анықтау қажет болса, олар жақсы көмек болады. Әр аудио файлда жиіліктің салыстырмалы стандарты бар.

Аудио файлдарға спектрлік талдау жүргізу үшін Adobe Audition ((Windows немесе Mac OS үшін), Audacity (Windows, Mac OS, Linux) немесе SoX (Windows, Mac OS, Linux, бірақ тек пәрмен жолынан) қолдануға болады. Жобадағы барлық спектрограммалар Audacity (Windows, Mac OS, Linux) бағдарламасы көмегімен қаралды.

Алынған спектрлік деректерді қолдана отырып, ән синтезін жасау жоспарлануда [10]. Спектрлік талдау үшін барлық қажетті скрипттер MathWorks MATLAB ортасында дайындалады.

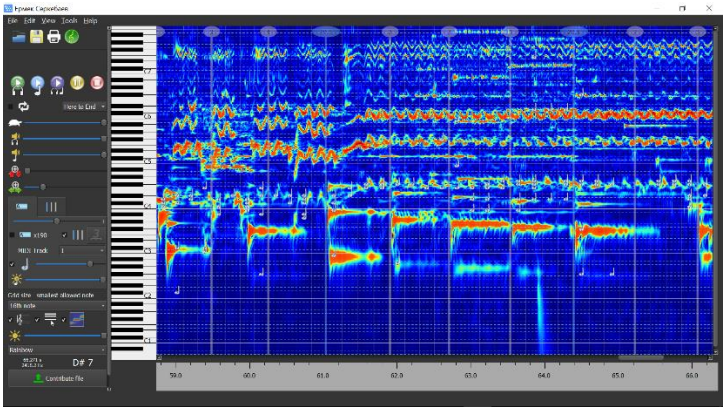
Біз әр түрлі ашық және жеке көздерден музыкалық аудиожазбалар жинауды жоспарлап отырмыз: 1) A Capella әнін жазамыз, 2) Әр түрлі деректер көздерінен түркі музыкасының цифрландырылған тарихи жазбалары, 3) VocalSet [11], The Million Song [12], 4) vocalset [11] сияқты қоғамдық мәліметтер жиынтығы. 5) ноталардан немесе MIDI файлдарынан жасалған. Сондай-ақ, Smithsonian Folkways Recordings сияқты көптеген коммерциялық көздер бар, олардың ішінде Орта Азия халықтарының музыкалық үлгілері бар. Біздің музыкалық корпустың кейбір мәліметтері GitHub платформасында орналастырылады және барлығына қол жетімді болады.

Адам дауысы-зерттеу үшін ең күрделі және қызықты музыкалық аспап. Ән айтудың көптеген аспектілері әлі күнге дейін әншілердің денесінде және денесінде физикалық өлшеу құралдарын қолданудың практикалық мүмкін еместігіне байланысты зерттелмеген, атап айтқанда резонанстық ән мен ән дірілінің ерекшеліктері толық түсінілмейді [13,14] – ән айту кезінде пульсирленген дыбыс/жиілік өзгерісі. Осыған байланысты, біз ән жазуға ерекше назар аударамыз капелла – сүйемелдеусіз.

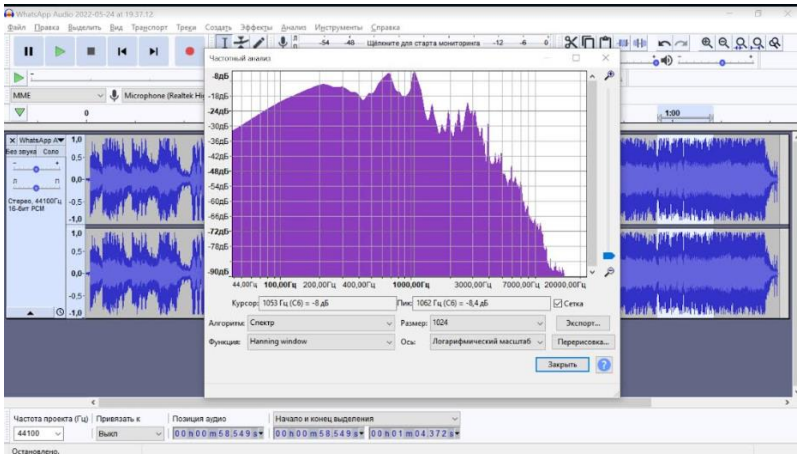
Жиналған музыкалық жазбалар үшін мәтіндер, Музыка және аудио деректерді қамтитын музыкалық аннотациялар жасаймыз: 1) сөздер немесе дауысты дыбыстар, 2) ноталар, 3) фразалар үшін уақыт белгілері және 4) жиілік спектрі мен спектрограммаларды қоса алғанда, дыбыстық сигналдың сипаттамалары.

Спектрлік талдау сигналдарды талдауға негіз болады. Мұнда спектрді талдау үшін үш қарапайым үлгіні (синусоидалы толқын, тікбұрышты толқын, Шу) қолданамыз. Қазақта лирикалық сезімге тұнып тұрған халық әндері қаншама? Соның бірі — бәрімізге белгілі «Япурай» әні. Сурет 1 – де халық әні «Япурай» әнінің «Маржан тағар-ай» фрагмент бөлігінің AnthemScore бағдарламасындағы спектрограмма

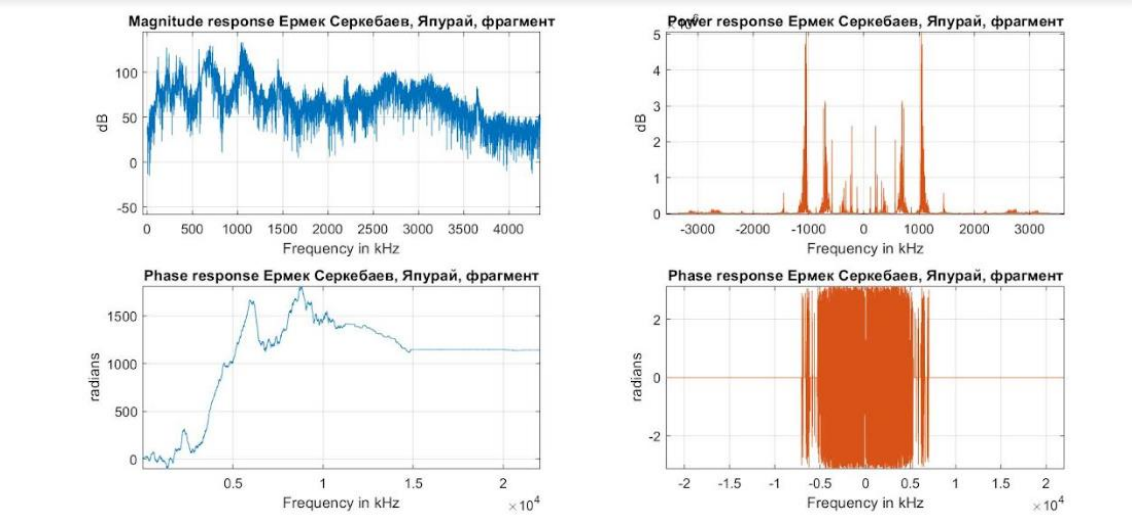
ұсынылып отыр, сонымен қатар қандай ноталарға сәйкес екені көрсетілген, алайда ноталарды анықтауда қателіктер туындайды, сол қателіктердің мүмкін болу ықтималдылығын азайту үшін, спектралды талдау жасайтын барлық бағдарламаларға талдау жүргізу аса қажет. «Audacity» бағдарламасын қолдана отырып, Фурье түрлендіруін әртүрлі терезе формаларымен, спектрлік талдаудың мысалы ретінде «Япурай» халық әнінің спектрлық диаграммасын Сурет 2– де ұсынылып отыр, ал «MathWorks Matlab» бағдарламасында скрипттар дайындалып жасалған спектрлік диаграмманы Сурет 3 – де көрсетілген.



Сурет 1 – Ермек Серкебаев, Япурай, «Маржан тағар-ай» бөлігінің (Тақиялы періште), AnthemScore спектрограмма



Сурет 2 - Ермек Серкебаев, Япурай «Маржан Тағар-ай» (тақиялы періште), «Audacity», бағдарламасындағы спектрлік диаграммасы




Сурет 3 - Ермек Серкебаев, Япурай «Маржан Тағар-ай» (тақиялы періште) халық әнінің «MathWorks Matlab» бағдарламасындағы спектрлік диаграммалар

Жоғарыда айтылған технологияларды қолдана отырып, біріңғай түркі елдерінің музыкасының тарихына арналған, музыкалық ерекшеліктерін арналған платформа жасау, яғни қазіргі таңда бар бағдарламалардың кемшіліктерін анықтау арқылы, барлық түркі халықтарының музыканттарының жазбаларының деректер қорын цифрландыру арқылы өскелең ұрпаққа өшпес мұра қалдыру. Яғни платформа ішінде спектограммалар арқылы аудио файлдан ноталар құрауды, сонымен қатар, мәтіндерді шығару бойынша жұмыстар жүргізіліп жатыр.

**4. Қорытынды**

Қорытындылай келе, бұл мақалада музыкалық ақпаратты зерттеуге арналған музыкалық корпустарын жасауды қарастырдық.

Корпусты құру - бұл күрделі және қымбат жұмыс, оны көптеген зерттеу жобаларында орындау оңай емес. Жұмыс әлі аяқталған жоқ, бірақ біз қазірдің өзінде жиналған коллекциялар зерттеу қоғамдастығы үшін және жобаның нақты міндеттерінен тыс музыкалық ақпаратты зерттеу үшін құнды болуы керек деп санаймыз. Себебі түркі халықтарының музыкалық өнері әлемгі үлгі болатын, ерекше өзіндік құрылымы бар дүние. Ата – бабамыздың мұрасын сақтап қалып, жастарға дәріптеу, цифрлық ақпараттардың кең ауқымда даму заманында тек жастардың қолында.

**Әдебиеттер тізімі**

1. Утегалиева С. (2017). Звуковой мир музыки тюркских народов (на материале инструментальных традиций Центральной Азии), 2017.- 527 с.

2. Карoматли, Ф. (1994) Музыкальное наследие тюркских народов в наши дни // Тезисы I Междуна-родного симпозиума «Музыка тюркских народов». Алматы.

3. Godfrey, J., Holliman, E., McDaniel, J. (March, 1992). SWITCHBOARD: telephone speech corpus for research and development. ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing.

4. Panteli, M., Emmanouil, B., and Simon, D. 2018. “A Review of Manual and Computational Approaches for the Study of World Music Corpora.” Journal of New Music Research 47 (2): 176–89.

5. Patrick, E. (2021). An overview of cross-cultural music corpus studies. Oxford Handbook of Music and Corpus Studies. New York: Oxford University Press, 2021, 2-7.

6. Nilsson, B., Kristina. (January, 2013). What is a corpus and why are corpora important tools? Nordic seminar: How can we use sign language corpora?, Copenhagen, Denmark, December 12-13, 2013, 2-4.

7. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017, March). Convolutional recurrent neural networks for music classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2392-2396). IEEE.

8. Алдошина, И., (1999). Основы психоакустики. Часть 1, 4-6, 1999.

9. Mitsui, K., Saito, Y., Koriyama, T., Tanji, J. and Saruwatari, H. (2019). «JVS corpus: free Japanese multi-speaker voice corpus» arXiv, vol. abs/1908.06248, 2019.

10. Lee, S. W., & Dong, M. (2011). Singing voice synthesis: Singer-dependent vibrato modeling and coherent processing of spectral envelope. In Twelfth Annual Conference of the International Speech Communication Association.

11. Wilkins, J., Seetharaman, P., Wahl, A., & Pardo, B. (2018, January). VocalSet: A Singing Voice Dataset. Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, 468-474.

12. McFee, B., Bertin-Mahieux, T., Ellis, D. P., & Lanckriet, G. R. (2012, April). The million song dataset challenge. In Proceedings of the 21st International Conference on World Wide Web (pp. 909-916).

13. Arroabarren, I., & Carlosena, A. (2004). Vibrato in singing voice: the link between source-filter and sinusoidal models. EURASIP Journal on Advances in Signal Processing, 2004(7), 1-14.

14. Loni, D.Y., Subbaraman, S. (2019). Timbre-Vibrato Model for Singer Identification. In: Satapathy, S., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems . Smart Innovation, Systems and Technologies, vol 107. Springer, Singapore.

***Abjalova Manzura Abdurashetovna***
*Tashkent State University of Uzbek Language and Literature named after Alisher Navoi*
*Tashkent, Uzbekistan*
*abjalova.manzura@gmail.com*

# THE AUTHOR'S CORPUS OF ALISHER NAVOI AND IT'S IMPORTANCE

**Abstract.** There are many great geniuses, talented creators and thinkers in the world. It is important to study their work and make their works readable for today's readers. In this regard, the author's corpus of Alisher Navoi was created so that the reader could understand the work of the world renowned Alisher Navoi. This article discusses the author's corpus of Alisher Navoi and its capabilities, created on the basis of semantic tags of 650 ghazals of the diwan named "Badae ul-vasat" of the poetry collection named "Khazain ul-maoniy" by Alisher Navoi.

**Keywords:** Alisher Navoi, author's corpus, semantic tag, database.

## Introduction

To increase the oral and written speech of students and pupils, to reveal the linguistic potential of the Uzbek language in the XV century to teachers and educators working in educational institutions and to promote the creative heritage of Alisher Navoi worldwide, to modernize our national heritage it is important to achieve widespread use in Information Technology field, to promote the works of our ancestors among young people, to create philological corpus that represent classics of literature in a readable and understandable way, and for this semantic analysis of Alisher Navoi's creative heritage. Therefore, in the first step, most of the textbooks published for secondary education contain samples of works from the "Badoe ul-vasat" diwan of Navoi's "Khazain ul-ma'oniy" poetry collection, in order to understand Navoi and acquaint students with Navoi, Alisher Navoi author's corpus was created on the basis of semantic tags of 650 ghazals in the "Badoe ul-vasat" diwan. "Xazoyin ul-ma'oniy" poetry collection includes diwans of which are: "G'aroyib us-sig'ar" (Wonders of childhood), "Navodir ush-habob" (Rarities of Youth), "Badoe ul-vasat" (Marvels of Middle Age) and "Favoyid ul-kibar" (Benefits of Old Age) which end up being large diwan altogether. It is widely named as "Chor diwan" among the Uzbek nation.

There are 650 ghazals, proverbs, phrases, archaisms, archaic words, talmikhs (to look lightly at something), taschbikhs (similarity), the art of using another word instead of closely related word, irsoil fables, in "Badoe

ul-vasat” diwan (diwan is collection of ancient literatures poetry). It is used firstly in secondary special and higher education under Alisher Navoi’s authority and also new format interfaces are created for new learners. So users can learn to work independently with the works of Navoi, and the sources of the XV century in general, to understand the vocabulary of that time period. In our modern life, we have lots of resources such as sites about Alisher Navoi and his biography and his works, mobile applications and his books are available in .pdf format. But it is not enough to learn all about his life, works, and books for our modern life students. It is necessary to provide a semantically labeled database in order to fill the gap in the field of Alisher Navoi so that the new student is able to understand the meaning of words in books that are difficult to understand.

There are 7 sections:

1. “Alisher Navoi’s autobiography” and its database of work and life of an author.

2. Simple and special searching feature in corpus

3. 8 diwans texts wchich belong to Alisher Navoi

4. 26 works by Alisher Navoi (poems; written on scientific, artistic, tazkira, historical, religious themes) which users can use as a database.

5. Users can access 650 ghazals and poems in “Badoe ul-vasat” that have been semantically tagged.

6. About Alisher Navoi’s corpus

7. Result of researches


Picture 1. The interface of author's corpus of Alisher Navoi

The created corpus is able to satisfy psychological needs of professionals and can provide educational material in the educational stages, and

information in the field. Interface of Alisher Navoi corpus is such as: Alisher Navoi's author’s corpus has educational, linguistic, historical, social, educational, and spiritual significance, the creation of this corpus creates the following opportunities as well:

– learn his personality

– apply an author's approaches

– linguistic analysis of the poet's work

– research the skill of an author of how to use contextual words

– create Alisher Navoi’s special vocabulary

– to Collect his collocations

– to find the authors of anonymous linguistic works through parameters which reflect the personality and style of writer from authority corpus

– to collect the author's paraphrase, paroemia and wisdom; the scope of application of figurative expressions can be determined from the creative context

In order to demonstrate the difference and significance of the corpus from other systems created in connection with the personality and creativity of Alisher Navoi, the following necessary steps will be taken within the project:

1. To identify words that demand further explanation

2. Pair words that require more detailed definition will be semantically tagged

3. Current meanings of the words (which are still being used with linguistic meaning) that require more definition will be provided

4. Current meanings of the words (which are no longer being used with linguistic meaning) that require more definition will be provided

5. The database which includes the geographical terms and names of places that require more definition will be implemented

6. Historical, literary and mythical names (prophets, kingdoms, heroes in poetry and so on) will be semantically tagged

7. It will be possible to work with contexts and move to the full text of the poem through similar variants of words.

The following types of actions are performed in this process: working with dictionaries, working with texts, working with words and phrases that are hardly understood as well as working with contexts. There were a lot of research related to the work of Alisher Navoi in world of philology and in Uzbek philology as well as about his personal life and the translation of his works. To this date, the autobiography of Alisher Navoi and websites which include the text of his works were consolidated ranging from [6,7], various mobile applications and his works which were made available via .pdf format were created ranging from [8,9,10,11], as well as there are Alisher Navoi's

language annotated dictionaries ranging from [1,2,3,4,5]. However, it is clearly seen that aforementioned works done for Alisher Navoi is not even adequate for current student who wants to learn more about Alisher Navoi's way of using contextual words.

The personal life and work life of Alisher Navoi are published via more than [13,14,15,16,17] websites by the world language department and IT spheres. Nevertheless, neither in world linguistics nor in literature has any work been done to create an Alisher Navoi author's corpus. On these terms, it is believed that this corpus will have both innovative sides and practical improvements.

**Conclusion.** The corpus will be available not only to literary critics, but also to representatives of all fields, as well as applicants and researchers, foreign people who are interested in the personality and the work of Alisher Navoi. To promote and develop Uzbek computer and corpus linguistics, to improve the educational process and research, for all ages a modern author's corpus has been created and put into practice to meet the educational needs of specialists, increase vocabulary and literacy.

**Reference:**

1. Explanatory Dictionary of the language of Alisher Navoi's works / E.Under the editorship of fazilov. Editorial board: Konanov A., Kayumov P., Shukurov Sh., Khayitmetov H., Bektemirov H., Karimov K. T. I-IV. – Uzbekistan: Science, 1983.
2. Berdak Yu. Dictionary of Navoi language. – Tashkent: East, 2018. – 496 b.
3. Berdak Yu. Dictionary of classical works. – Tashkent: East, 2010. – P. 6.
4. Muhammad H. Dictionary of works of Alisher Navoi. – Tashkent: Akademnashr, 2017. – 407 b.
5. Shamsiev P., Ibrakhimov S. Dictionary of works of Navoi. – Tashkent: Gafur Gulam, 1972. – 784 p.
6. https://navoi.uz
7. http://alisher.navoiy-uni.uz/
8. https://kitobxon.com/oz/yozuvchi/alisher_navoi
9. https://n.ziyouz.com/kutubxona/category/40-alisher-navoiy-asarlari
10. https://arboblar.uz/uz/people/alisher-navoi
11. https://n.ziyouz.com/kutubxona/category/40-alisher-navoiy-asarlari
12. https://www.opensourceshakespeare.org
13. https://ru.wikipedia.org/wiki/alisher_navoi
14. https://eurasia.expert/rodilsya-uzbekskiy-poet-i-filosof-alisher-navoi
15. https://biographe.ru/znamenitosti/alisher-navoi
16. https://msu.uz/articles/2022-02-12-alisher-navoi-pevets-radosti-i-pechali-chelovecheskoi-dushi
17. https://www.asia-travel.uz/uzbekistan/outstanding-people/alisher-navoi

***[1]Хакимов Б.Э., [2]Шаехов М.Р.***
*[1]Институт прикладной семиотики*
*Академии Наук Республики Татарстан*
*[2]Казанский федеральный университет*
*Казань, Татарстан, Россия*
*[1]khakeem@yandex.ru, [2]q-mir-bey@list.ru*

# СРАВНЕНИЕ КАЧЕСТВА МАШИННЫХ ПЕРЕВОДЧИКОВ В РУССКО-ТАТАРСКОЙ ПАРЕ С ПОМОЩЬЮ ТЕСТОВОГО ПАРАЛЛЕЛЬНОГО КОРПУСА

**Аннотация.** Квантитативные и квалитативные методы оценки машинного перевода имеют собственные ограничения. Количественные метрики дают лишь общую условную оценку, не учитывают степень серьезности ошибок, ориентированы на единственный принятый эталонный перевод и не допускают возможность существования множества правильных переводов. В свою очередь, качественные методы экспертной оценки требуют больших затрат времени и труда, дают в определенной степени субъективную и фрагментарную оценку. С учетом необходимости более точной и детализированной оценки результатов машинного перевода в русско-татарской языковой паре с использованием как количественных, так и качественных методов, нами была поставлена задача создания сбалансированного аннотированного параллельного тестового корпуса. В данной статье предлагается подход к сравнительному анализу разных машинных переводчиков с использованием двух разновидностей разметки тестового корпуса: лингвистической и аннотации ошибок, а также методов краудсорсинга с помощью социальных сетей. Исследование проводится на материале русско-татарских машинных переводчиков.

**Ключевые слова:** машинный перевод, методы оценки качества МП, параллельный тестовый корпус, аннотация ошибок

***[1]Khakimov B.E., [2]Shaekhov M.R.***
*[1]Institute of Applied Semiotics of the Tatarstan Academy of Sciences*
*[2]Kazan Federal University, Kazan, Russia*
*[1]khakeem@yandex.ru, [2]q-mir-bey@list.ru*

# COMPARATIVE QUALITY EVALUATION OF THE RUSSIAN-TATAR MACHINE TRANSLATION SYSTEMS USING THE PARALLEL TEST CORPUS

**Abstract.** Quantitative and qualitative methods for machine translation

evaluation have limitations. Quantitative metrics can give only a general assessment, do not directly take into account the severity of errors, focus on a single accepted reference translation. Qualitative expert-oriented methods require a lot of time and labor, the assessment is subjective and fragmented to a certain extent. Taking into account the need for a more accurate and detailed assessment of the results of machine translation in the Russian-Tatar language pair using both quantitative and qualitative methods, we set the task of creating a balanced annotated parallel test corpus. This article proposes an approach to a comparative analysis of different machine translators using two types of annotation: linguistic and error annotation, as well as crowdsourcing methods using social networks. The study is based on the material of Russian-Tatar machine translators.

**Keywords:** machine translation, MT quality assessment methods, parallel test corpus, error annotation

**Введение**

В настоящее время есть разные подходы к оценке качества машинного перевода. В первую очередь, противопоставляются количественные (квантитативные) и качественные (квалитативные) методы. Их также можно назвать автоматическими и экспертными, хотя при количественном подходе могут быть в некоторой степени использованы экспертные оценки, а качественный анализ не исключает определенной автоматизации. Среди автоматических методов преобладает использование метрик, основанных на сравнении с эталоном (BLEU [Papineni et al, 2002], METEOR [Banerjee, Lavie, 2005], chrF [Popovic, 2015] и др.) или на оценке объема требуемого постредактирования (расстояние Левенштейна [Левенштейн, 1965] и др.). Экспертные методы тесно связаны с разнообразными методиками оценки «человеческого» перевода, обоснованными в традиционной теории перевода [Комиссаров, 2000], разрабатываются оригинальные подходы, учитывающие разные языковые аспекты [Куниловская, 2013].

Использование количественных метрик типа BLEU является одним из наиболее распространенных подходов к сравнению качества машинных переводчиков, однако при использовании тестовой выборки из тех же источников, что и обучающие данные, показатели могут оказаться завышенными и не отражать действительное качество перевода, особенно по отдельным стилям. Справедливо полагать, что этот фактор играет еще более значительную роль для малоресурсных языков с ограниченным объемом доступных языковых данных.

Собственно лингвистические факты используются для оценки результатов перевода намного реже по причине трудоемкости процесса

оценки и потенциальной субъективности экспертных мнений. Среди преимуществ этого подхода более точная и детализированная оценка и фокус на конкретных предложениях и языковых явлениях.

Нами был разработан тестовый корпус для оценки качества русско-татарского машинного перевода объемом более 2000 пар предложений с разметкой по более чем 60 лингвистическим явлениям, включая грамматические категории [Хакимов, Шаехов, 2020].

**Подготовка данных**

Было отобрано 2184 пар предложений (27613 слов в татарском и 28361 слов в русском) на русском и татарском языках, полученных при помощи ручного перевода в большей степени с русского на татарский. Публицистический стиль был представлен текстами новостей, литературный стиль – произведениями художественной литературы на обоих языках, официально-деловой стиль – нормативно-правовыми документами, научный стиль – учебно-методическими текстами, религиозный стиль – текстами канонических книг, разговорный стиль – фразами из литературных произведений.

Подготовка тестового параллельного корпуса осуществлялась в несколько этапов:

1) Подборка списка параллельных предложений объемом не менее 2 тысяч из различных источников.

2) Проверка эквивалентности параллельных предложений и внесение необходимых изменений.

3) Составление списка языковых явлений для аннотации предложений.

4) Автоматическая разметка предложений на татарском языке при помощи морфоанализатора.

5) Ручная проверка и корректировка всех размеченных предложений на татарском языке, снятие многозначности.

6) Автоматическая аннотация предложений на татарском и русском языке предмет наличия отдельных явлений, не затрагиваемых морфоанализатором.

7) Ручной анализ всех предложений на предмет наличия явлений, которые не могут быть найдены автоматически.

8) Финальная проверка и корректировка полученных данных.

**Аннотация и редактирование параллельного корпуса**

С целью повышения точности и детализации тестирования, была выполнена лингвистическая аннотация тестового корпуса. Комбинированным способом (автоматически и вручную) были аннотированы

следующие языковые явления:

1) морфологические категории татарского языка (полный морфоанализ)

2) морфологические категории русского языка (на данном этапе ограниченный набор явлений)

3) наличие явлений омонимии и полисемии (многозначности) в предложениях

4) синтаксические явления

5) наличие некоторых пунктуационных и орфографических явлений (кавычки, сокращения и др.)

Аннотация тестового корпуса осуществлялась путем маркирования наличия/отсутствия определенного явления в каждой отдельной паре предложений. Часть разметки явлений в татарской части была получена после автоматической обработки морфоанализатором и ручной проверки корректности разборов и снятия многозначности. Другая группа явлений, не затрагиваемая морфоанализатором (лексические, синтаксические и др. явления) была аннотирована экспертами вручную. Для некоторых из явлений, размечаемых на данном этапе вручную (составные глаголы, редуплицированные глагольные формы и др.) были предложены правила автоматического определения в предложении для использования в дальнейшей работе.

Для предложений на русском языке на данном этапе были размечены отдельные явления, в дальнейшем планируется дополнить аннотацию за счет полного морфологического анализа. Наличие таких явлений как, например, отрицательные конструкции с *не/түгел* и *нет/юк* также было аннотировано автоматически.

При аннотировании явлений омонимии и полисемии (многозначности) размечались не только случаи «чистой» омонимии и полисемии, а также и по возможности ситуации, когда определенное слово в предложении может потенциально иметь несколько типичных вариантов перевода в зависимости от контекста.

На этапе ручной обработки, помимо собственно снятия морфологической неоднозначности и исправления ошибок морфоанализатора, были обработаны также следующие случаи, связанные с орфографией и пунктуацией:

1) *Разные типы тире и кавычек*. В разных предложениях встречались разные типы тире (длинные, короткие, дефисы) и кавычек, которые были преобразованы к одному типу или вручную на этапе снятия многозначности, или автоматически - после выгрузки данных.

2) *Кавычки+аффикс.* Анализатор не распознает аффиксы, которые присоединяются к словам после кавычек. Например: *«Татнефть»**не**,*

*«Казан утлары»**на**; Орбитага «СириусСат-1»**ны** (иярченнең рәсми атамасы) ачык галәмгә чыккач ХКС Россия экипажының әгъзалары кулдан җибәрәчәк*. Для решения этой проблемы кавычки были временно переставлены на конец слова, чтобы не нарушать целостность морфологической структуры слова. Например: *«17 октябрь манифест**ын**» Габдулла, барыннан да элек, милләт мәнфәгатеннән чыгып кабул итә, җыелачак Дәүләт думасына өмет баглый*.

3) *Цифры+аффикс.* По правилам правописания в этих случаях между цифрой и аффиксом ставится пробел, поэтому к такому виду были приведены все случаю такого рода.

4) *Слово на латинице+аффикс.* Слова, заимствованные и написанные на латинской графике, распознаются как отдельная категория, поэтому аффиксы татарского языка пришлось дописывать вручную. Например: *Алты ел үткәннән соң, ул Габонда француз илчесе итеп билгеләнә, 1982 елда Elf**'ка** кайтып, Африка идарәсен җитәкли*. Причем между такими словами и татарским аффикосм ставится апостроф для разграничения разных алфавитов.

5) *%+аффикс.* Символ процента не распознается как слово, поэтому в этом случае также аффиксы были расставлены вручную, причем между символом и аффиксами по правилам правописания ставился дефис. Например: *Чиста комиссион керемнәр **14,7 %-ка** үсте, 55,4 млрд сум тәшкил итте*. В некоторых случаях предложения были перефразированы так, что символ не применялся: *Бюджетның барлык тармакларына 18,35 млрд сумлык салым күчерелгән - **17 процентка** үсеш белән*.

6) *Сокращения.* Еще одной проблемной зоной для автоматического анализа являются сокращения. Например: *Дәүләт бурычы **29 млрд-ка** кадәр артты (чагыштыру өчен - 2012 елда ул **6,3 млрд** тәшкил иткән, ике ел узгач - 14,8), бу керемнәрнең 90 %-нан артып китә*. В некоторых случаях эти сокращения были расшифрованы для сохранения полноценности содержания: *Берьюлы өч татарстанлы үз үлчәү категорияләрендә иң югары дәрәҗәле медальләр яулады: Ринат Әхмәтшин (90 **килограммга** кадәр), Илдар Аббасов (100 **килограммнан** артык) һәм Елена Михайлова (76 **килограммнан** артык)*.

**Разметка ошибок**

В процессе оценки качества машинного перевода предложения на одном языке переводятся с помощью анализируемого переводчика. Полученные переводы, в дополнение к автоматической количественной оценке в сравнении с эталонами, размечаются экспертами вручную по специально разработанной классификации ошибок. Наличие классификации ошибок, составленной с учётом особенностей перевода в

русско-татарской языковой паре, позволяет осуществлять более точный анализ качества работы машинных переводчиков.

Мы отошли от общепринятой и традиционной в переводоведении классификации ошибок по лингвистическому принципу на лексические, грамматические и стилистические, так как при анализе текстов машинных переводчиков такой подход не является продуктивным.

При разработке классификации ошибок была использована типология MQM (Multidimensional Quality Metrics), разработанная для более универсальной оценки качества перевода [Lommel et al, 2014]. Данная метрика нашла свое применение в области оценки машинного перевода [Freitag et al, 2021; Specia et al, 2020 и др.] и позволяет оперативно и довольно точно определить качество как машинного, так и ручного перевода, в том числе в коммерческой сфере на разных языках в разных вариациях. Как правило, перечисленные в классификации ошибок MQM критерии оценки являются избыточными и могут не применяться в полной мере в каждом конкретном случае в силу трудоемкости процесса. Мы также пошли путем оптимизации и выбрали из имеющихся критериев наиболее подходящие для машинного переводчика в конкретной (русско-татарской) языковой паре. В частности, для данного направления не характерно допущение ошибок при переводе терминологии, даты и времени, локализации понятий и форматов и других. Это связано с межъязыковыми связями в едином информационном пространстве, довольно хорошим качеством тестируемых переводчиков, а также с особенностями анализируемого тестового корпуса.

Таким образом, по итогам обсуждения и апробации для анализа предложений тестового корпуса был взят за основу следующий список типичных ошибок, представленный в Таблице 1.

Таблица 1. Список типов ошибок

| **Индекс** | **Тип ошибки** | **Соответствие в MQM** |
|---|---|---|
| 1.1 | Добавление слов | Accuracy - Addition |
| 1.2 | Прямой перевод | Accuracy - Mistranslation - Overly literal |
| 1.3 | Лишний перевод | Accuracy - Mistranslation - Should not have been translated |
| 1.4 | Неправильный перевод слова | Accuracy - Mistranslation |
| 1.5 | Пропуск слов | Accuracy - Omission |
| 1.6 | Оставлено без перевода | Accuracy - Untranslated |
| 2.1 | Повтор слов или части | Fluency - Duplication |
| 2.2 | Порядок слов | Fluency - Grammar - Word order |

| Индекс | Тип ошибки | Соответствие в MQM |
|---|---|---|
| 2.3 | Грамматическая | Fluency - Grammar |
| 2.4 | Согласование | Fluency - Grammar - Agreement |
| 2.5 | Словообразование | Fluency - Grammar |
| 2.6 | Орфографическая ошибка | Fluency - Spelling |
| 2.7 | Смешение алфавитов | Fluency - Spelling |
| 2.8 | Пунктуация | Fluency - Typography - Punctuation |
| 3 | Стилистическая | Style |
| 4 | Совсем не тот перевод | Non translation |
| 5 | Ошибка в исходнике | Source error |

В процессе анализа ошибок было решено не размечать критичность всех ошибок, при этом допускается дополнительно отмечать грубые (со знаком +) и незначительные (со знаком -) ошибки в каждом предложении. Соответственно, ошибки без пометок означают средний, обычный уровень ошибочности. Это позволило сократить затраты на оценку качества перевода, собирая при этом отдельную информацию по выделяющимся случаям.

С целью апробации разработанной классификации ошибок был проведен ручной анализ качества автоматического перевода предложений тестового корпуса одним из функционирующих и машинных переводчиков - TatSoft - на предмет ошибок по приведенной таблице. Переводчик TatSoft был разработан специально для русско-татарской языковой пары с использованием специальных языковых моделей, учитывающих особенности татарского языка [Khusainov et al, 2018].

Русские предложения из тестового корпуса были переведены на татарский с помощью TatSoft (экспериментальная модель) и проверены экспертами с использованием приведенной выше классификации. Согласно предварительным итогам, 1053 (48%) предложения из корпуса переведено анализируемым переводчиком без ошибок, причем 157 (7%) идентичны эталонному переводу. Статистика по отдельным типам ошибок приведена в Таблице 2.

В дальнейшем планируется осуществить экспертную оценку и разметку переводов предложений из тестового корпуса для основных русско-татарских машинных переводчиков (TatSoft, Яндекс.Переводчик, Google Translate).

Таблица 2. Частотность ошибок машинного переводчика TatSoft

| Индекс | Тип ошибки | Количество предложений с этой ошибкой |
|---|---|---|
| 1.1 | Добавление слов | 31 |
| 1.2 | Прямой перевод | 203 |
| 1.3 | Лишний перевод | 14 |
| 1.4 | Неправильный перевод слова | 148 |
| 1.5 | Пропуск слов | 101 |
| 1.6 | Оставлены без перевода | 17 |
| 2.1 | Повтор слов или части | 66 |
| 2.2 | Порядок слов | 74 |
| 2.3 | Грамматическая | 112 |
| 2.4 | Согласование | 12 |
| 2.5 | Словообразование | 20 |
| 2.6 | Орфографическая ошибка | 5 |
| 2.7 | Смешение алфавитов | 0 |
| 2.8 | Пунктуация | 66 |
| 3 | Стилистическая | 28 |
| 4 | Совсем не тот перевод | 2 |
| 5 | Ошибка в исходнике | 2 |

## Сравнение качества переводчиков

При использовании лингвистически аннотированного параллельного тестового корпуса и экспертной разметки по типам ошибок становится возможным применять для сравнения качества перевода в направлении с русского языка на татарский комбинированный подход с использованием двух разновидностей разметки. Дополнительно мы решили обратиться к методам краудсорсинга с помощью социальных сетей. Такой подход предполагает комбинацию двух оценок:

1) на основе экспертного лингвистического анализа;

2) на основе пользовательского рейтинга, полученного методом краудсорсинга.

Оценка на основе разметки результатов перевода по ошибкам и лингвистическим явлениям может быть представлена такими показателями, распределенными по группам предложений с определенными явлениями, как количество и доля предложений с ошибками, количество и доля предложений с грубыми ошибками, количество и доля предложений без ошибок, число ошибок на 1 предложение и др. Пример таких оценок для экспериментальной модели переводчика TatSoft приведен в таблице 3.

Таблица 3. Распределение показателей ошибочности (фрагмент)

| **Явление** | **Всего предложений** | **Доля без ошибок** | **Число ошибок на 1 предложение** |
|---|---|---|---|
| Посессив, 3 лицо | 868 | 0,58 | 0,64 |
| Множественное число (сущ.) | 631 | 0,58 | 0,66 |
| Директив | 551 | 0,56 | 0,68 |
| Аккузатив | 538 | 0,57 | 0,65 |
| Локатив | 533 | 0,58 | 0,64 |
| Прошедшее категорическое время | 491 | 0,57 | 0,68 |
| Составное сказуемое | 468 | 0,53 | 0,71 |
| Генитив | 463 | 0,58 | 0,67 |
| Настоящее время | 445 | 0,58 | 0,61 |
| имя действия на -у | 398 | 0,58 | 0,60 |
| Наречие на -ып | 359 | 0,51 | 0,74 |
| Причастие на -ган | 357 | 0,51 | 0,80 |
| Составные глаголы | 348 | 0,51 | 0,76 |
| Сложноподчиненное предложение | 315 | 0,52 | 0,74 |
| Аблатив | 238 | 0,54 | 0,68 |
| Глагольное отрицание (-ма) | 222 | 0,50 | 0,70 |
| Каузатив | 219 | 0,58 | 0,62 |
| Пассивный залог | 213 | 0,63 | 0,59 |

Как видно из таблицы, показатели довольно стабильны для разных языковых явлений. В то же время, отклонения от средних значений могут свидетельствовать о сравнительно большей или меньшей вероятности возникновения ошибки.

Как было указано выше, экспертные оценки можно дополнить данными обратной связи от пользователей. Для сбора пользовательских оценок эффективным инструментом является использование Телеграм-бота. В настоящий момент осуществляется разработка бота, который показывает пользователю оригинал и варианты перевода предложений из тестового корпуса, предлагая выбрать лучший из вариантов. Оценки пользователей собираются в базу данных, при этом каждая пара переводов должна получить несколько независимых оценок. На основе полученных данных рассчитываются суммарные показатели голосов, отданных за отделные переводы и по разным машинным переводчикам в целом, с учетом степени согласия оценщиков по отдельным

предложениям. Пользователям не предлагается каким-либо образом конкретизировать и аргументировать свои оценки, учитывается лишь восприятие перевода в целом. Полученные результаты помогут установить корреляцию между субъективным восприятием людей и детализированным анализом по объективным характеристикам.

**Заключение**

Наличие подготовленного и аннотированного экспертами тестового параллельного корпуса повышает достоверность и интерпретируемость результатов автоматической оценки качества машинных переводчиков. Тестовый параллельный корпус дает возможность проводить более точную оценку качества машинного переводчика, выбирать ту модель перевода, которая показывает наилучшее качество.

Предлагаемая классификация переводческих ошибок позволит создавать для русско-татарской языковой пары на основе тестового параллельного корпуса аннотированные наборы данных для анализа ошибок в результатах отдельных машинных переводчиков, обеспечивая эффективный инструмент сравнения разных моделей перевода.

**Список литературы**

1. Banerjee, S., Lavie, A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements. In: Proceedings of the ACL 05 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, pp. 65–72.

2. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., Macherey, W. (2021) Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. In: Transactions of the Association for Computational Linguistics 9(1), December, 2021, pp. 1460-1474.

3. Khusainov, A., Suleymanov, D., Gilmullin, R. (2020) The Influence of Different Methods on the Quality of the Russian-Tatar Neural Machine Translation. In: Russian Conference on Artificial Intelligence, Lecture Notes in Computer Science, vol. 12412, Springer, Cham. 2020, pp. 251-261.

4. Khusainov, A., Suleymanov, D., Gilmullin, R., Gatiatullin, A. (2018) Building the Tatar-Russian NMT System Based on Re-Translation of Multilingual Data. In: Proceedings of the 21st International Conference TSD 2018, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11107 LNAI, pp. 163-170.

5. Lommel, A., Uszkoreit, H., and Burchardt, A. (2014) Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. In: Tradumàtica, Vol. 12 (2014), pp. 455–463.

6. Multidimensional Quality Metrics (MQM). URL: http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html

7. Papineni, K., Roukos, S., Ward, T., Zhu, W-J. (2002) BLEU: a method for

automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of Error classification and analysis for machine translation quality assessment 29 the Association for Computational Linguistics (ACL 2002), Philadelphia, PA, pp. 311–318.

8. Popovic, M. (2015) chrF: character n-gram F-score for automatic MT evaluation. In: ´ Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 2015), Lisbon, Portugal, pp. 392–395.

9. Specia, L., Blain, F., Fomicheva, M., Fonseca, E., Chaudhary, V., Guzman, F., Martins, A.F.T. (2020) Findings of the WMT 2020 Shared Task on Quality Estimation. In: Proceedings of the 5th Conference on Machine Translation (WMT), pp. 743–764.

10. Комиссаров В.Н. Общая теория перевода (лингвистические аспекты): Учеб. для ин-тов и фак. иностр. яз. – М.: Высшая школа, 1990. – 253 с.

11. Куниловская М.А. Классификация переводческих ошибок для создания разметки в учебном параллельном корпусе Russian Learner Translator Corpus // Lingua Mobilis. – 2013 - №1(40). - С.141-158.

12. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. – 1965 – Т.163. - №4. – С.845–848.

13. Хакимов Б.Э., Шаехов М.Р. К вопросу создания параллельного тестового корпуса для задачи машинного перевода в русско-татарской паре // Восьмая Международная конференция по компьютерной обработке тюркских языков «TurkLang-2020». (Труды конференции). Уфа: ИИЯЛ УФИЦ РАН, 2020 – С. 283-292.

14. Переводчик Google. URL: https://translate.google.com/?sl=ru&tl=tt&op=translate

15. Русско-татарский переводчик TatSoft. URL: https://translate.tatar/

16. Яндекс.Переводчик. URL: https://translate.yandex.ru/translator/Russian-Tatar

*ӘОК 004.89.*
***[1]Леспекова А.А., [2]Муканова А.С., [3]Елибаева Г.К.***
*[1, 3] Л.Н. Гумилев атындағы Еуразия ұлттық университеті,*
*[2]Астана Халықаралық университеті,*
*Қазақстан, Нұр-Сұлтан,*
*[1]azizalespekova1998@gmail.com, [2]asiserikovna@gmail.com ,*
*[3]gaziza_y@mail.ru*

## ТЫЙЫМ САЛЫНҒАН КОНТЕНТТІ АНЫҚТАУ ҮШІН МӘТІНДІК КОРПУС ҚҰРУ

**Аңдатпа:** Internet технологияларының дамуына байланысты желіде адам өміріне қауіп тудыратын және мемлекетпен рұқсат етілмеген ақпараттар көптеп таралып жатыр. Сайттардың саны халықтың жартысынан да көп және тез тарауда. Сондықтан ақпараттың үлкен көлемін өңдеу қажеттілігі туындауда және ол күрделі жұмыс. Бұл мәселені ішінара шешуге қазіргі уақытта белсенді түрде құрылған мәтіндер корпусы қызмет етеді. Бұл жұмыста тыйым салынған контентті анықтау үшін қажетті мәтіндік корпусты құру қарастырылады.

**Кілттік сөздер:** мәтіндер корпусы, тыйым салынған контент, интернет

*УДК: 004.89.*
***[1]Леспекова А.А., [2]Муканова А.С., [3]Елибаева Г.К.***
*. [1,3]Евразийский национальный университет Л.Н. Гумилева*
*[2]Международный университет Астана*
*Нур-Султан, Казахстан,*
*[1]azizalespekova1998@gmail.com , [2]asiserikovna@gmail.com ,*
*[3]gaziza_y@mail.ru*

## СОЗДАНИЕ ТЕКСТОВОГО КОРПУСА ДЛЯ ОБНАРУЖЕНИЯ ЗАПРЕЩЕННОГО КОНТЕНТА

**Аннотация:** В связи с развитием технологий Internet в сети все больше распространяется информация, представляющая опасность для жизни людей и не разрешенная государством. Количество сайтов быстро растет. Поэтому возникает необходимость обработки большого объема информации, и это сложная работа. Частичному решению этой проблемы служит активно созданный в настоящее время корпус текстов. В данной работе рассматривается создание текстового корпуса,

необходимого для обнаружения запрещенного контента.

**Ключевые слова:** тексты, запрещенный контент, интернет.

*UDC 004.89.*

***[1]Lespekova A., [2]Mukanova A., [3]Yelibayeva G.***
*[1, 3]L. N. Gumilyov Eurasian National University,*
*[2]Astana International University,*
*Kazakhstan, Nur-Sultan,*
*[1]azizalespekova1998@gmail.com, [2]asiserikovna@gmail.com ,*
*[3]gaziza_y@mail.ru*

## CREATING A TEXT CORPUS TO IDENTIFY PROHIBITED CONTENT

**Abstract:** In connection with the development of Internet technologies, a large number of information that poses a threat to human life and is not authorized by the state is being distributed on the network. The number of sites is more than half of the population and is rapidly gaining popularity. Therefore, there is a need to process a large amount of information, and this is a complex work. A partial solution to this problem is the currently actively created corpus of texts. This paper examines the creation of the necessary text corpus to identify prohibited content.

**Keywords**: text corpus, prohibited content, internet

### Кіpспе

Мақалада тыйым салынған мәтіндерді анықтау әдістерін оқыту және тестілеу үшін мәтіндер корпусы сипатталды. Сонымен қатар қазақ тіліндегі тыйым салынған мәтіндерге немесе заңсыз мазмұндағы материалдар жиынтығы жасалды. Жинақталған мәтіндер көмегімен тыйым салынған контентті анықтау көрсетілді.

Тыйым салынған контент – бұл мемлекеттен тыйым салынған ақпараттық ресурстың немесе веб-сайттың кез келген деректерін адамдарға көшіруге, таратуға және көруге рұқсат етілмеген мазмұнды айтамыз [1-2]. Тыйым салынған контентке ұятсыз мәтіндер, мультимедиа, құмар ойындар, митингке шақыртулар, терроризм ұйымдастырушылық, қатыгездік, кісі өлтіру және т.б. тақырыптары бар желідегі ақпаратты жатқызамыз. Бұл тақырыпқа сәйкес барлық ақпараттар бақылауға алынады, ал қауіпті деп танылса бірден бұғатталады. Тыйым салынған ақпаратты таратушылар заң бұзғаны үшін айыппұл төлеуі керек, ал қасақана ұйымдастырылған жағдайда қауіпсіздік күшейтіліп, қолайсыз мазмұнды насихаттағаны үшін

қылмыстық жауапкершілікке тартылады.

Порнография және жыныстық қанағаттануға арналған барлық деректер немесе сексуалдық сипаттағы қызмет көрсетуді насихаттайтын ақпараттарды жариялауға тыйым салынады. Сонымен қатар, ақша үшін жыныстық қызметтерді ұсынатын қосымшаларға да мемлекетпен тыйым салынады.

Нәсілдік, ұлттық, діни, жыныстық, мүгедектік, ардагер мәртебесі, жыныстық бағдар, гендерлік сәйкестілік және басқа да белгілер негізінде кез-келген адамдар мен әлеуметтік топтарға зорлық-зомбылықты насихаттайтын қосымшаларды жариялауға мемелекетпен тыйым салынады

Жарылғыш заттарды, атыс қаруын, патрондарды, атыс қаруына арналған кейбір бөлшектерін сатып алуға болатын қосымшаларды жариялауға мемлекет рұқсатынсыз сатуға тыйым салынады.

Тыйым салынған контентті анықтаудың негізгі мақсаты - заңсыз таратуға тыйым салынған ақпаратты қамтитын "Интернет" желісіндегі сайттарға кіруді шектеу.

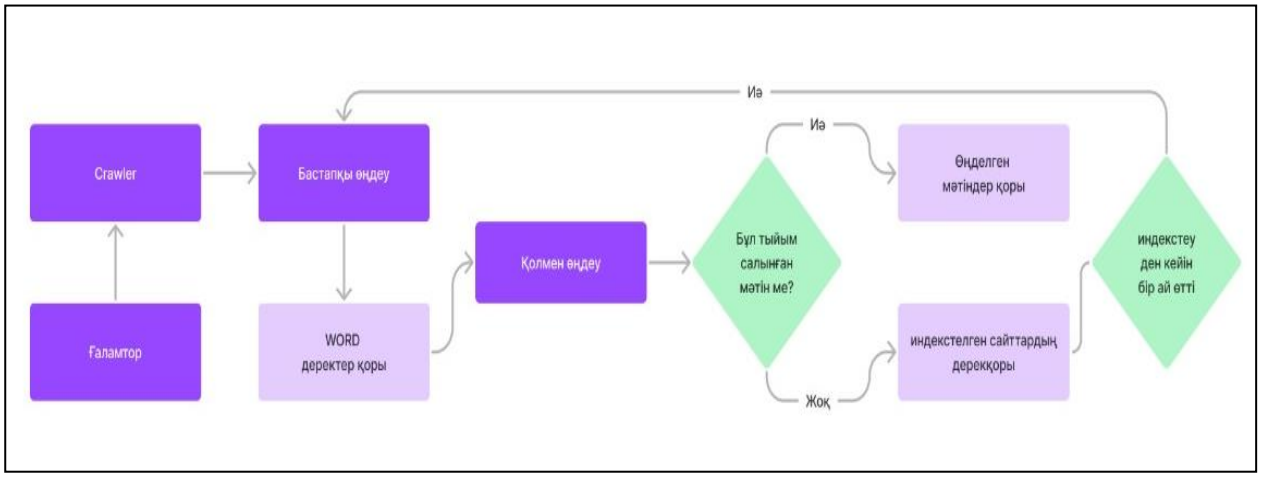
Мәтіндік корпус – сөйлемдерді және лингвистикалық ақпараттарды басқару және жинақтау жүйесі [3]. Оны көбінесе біздер корпусты басқару жүйесі немесе корпустық менеджер деп атаймыз. Бұл корпустағы сөйлемдер мен сөздерді іздеуге, бізге қажетті ақпаратты алуға негізделген жүйе. Корпустарды құрудың мақсаты мен пайдаланудың мәні келесі алғышарттармен анықталады:

– бір рет құрылған және дайындалған мәтіндік корпусты бірнеше рет, әр түрлі зерттеулерде және бірнеше жеке мақсаттарда пайдалануға мүмкіндік береді;

– әртүрлі тоналдылыққа ие деректер корпуста өзінің шынайы контекстік формасында болады. Бұл оларды кеңінен және жан-жақты зерттеуге мүмкіндікті тудырады;

Жоғарғы репрезентативтіктілікке ие толық өңделген корпусымыз, деректердің шынайылығана ақпараттың дұрыстығына кепілдік береді.

Бұл жұмыста сипатталатын корпус жартылай автоматты түрде жасалынған. Ол үшін бірнеше модульдерді қолданатын боламыз. Олар: Crawler, requestGenerator, pagePreprocessing. Мәтіндік корпус архитектурасы төмендегідей. Тыйым салынған контентті анықтауға мүмкіндік беретін мәтіндік корпус архитектурасы ашып көрсетілген.




Сурет 1 – Тыйым салынған контентті анықтауға арналған мәтіндік корпус архитектурасы

Figure 1- Text corpus architecture for detecting prohibited content

Суретте мәтіндік корпус архитектурасының сипаттамасы берілген:

– мәтіндік корпуста ең бірінші crawler кілттік сөздер бойынша сайттарды іздей бастайды.

– екінші бастапқы өңдеу басталады және html тегтері жойылып кіші регистрге жазылады.

– өңделген мәтін WORD деректер базасына жазылады.

– адамдар WORD деректер базасындағы мәтіндердің мағынасына қарай тыйым салынғандыққа анықтайды.

– егер мәтін мағынасына қарай тыйым салынған болса, негізгі деректер базасына қосылады.

– егер мәтін мағынасына қарай тыйым салынбаған болса, индекстелген сайттар деректер базасына қосылады.

Егер сайт индекстелген сайт деректер базасында 30 күннен артық күн жатса, ол бастапқы өңдеуге қайта қосылады. Оның мәні – сайтқа қауіпті ақпарат қосылмағанына көз жеткізу болып табылады.

Мәтіндер корпусында заңсыз мәтіндер жеті санатқа жіктеледі: терроризм, идеологиялық мәтіндер, діни өшпенділік, сепаратизм, ұлтшылдық, агрессия және тәртіпсіздікке шақыру, фашизм, сондай-ақ ұқсас лексикасы бар бейтарап мәтіндер. Тыйым салынған контентті мәтіндік корпус арқылы анықтауға жүргізілген жұмыстар:

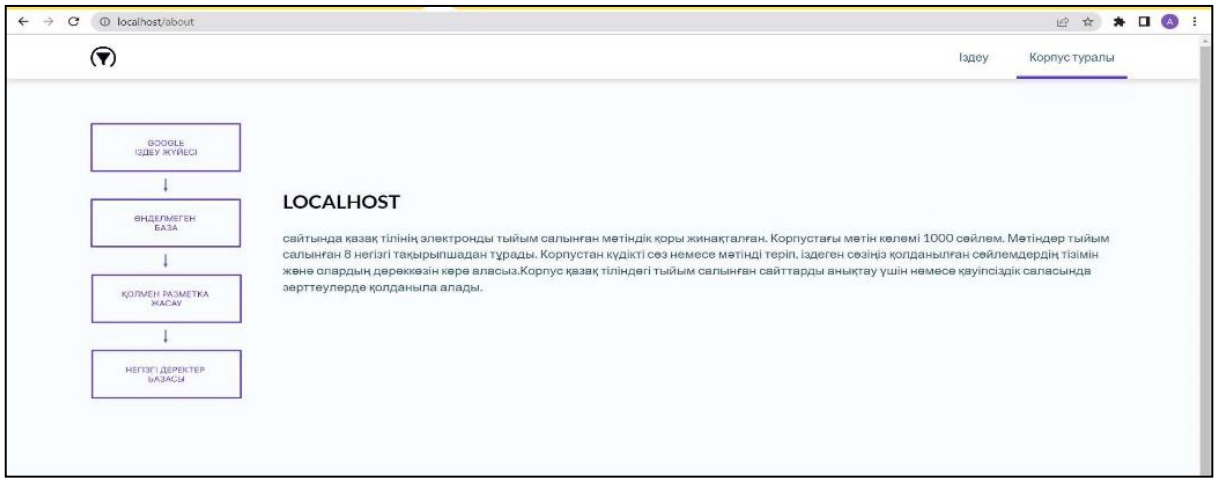
– Ағылшын тіліндегі экстремистік мәтіндердің дайын корпусы сипатталған [4]. Корпусы (Narrative Networks Corpus) діни әңгімелерден, интернеттегі материалдардан және жарнамалық журналдардан алынған исламистік экстремизмге қатысты 100 мәтінді құрайды (42 480 сөз). Барлық мәтіндер корпусындағы сөйлемдер араб тілінде жазылған. Бірақ қазіргі уақытта толығымен ағылшын тіліне аударылып жазылды.

Корпуста сөйлемдерді тоналдылыққа талдау жүйелері бар. Ол автоматты түрде жинақталып, содан кейін қолмен жеке тексерілді.

– Орыс тіліндегі экстремистік мәтіндердің дайын корпусы сипатталған [5-8]. Жұмыста экстремистік бағыттағы мәтіндерді анықтау әдістерін оқыту және тестілеу үшін мәтіндер корпусы жасалған. Қазіргі уақытта корпустың жалпы көлемі – 493 мәтін (650 000 сөз), оның ішінде 368 мәтін экстремистік материалдар санатына жатады. Барлық мәтіндер қолмен жиналған.

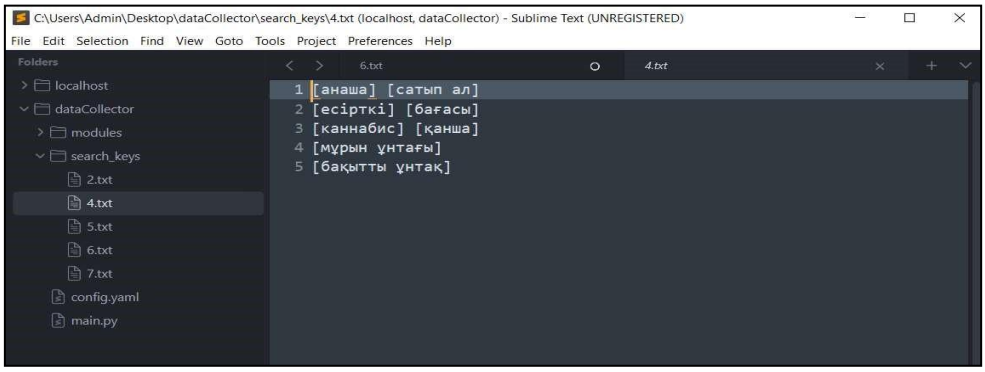
– Қазақ тіліндегі WEB-ресурстарда экстремистік бағытты анықтау үшін түйінді сөздер жинағын құру сипатталған [9]. Бұл жұмыста мәтіндегі тыйым салынған контентті анықтау жүзеге асыру сипатталған. Осы мәселені шешу негізгі бес кезеңге бөлінген: тыйым салынған аймақтың веб-сайттарын анықтау, мәліметтерді алуға дайындықты қарастыру, мәліметтерді өндеп алу, мәліметтерді бөлу және талдау.

Тыйым салынған контентті анықтау үшін мәтіндік корпустың жұмысын талдамас бұрын корпустың өзіне тоқталып өтейік. Корпус тыйым салынған сөйлемдер бойынша белгілі бір категорияларға бөлініп жинақталған. Сөйлемдер интернет желісінен ізделініп, өңделмеген базаға тіркеледі. Әрі қарай қолмен разметка жасалып өзінің категориясына анықталады. Негізгі базаға анықталған категория бойынша тіркеледі. Төмендегі суретте корпустың сайттағы сипаттасы көрсетілген. Localhost сайтында қазақ тілінің электронды тыйым салынған мәтіндік қоры жинақталған. Корпустағы мәтін көлемі 1000 сөйлем. Сипаттамада гугл жүйесіне іздеу жүйесінен мәтіндер өңделмеген базаға тіркеледі. Қолмен анықтау жүйесінде мәтіндерді өңделген базаға тіркеледі.



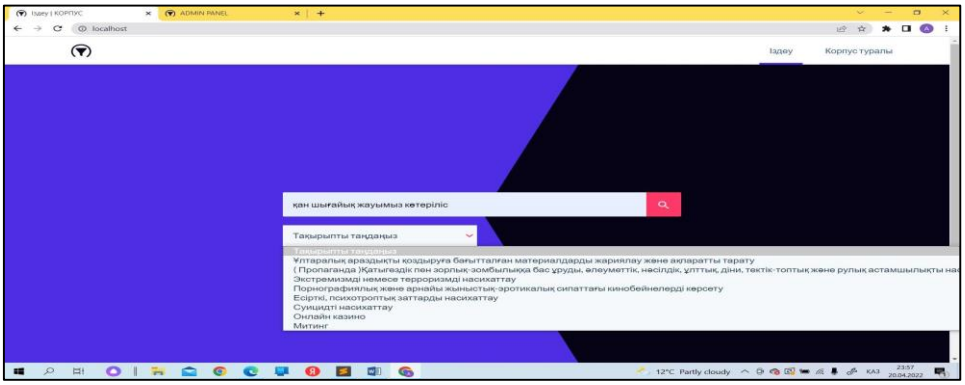

Сурет 2 – Корпустың сайттағы сипаттамасы
Figure 2-Description of the building on the site

Мәтіндер тыйым салынған 7 негізгі тақырыпшадан тұрады. Корпустан күдікті сөз немесе мәтінді теріп, іздеген сөзіңіз қолданылған сөйлемдердің тізімін және олардың дереккөзін көре аласыз. Корпус қазақ тіліндегі тыйым салынған сайттарды анықтау үшін немесе қауіпсіздік саласында зерттеулерде қолданыла алады.




Сурет 3 – Кілттік сөздердің жинақталуы
Figure 3-Accumulation of key words

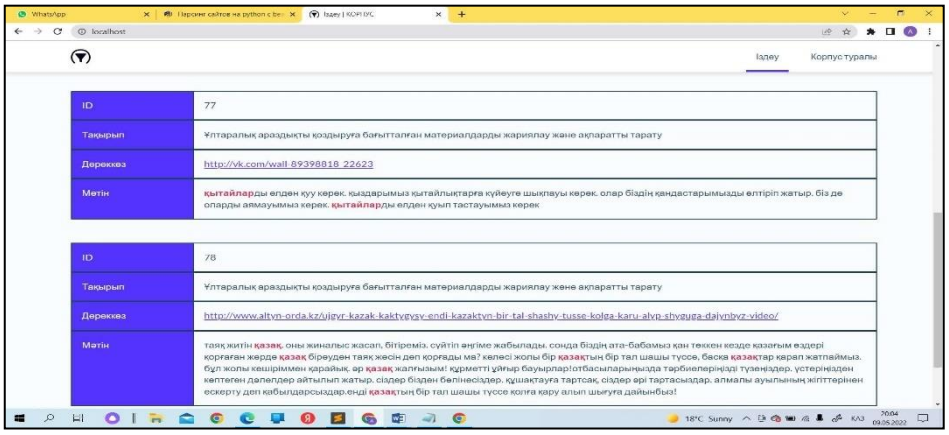
Жоғарыдағы суретте тыйым салынған сөздердің әрбір категориясына кілттік сөздерді жинақтадық. Экраннан көріп тұрғаныңыздай, ол 7 негізгі мәтіндік форматқа бөлінген. Ол жерде кілттік сөздер немесе дәлдік сөйлемдер арқылы да іздестіруімізге болады. Әрбір сөз тік жақша көмегімен бөлінген түрде жазылады. Әрбір жолдағы сөздер интернет желісіне тек бір сұрау болып жүргізіледі.




Сурет 4 – Мәтіндік корпустың қолданылуы
Figure 4-Using a text corpus

Мәтіндік корпусты қолдану үшін өзімізге қажетті қауіпті деп танылатын сөздер немесе сөйлемдерді енгіземіз. Енгізілген ақпаратқа сәйкес қай категорияға жататынын (порнографиялық және арнайы жыныстық-эротикалық сипаттағы кинобейнелерді көрсету ұлтаралық

араздықты қоздыруға бағытталған материалдарды жариялау және ақпаратты тарату қатыгездік пен зорлықзомбылық жасауды насихаттау әлеуметтік, нәсілдік, ұлттық, діни, тектіктоптық және рулық астамшылықты насихаттау экстремизмді немесе терроризм жолына түсуге үгіттеу есірткі сияқты психотроптық заттарды насихаттау суйцидті насихаттау және лицензиясы жоқ онлайн казино, митингке үгіттеу) таңдаймыз. Соңында іздеу батырмасын басамыз.




Сурет 5 – Ұлтаралық қақтығысқа категориясы бойынша табылған мәліметтер

Figure 5 - Found data on the category of interethnic conflict

5-суреттен ұлтаралық араздықты қоздыруға бағытталған материалдарды жариялау және ақпаратты тарату тақырыбына қатысты сөздер бойынша табылған сайттардың ақпараттары көрсетілген. Басқа ұлт өкілдеріне қарсы сөздер және өзге ұлтты елімізден қуып шығу секілді ақпараттар жинақталды. Қазақ ұлтына қатысты кері айтылған ақпараттар да тіркелді. Мысалы:

77 номерде қытай халқын арандату туралы;

78 номерде қазақ халқын арандату туралы ақпараттар тіркелді.

Қорытындылай келе бұл мақалада интернет желісінен тыйым салынған контентті мәтіндік корпус арқылы анықтау жүйесінің архитектурасы толығымен сипатталды. Тыйым салынған контентті анықтауға мүмкіндік беретін мәтіндік корпусты құрудың жолы қарастырылды.

**Әдебиеттер тізімі**

1 Ельчанинова Н.Б. Проблемы совершенствования законодательства в сфере ограничения доступа к противоправной инфорамции в сети Интернет// Общество: политика, экономика, право-2017.-№12. – С. 119-121

2 Марценюк А.Г. Запрещенная информация и ее место в системе информационных отношений// Гражданин и право-2018. - № 5-С.62

3 Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., переработанный и дополненный., – СПб.: СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

4 learning techniques for sentiment classification. InACL. The Association for Computer Linguistics

5 Богуславский И. М. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации” // Труды Международного семинара по компьютерной лингвистике и её приложениям "Диалог-2000". Протвино, 2000.

6 Корпусная лингвистика и контекст (в соавт. с Ю. Н. Марчуком) // Межвузовский сборник научных трудов "Теоретические и практические аспекты лингвистики и лингводидактики". - Сургут: Изд-во СурГУ, 2002. - С. 123-128.

7 Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора / Ю. В. Рубцова // Программные продукты и системы. –2015. – № 1. – С. 72–78

8 Захаров, В. П. (2015). Оценка качества Интернет-корпусов русского языка. В Труды международной конференции «Корпусная лингвистика2015» (стр. 218-229). Издательство Санкт-Петербургского университета

9 Bolatbek M. A., Mussiraliyeva S. Z., Tukeyev U. A. Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language // KazNU Bulletin. Mathematics, Mechanics, Computer Science Series.

***[1]Abdurakhmonova N., [2]Tuliyev U., [3].Ismailov A, [4]Abduvahobo G.***
*[1, 2]National university of Uzbekistan, Tashkent, Uzbekistan,*
*[3]Andijan Machine Building institute, Andijan, Uzbekistan*
*[4]Fergana state university, Fergana, Uzbekistan*
*[1]abdurahmonova.1987@mail.ru, [3]alisherismailov1991@gmail.com*

# UZBEK ELECTRONIC CORPUS AS A TOOL FOR LINGUISTIC ANALYSIS

**Abstract.**This article analyzes the theoretical and applied foundations of Uzbek electronic corpus applying as a linguistic tool in computational linguistics. Information is provided on the functional capabilities of the Uzbek language electronic corpus. It is discussed that experiences in building the linguistic and software development of the corpus and ideas for the overall conceptual architecture of the corpus.

Structure and its linguistic annotation and metadata, and corpus manager are important for usage for many purposes. The fact that the platform allows users to address linguistic analysis issues in the domain of computational linguistics and NLP.

The Uzbek corpus based on structural and sub corpus models, which partially represented in this paper, is going on process to develop Uzbek language technology.

The functional capabilities of the Uzbek corpus http://uzbekcorpus.uz/ are represented as a tool of language analysis of NLP.

**Keywords**: Uzbek electronic corpus, software, computational linguistics, morphoanalyzer, metadata, parallel corpora, text analysis, corpus manager

## 1. INTRODUCTION

It can be said at the present time that corpus is one is main tools for natural language processing and as a research object for other cross fields to study language and speech knowledge. Particularly, corpus is important tool if the language is considered lack of resources of language as linguistics resource and platform. Hence creation corpus is necessary all aspects of development language technology. Our scientific achievement implemented in platform of http://uzbekcorpus.uz/ . Our project focused on not only text fragment compilation but also to achieve semantic and grammatical knowledge of the Uzbek language in order to construct models of the applications of NLP. In our researches multispectral issues have been discussed in the the number of works [8, 9,10,11,12,13,14,15,16,17, 18].

As according to principles of corpus manager whose search system should be considered pilot formal-functional models of corpus linguistics. However, corpus differentiate with annotation (existing or not), to develop forward the next stages not only for linguistic approach but also other spheres as well. Considering experience of a number of corpora models, there are different kind of corpus manager, linguistic annotation and metadata is required for corpus to apply information for a number purposes. Creation corpus is multilevel process for language processing. V. Zakharov [1] points out some requirements for contemporary corpus manager to be construct concordance, search by not only word but also collocation, search by fragments (complex query), sort list according to some requirement by user chosen, make it possible to display the found word forms in an extended context; provide statistical information on individual elements of the corpus; display lemmas, morphological characteristics of word forms and metadata (bibliographic, typological), which depends on the degree of markup of the corpus; save and print the results; work both with separate files and with corpora of unlimited size.

Each scenario of algorithm of analysis of corpus belongs to language characteristics and grammatical rules. For example, for European languages there are ready tools for morphological analysis and parsing as a stage of stemming, tokenization and lemmatization. These tools are accessible to users via WebLicht2, a tool chainer providing both infrastructural services and a GUI for combining the individual tools [2].

Encoding text of corpus is important for representation of language analysis. There are several format to input data in corpus. A corpus is prepared by whole texts or of fragments or text samples by different genres oral or written language texts.

Parameters of the text given in metadata with the name of author, name of the work, genre, publishing house and year, Latin or Cyrillic.

## 2. CORPUS MANAGER SYSTEM

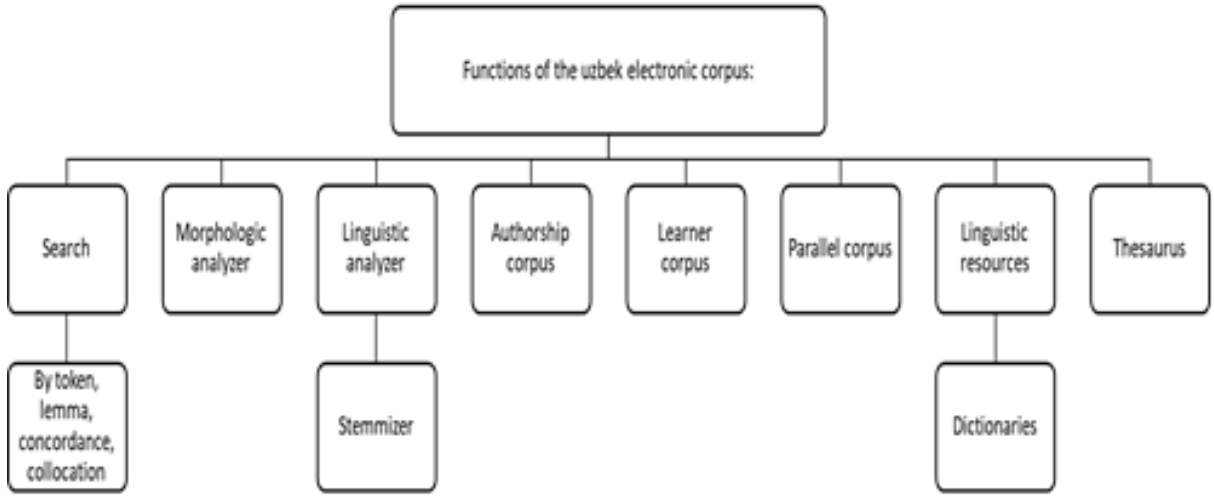
Mostly, often, such systems are based on ready-made solutions, which leads to problems with the speed of search queries (data samples), system flexibility, scalability. Our observation shows that the corpus manager of many language corpus is controlled by ready-made technologies. For example, the national corpus of the Russian language is Yandex. Uses which owns the servers fast and multifunctional search engine. This search engine consists of creating direct and inverse queries, as well as logical operations when searching by logical operators and, or (conjunction, disjunction). Another is a Sketch Engine system that supports document metadata independently and uses a special, query language (CQL - Corpus Query

Language) to view corpus statistics. The search engine of the case, created for the Tatar language, consists of a control panel for checking and filtering data. After this step, the following models will be launched: SinglePageModel, SearchModel, QueryModel (the request is sent directly to the SearchModel system, not from the controller), ContextModel, SinglePageEditModel, StatisticsModel, DataManagementModel. Functional models such as DocumentModel, SentenceModel, WordModel, SecurityModel were also used effectively [3].

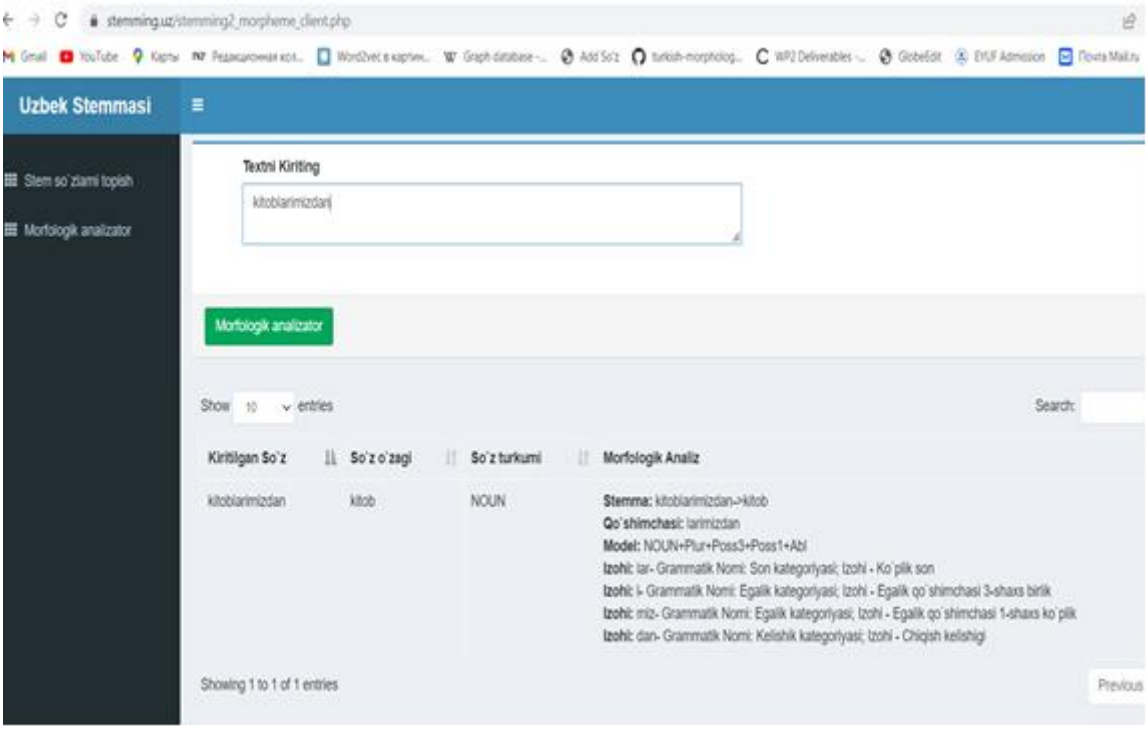
Uzbek language corpus manager is available at http://uzbekcorpus.uz/ (Pic.1). Formal-functional model of the Uzbek language corpus consists of architecture of the text types, corpus manager system, and language analysis. Search engine are intended for use as objects of various fields. In the field of pedagogy and computer linguistics, in particular, generated

texts can be used effectively through a search query provided in the interface using the corpus manager. Search by lemma, token and concordance organized by n-gram by lexical units. By token search system algorithm count each words by word in lexicon and rest of parts analyzed at tokenization stage. In Uzbek words might be derivational and grammatical forms:

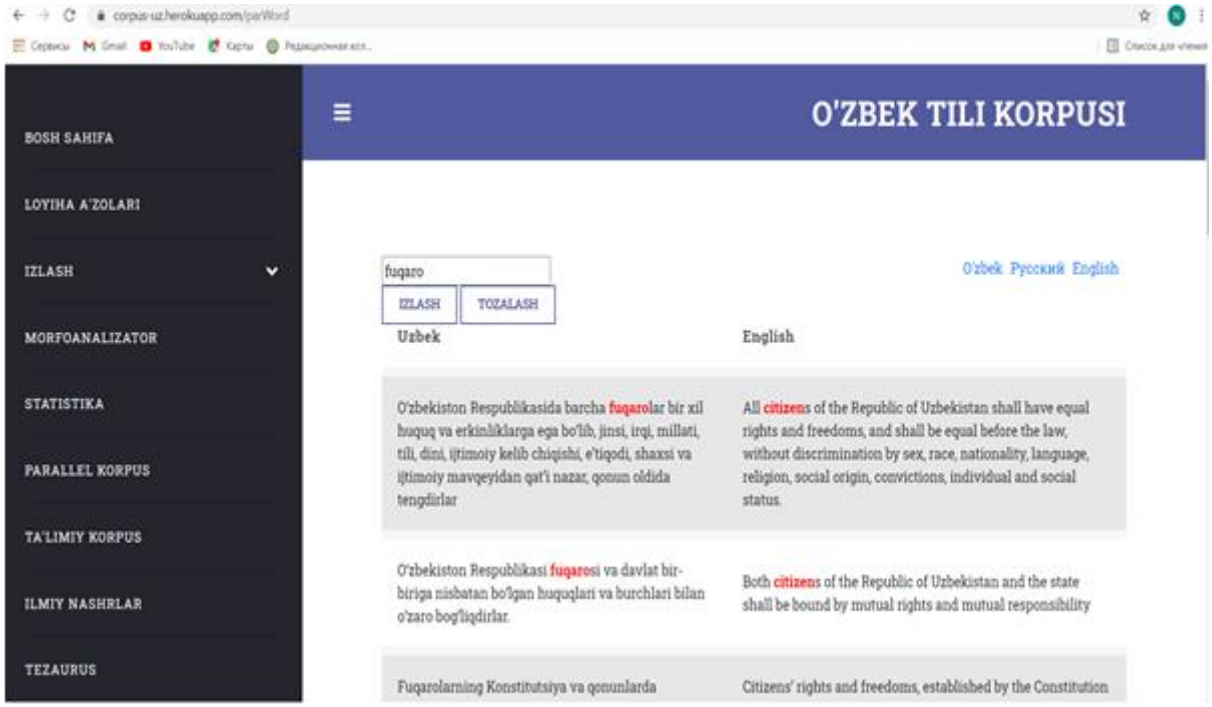
[gul]=>{gulchilik-gulli-guldor, gulzor}=>derivation words




[gul]=>{gullar; gulag; gulgami; gulniki, gullardanmi…}=>grammatical categorization




Parallel corpus includes Uzbek-English parallel texts in official and scientific style. After segmentation text and word forms alignments input database (pic.2)




Pic.2. Subcorpus –Parallel corpus between Uzbek and English

The use of a translation memory environment not only speeds up the human translation process, but also has a positive effect on its quality. This is because usage the features of this program could give opportunity to show

mistakes and correct them in the text, as well as to match the meanings of different styles of bilingual texts. The definition of concordances in parallel texts is based on the alternative equivalence of a word, phrase, or stable combination in a given language. In this case, adequate translation does not always justify itself, as some words are dropped or a component is added to change their lexical and grammatical model. Therefore, for keywords in parallel texts, active words and terms that are frequently encountered in the context and have not undergone specific changes in translation are aggregated into a database. In this case, auxiliary word groups in Uzbek (connective, auxiliary and preposition), words that do not have an independent meaning (imitation, adverb, auxiliary verbs, independent verbs, etc.) and we will exclude from database the words which are often used as homonyms in the text. Because it is impossible to predict their meaning in translation. The number of compound words in Uzbek language is relatively large and can be as follows: {n… S / S… n1} => compound word (point of view / disregard) {n… SWn} => phrase (look down)

For example, the following keywords are considered as frequently used normative templates in abstracts: word base => derivation form => word form => frequency of frequently used words, phrases or terms => concordance => translation. The text in the sample is a "long" sentence because it has a large number of punctuation and predicative elements, and the text has been translated into two parts by the translator. The generation of parallel sentences separated from the large text context can be done in the following steps [5]: Information about alternative pairs of tokenized words from parallel text is obtained; A long sentence is divided into separate segment sentences up to the part with certain punctuation marks “, ”, “; ”, “: ”; Basis - the amount of content of the segments of the translated sentences is counted and the state of conformity is determined; Translation - the alternative status of the main texts is determined; If more or more relationships are observed between segments, several segment units are interconnected or attached in the form of a single relationship.

Learner corpus is focused to study lexical minimum of school pupil or language acquisition Uzbek as a second language. It has literatures 5-11 classes with metadata. Corpus is for specialized purposes corpus 3content scientific and publicity style.

Our corpus includes language analysis system: morphoanalyzer, parsing, and semantic analysis. Linguistic annotation as a level of morphology we intended to tag universal markup tagging system. Therefore, there are two graphemes of parts of speech and their POS tagging set. Search system of corpus is enclosed search system by lemma, token, KWIK and collocation. Preparation of language query system for corpus as mentioned before we

included 85 thousand words in Uzbek lexicon for both graphemes. Here showed POS tagging set for morphology:

| Example of POS | TAG | National name |
|---|---|---|
| *o‘quvchi (pupil)* | **NOUN** | **Ot** |
| *bor (go)* | **VERB** | **Fe’l** |
| *go‘zal (beatiful)* | **ADJ.** | **Sifat** |
| *tez (fast)* | **ADV.** | **Ravish** |
| *bir (one)* | **NUM.** | **Son** |
| *hamma (all)* | **P** | **Olmosh** |
| *o‘qish (reading)* | **V_N** | **Harakat nomi** |
| *kulayotgan (laughing)* | **V_S** | **Sifatdosh** |
| *borgiz (s+make go)* | **V_O** | **Orttirma nosbatdagi fe’l** |
| *ko‘ril (obj+was seen)* | **V_PASS** | **Majhul nosbatdagi fe’l** |
| *orqali (by, through)* | **ADP** | **Ko‘makchi** |
| *-mi (question meaning of the word)* | **PART** | **Yuklama** |
| *agar (if)* | **CONJ** | **Bog‘lovchi** |
| *vov (av-av)* | **Imit.** | **Taqlid so‘z** |
| *ehtimol (probably)* | **MW** | **Modal so‘z** |
| *voy! (wow)* | **EXL** | **Undov so‘z** |

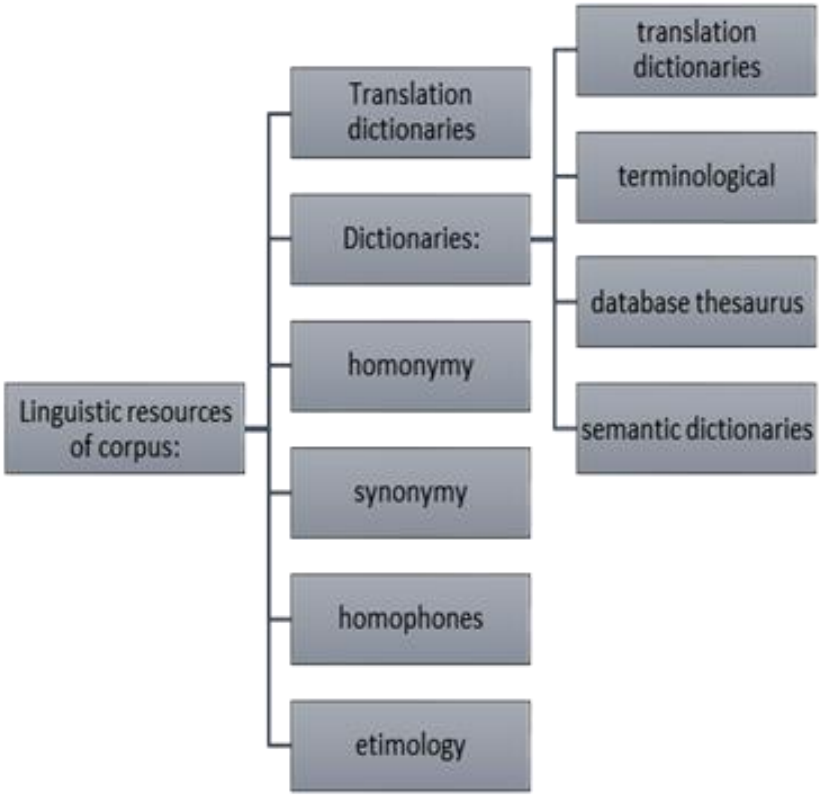
Concordance is searched by left and right side algorithm. Due to the corpus contained documents in both Cyrillic and Latin alphabets, lexicon was also stored in two alphabets. The algorithm for finding the lemma is shown in the example of the Latin alphabet. a list that contains all the words of the Latin variable. Looking for a lemma searched from left to right at the beginning of a word, the rest of the word is searched within the suffix models corresponding to the current word (suffix models are stored in the list of suffixes). The lemma is considered correct only if the search among the attachment models is successful result.

Corpus technology allows to analyze more deeply. As a result of our research, an electronic corpus of Uzbek language has been developed, which will serve as an object of study for a number of areas, such as computer linguistics, applied linguistics, pedagogy, translation theory and practice, Uzbek linguistics and literature. Foreign researches in the conceptual and structural design of the electronic corpus of the Uzbek language has been learnt, linguistic structure of the language was translated into machine language by applying automatic methods of morphological and syntactic

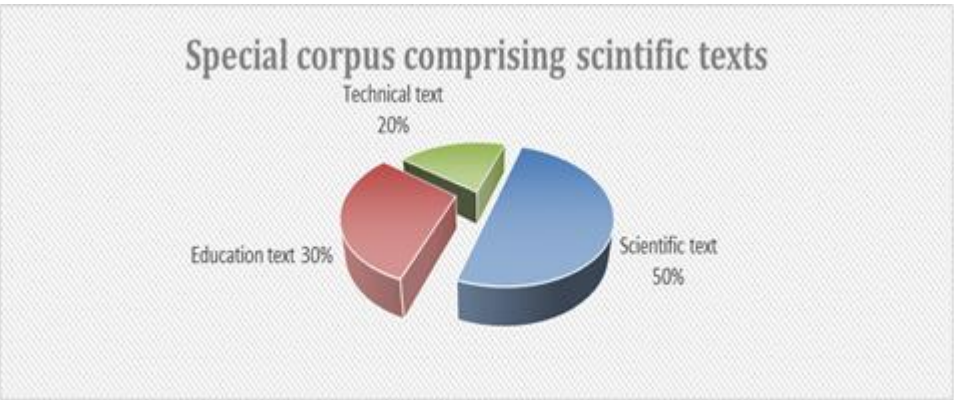
tagging and analysis such as FST and UdPipe in the development of the linguistic corpus of the language. Linguistic and software of the text corpus was created to analyze the representativeness of the text fragment and search units (lemma and token). The corpus interface was formed on the basis of formal-functional models of the corpus manager. Corpus search engine for concordances on the lemma, token, phrase and n-gram model of the Uzbek language corpus manager (search engine).

The electronic corpus of the Uzbek language consists of the following parts: SEARCH-> search interface (where a total of 150 million tokens of text are searched by lemma, token, phrase and n-gram model). Each file clissified according to styles of Uzbek: literally, scientific, official, publical and spoken texts. Moreover each literal works classified by epoch of Uzbek literuteres: Temuries, Khanlik, Jadidism, Soviet and Independence.

In linguistic resource section there are Uzbek translation of 120,000 English verb phrases, bilingual translation dictionaries, terminological database (belonging to 12 fields), synonym, homonym, paronym, antonym, phrase, etymological dictionary such as electronic dictionaries.

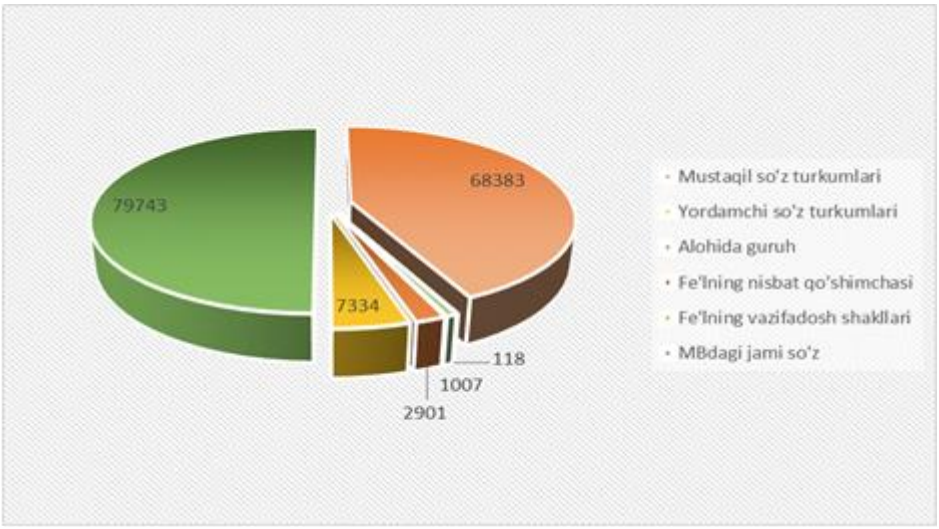



In the Thesaurus section, a database of comprehensive concepts based on semantic relationships (hierarchical, equivalent, equivalent), as well as language units including proverbs and phrases, is created on the example of the concept of “family spirituality”. A number of concepts in the Uzbek language are currently being studied.




The linguistic processor part of the case contains a morphological analyzer and a linguistic analyzer, which are used to separate words into stems, morphologically tag them, and identify attributes.

The scientific significance of the research results is the representation of the proposed and developed structured text structures and suggestions for the analysis of the text at different linguistic stages.In the future, the technology of processing Uzbek texts, the corpus of parallel texts, the object of study for representatives of a narrow field will help to conduct research focused on language and speech phenomena. The following statistical results were achieved in the morphological classification of words.




In conclusion, as one of the global challenges of the 21st century, it is important to consistently conduct research on NLP and language technologies in the creation and development of electronic corpuses of world natural languages. In this case, the most optimal form of using natural language is the digital (electronic) form, and it is important to present convenient and fast way of processing and analysis on a machine (computer).

## 3. Further works and discussion

The automatic construction of ontologies for specific subject areas, or at least the construction of a list of terms and hypotheses about possible relationships between them for subsequent manual processing, is an important and urgent task of modern computational linguistics [7]. According to the opinion of the above scientists, “Currently, there is no generally accepted method for constructing ontologies.”

In computational linguistics, the introduction of terminology automation using corpus technologies gives positive results in the field of science. This has practical implications for translation studies, machine translation, terminology, and other areas of computational linguistics. In our study, dedicated to the analysis of medical terminology and its functioning in scientific texts are of particular importance for linguists and programmers.

In order to extract the key terminology, we used, describing the subject area of the development of different subjects’ area texts, carried out on the basis of a corpus of special texts.

In the perspective we are going to expand functional capabilities of corpus as a tool and open resource for the issues of computational linguistics and NLP.

**References:**

[1] В. П. Захаров, И. В. Азарова, О. А. Митрофанова, А. М. Попов, М. В. Хохлова (2019) Моделирование в корпусной лингвистике Специализированные корпусы русского языка, Санкт-Петербургский государственный университет. - С. 19.

[2] Erhard Hinrichs, Marie Hinrichs, Thomas Zastrow, Gerhard Heyer, Volker Boehlke, Uwe Quasthoff, Helmut Schmid, Ulrich Heid, Fabienne Fritzinger, Alexander Siebert, and Jorg Didakowski.(2009) Weblicht: Web- based LRT services for German. In Workshop on linguistic processing pipelines, GSCL Jahrestagung, Potsdam.

[3] Мухамедшин, Д.Р., Сулейманов Д.Ш. (2018) Система корпус-менеджер: архитектура и модели корпусных данных Программные продукты и системы / Software & Systems 4 (31) – С. 6.

[4] Wynne, Martin (ed.) (2005), Developing Linguistic Corpora: A Guide to Good Practice, OxfordBooks. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/

[5] Jinyi Zhang, Tadahiro Matsumoto (2019) Corpus Augmentation for Neural Machine Translation with ChineseJapanese Parallel Corpora / Applied sciences (9), 2036.

[6] Zinsmeister, Heike, Erhard Hinrichs, Sandra Kübler and Andreas Witt (2009), “Linguistically annotated corpora: Quality assurance, reusability and sustainability,” in A. Lüdeling and M. Kytö (eds), Corpus Linguistics: An International Handbook, Vol. 1, Berlin: Mouton de Gruyter, pp. 759–76.;

[7] Гельбух А. Ф., Сидоров Г. О., Лавин-Вийа Э. Автоматический поиск и классификация однословных терминов в корпусе предметной области с использованием логарифмической меры сходства с неспециализированным корпусом / Компьютерная лингвистика и интеллектуальные технологии по материалам ежегодной Международной конференции «Диалог» (2010) Выпуск 9 (16) – С. 82.

[8] Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari: filol. fan. dok. (DSc)…diss. – Toshkent, 2021. – 220 b.

[9] Sandra Kubler, Heike Zinsmeister Corpus linguistics and linguistically annotated corpora New York: Bloomsbury, 2015. - P.321.

[10] Копотев М. Введение в корпусную лингвистику (учебное пособие) Прага, 2014. 264 с.

[11] Atkins B., Zampolli A. Computational approach to the lexicon Oxford, 1994, 494 p.

[12] N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.

[13] Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In: Abstract of the VI International Conference “Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018”, pp. 37–38, Tashkent, Uzbekistan (2018)

[14] Abduraxmonova, N. Z. "Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref." (2018).

[15] Abdurakhmonova N. The bases of automatic morphological analysis for machine translation. Izvestiya Kyrgyzskogo gosudarstvennogo tekhnicheskogo universiteta. 2016; 2 (38):12-7.

[16] Abdurakhmonova N, Tuliyev U. Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education. 2018(3):68.

[17] Abdurakhmonova N, Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL). 2019;6(1-2019):131-7.

[18] Abdurakhmonov N. Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation. Journal of Social Sciences and Humanities Research. 2017;5(03):89-100.

*УДК.004.934.5+004.032.26*
***[1]Кутдусова Е.П., [2]Прокопьев Н.А.***
*Казанский Федеральный Университет*
*Казань, Татарстан, Россия*
*[2]nikolai.prokopyev@gmail.com*

# СИНТЕЗ ТАТАРСКОЙ РЕЧИ ПРИ ПОМОЩИ ГЛУБОКОГО ОБУЧЕНИЯ НА ОСНОВЕ МОДЕЛИ VITS

**Аннотация.** Синтез речи является одной из важнейших задач речевой обработки и имеет широкое применение в современных информационных технологиях. Так как татарский язык относится к малоресурсным языкам, то есть мало обеспечен различными электронными базами, словарями и лингвистическими ресурсами, то наиболее оптимальным методом синтеза речи должен быть тот, который использует для реализации минимальные ресурсы.

Реализация такого метода оказалась возможной благодаря недавно разработанным системам нейронных сетей сквозного тестирования (end-to-end) для автоматического синтеза речи. Такие системы обладают рядом преимуществ, так как объединяют в себе сразу все модули стандартных систем, не требовательны предобработке исходных данных, что сокращает время обработки и объем требуемой памяти. Одной из эффективных и высоко оцененных end-to-end моделей для синтеза речи является архитектура Vits.

В данной работе представлена платформа для обучения модели Vits на наборах данных для татарского языка, состоящих из более чем 10000 пар текст-аудио различных дикторов. Проведены эксперименты по обучению модели при помощи библиотеки coqui-ai/TTS, выявлена наиболее эффективная конфигурация для обучения. Получены наилучшие варианты обученных моделей отдельно для дикторов мужских и женских голосов. Произведена оценка использования модели Vits для синтеза речи малоресурсных языков, в частности, для татарского языка.

Модель Vits показала достаточно высокие результаты для татарского языка уже по результатам начальных экспериментов по обучению. Однако, требуется дальнейшая разработка и обучение на большем объеме данных, иные варианты их группирования по дикторам и уточнение оценки MOS от большего количества экспертов.

**Ключевые слова:** синтез речи, глубокое обучение, модель end-to-end

*UDC.004.934.5+004.032.26*
***[1]Kutdusova E.,[2]Prokopyev N.***
*Kazan Federal University*
*Kazan, Tatarstan, Russia*
*[2]nikolai.prokopyev@gmail.com*

# TATAR SPEECH SYNTHESIS WITH DEEP LEARNING BASED ON VITS MODEL

**Abstract.** Speech synthesis is one of the most important problems of speech processing and is widely used in modern information technologies. Since Tatar language belongs to low-resource languages, that is, it is lacking of electronic databases, dictionaries and linguistic resources volume for standard machine learning techniques, the most optimal speech synthesis method should be the one that uses minimal resources for implementation.

The implementation of this method was made possible due to recently developed end-to-end neural network systems for automatic speech synthesis. Such systems have a number of advantages, since they combine all modules of standard speech synthesis systems, they do not require preprocessing of the training dataset, which reduces the processing time and the amount of memory required. One effective and highly scored end-to-end model for speech synthesis is the Vits architecture.

This paper presents a platform for training the Vits model on datasets for Tatar language, consisting of more than 10,000 text-audio pairs of various speakers. Experiments were carried out to train the model using coqui-ai/TTS library, the most effective configuration for training was identified. Best variants of the trained models are obtained separately for the male and female speaker voices. An assessment of trained Vits models was made for speech synthesis of low-resource languages, in particular, for Tatar language.

The Vits model showed quite good results for Tatar language already in the results of initial training experiments. However, further configuration and training on a larger amount of data, other options for grouping by speakers and refinement of MOS score from a larger number of experts are required.

**Keywords:** speech synthesis, deep learning, end-to-end model

## 1. Введение

Синтез речи является одной из важнейших задач речевой обработки и имеет широкое применение в современных информационных технологиях. Синтез речи по тексту является необходимым шагом в направлении более тесного общения человека с компьютером и может

потребоваться во всех случаях, когда получателем информации является человек.

Наиболее развитые лингвистически информационные технологии в основном связаны с теми языками, для которых доступны достаточно большие лингвистические электронные ресурсы, или же с языками, которые стали по какой-либо экономической или политической причине представлять интерес для мирового сообщества. Большая же часть языков сегодняшний день является малоресурсными. Так как татарский язык относится к малоресурсным языкам, то есть мало обеспечен различными электронными базами, словарями и лингвистическими ресурсами, то наиболее оптимальным методом синтеза речи должен быть тот, который использует для реализации минимальные ресурсы.

Реализация такого метода оказалась возможна благодаря сравнительно недавно разработанным системам нейронных сетей сквозного тестирования (end-to-end) для автоматического синтеза речи. Такие системы обладают рядом преимуществ, так как объединяют в себе сразу все модули стандартных систем, что сокращает время обработки и объем требуемой памяти.

## 2. Анализ моделей глубокого обучения для синтеза речи

На сегодняшний день можно выделить самые актуальные архитектуры нейронных сетей для синтеза речи, подробное описание которых указано в статье [Киреев, 2020]:

1. WaveNet – разработана компанией Google;
2. Deep Voice – состоящая из глубоких нейронных сетей;
3. Tacotron – end-to-end модель с модулем «внимания»;
4. Vits – end-to-end модель.

WaveNet – генеративная модель для синтеза необработанного звука. Это полностью сверточная нейронная сеть, в которой каждый новый образец зависит от предыдущего. Основным отличием этой архитектуры является причинно-следственные и расширенные (дилатационные) свертки. Оценка MOS (средняя экспертная оценка) для данной модели составляет 4.21 для английского языка, что является достаточно высоким результатом. Недостатки WaveNet:

- сложная система, которая требует подготовки большого объема размеченных текстов;
- требует дополнительные лингвистические функции (например информацию об ударении или основной частоте);
- вычислительно сложный синтез.

Система DeepVoice состоит из нескольких независимо обученных моделей, объединённых в вычислительный конвейер. Предварительно

обученные независимо друг от друга модели “Graphem-to-Phoneme” и “Segmentation” используются для обучения моделей “Audio-Synthesis”, “Duration Prediction” и “Fundamental Frequency”. Модель “Graphem-to-Phoneme” также используется при финальном синтезе речи. Частично синтез речи использует модифицированную архитектуру WaveNet, поэтому модель обладает тем же недостатком вычислительно сложного синтеза. Оценка MOS для данной модели достаточно низка и составляет 2.67 для английского.

Tacotron – модель типа end-to-end состоящая из кодера и декодера с механизмом внимания. Тренировка этой архитектуры достаточно проста, так как требует только пары текст-аудио. Генерация речи начинается с подачи необработанного текста на вход кодера. Первый уровень кодировщика – встраивание символов. Вложения (embedding) для текста передаются в двухслойную сеть для предобработки. Следующим этапом является применение модуля CBHG – рекуррентного блока одномерной свертки, изначально разработанного для задачи перевода. Оценка MOS для Tacotron составляет 3.82 для английского.

Vits – модель типа end-to-end, которая генерирует более естественный звук, чем иные модели, при этом не требует последовательного обучения промежуточных моделей или тонкой настройки. При помощи Vits сети обучаются с использованием оптимизатора AdamW, также используется авторегрессивная модель Tacotron 2 и неавторегрессивная модель Glow-TTS. Оценка MOS модели Vits, согласно статье [Jaehyeon, 2021] достаточно высока и составляет 4.43.

Поскольку наибольшую оценку MOS имеет модель Vits, а также она является нетребовательной к объему исходных данных для обучения, к техническому обеспечению и к тонкой настройке, было решено произвести экспериментальную реализацию синтеза татарской речи на основе именно этой модели.

## 3. Реализация синтеза речи на основе модели Vits

### 3.1. Архитектура Vits

В модели Vits в качестве целевой формы данных функции потерь используется mel-спектрограмма вместо необработанной формы сигнала, обозначаемая xmel. Восстанавливающая функция потерь для данной модели выглядит следующим образом:

$$L_{recon} = Y_{mel} - \hat{Y}_{mel}$$

Эту функцию можно рассматривать как оценку максимального правдоподобия в виде суммы распределения Лапласа для mel-спектрограммы.

Для вычисления сопоставительной оценки близости (выравнивания) входного текста и целевой речи, обозначаемой A, в модели Vits используется подход MAS (монотонный поиск выравнивания), который максимизирует вероятность параметризации данных при ограничении монотонности и отсутствия пропуском между словами в речи в связи с тем, что люди читают текст по порядку, не пропуская ни одного слова.

Для состязательного обучения в модели Vits добавлен дискриминатор D, который различает выходные данные, сгенерированные декодером G. Также применяется еще две функции потерь: метод наименьших квадратов для состязательного обучения и метод дополнительных потери при сопоставлении признаков для обучения генератора.

$$L_{adv}(D) = \mathbb{E}_{(y,z)}\Big[(D(y)-1)^2 + (D(G(z)))^2\Big],$$

$$L_{adv}(G) = \mathbb{E}_{z}\Big[(D(G(z))-1)^2\Big],$$

$$L_{fm}(G) = \mathbb{E}_{(y,z)}\Big[\sum_{l=1}^{T}\frac{1}{N_l}\|D^l(y) - D^l(G(z))\|_1\Big]$$

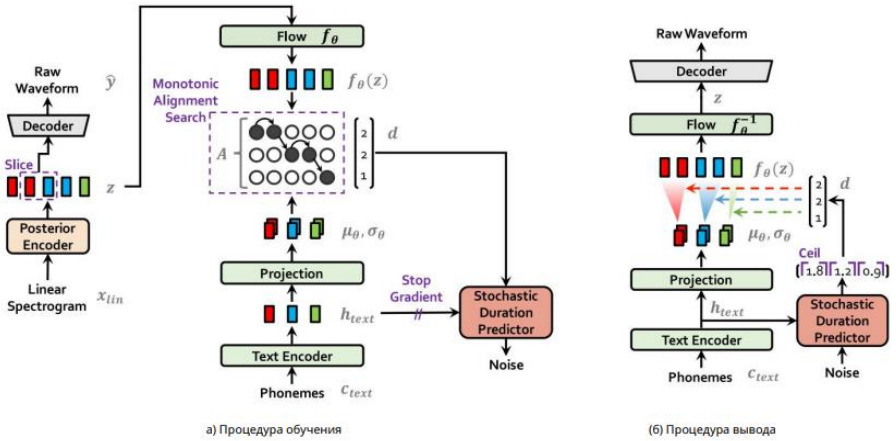



Рисунок 1. Архитектура Vits

Общая архитектура Vits (представлена на рисунке 1) состоит из апостериорного кодировщика, априорного кодировщика, декодера, дискриминатора и предиктора стохастической длительности. Апостериорный кодировщик и дискриминатор используются только для обучения, но не для синтеза. Сети обучаются с помощью оптимизатора AdamW [Loshchilov, 2019] с β1=0.8, β2=0.99 и уменьшением веса

λ=0.01. Начальная скорость обучения lr=0.0002, снижается каждую эпоху.

### 3.2. Подготовка модели и данных к обучению

Поскольку модель Vits относится к типу end-to-end моделей, исходные данные для ее обучения представляют собой пары из звуковых файлов записанной речи дикторов и соответствующих им текстов на татарском языке. Данные для модели требуется поместить в отдельные папки txt (для текстовых файлов) и wav48_silence_trimmed (для аудиофайлов). Внутри них файлы, в свою очередь, должны размещаться в отдельные папки для разных дикторов, при этом пары текст-аудио должны иметь одинаковые названия файлов. Для размещения данных в нужной форме был подготовлен скрипт на Python, общая схема представлена на рисунке 2.

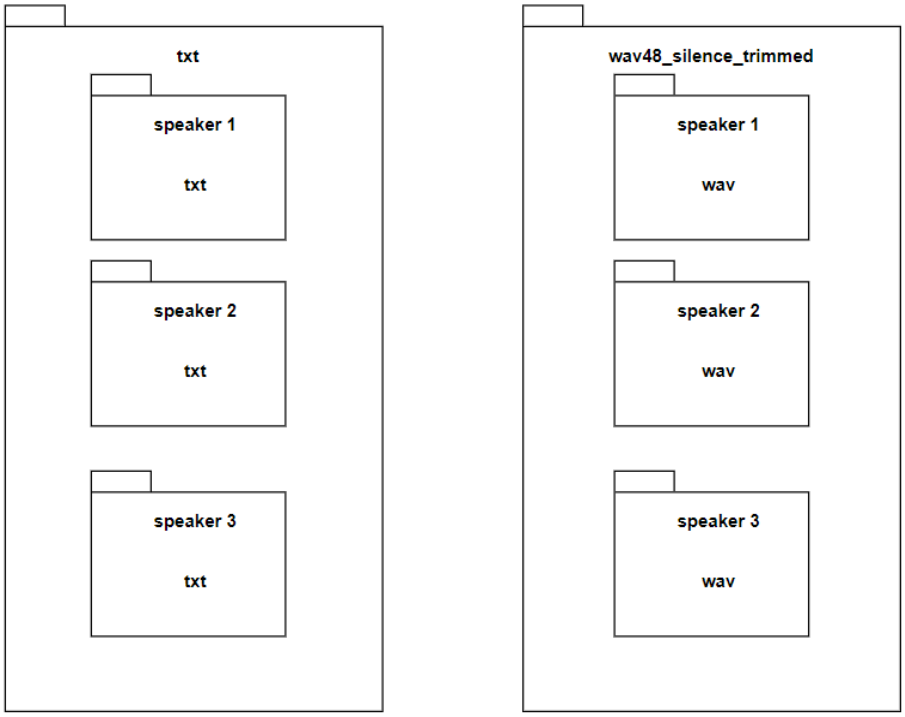



Рисунок 2. Файловая структура исходных данных для обучения

В дальнейшем, после экспериментов с обучением, дикторы были объединены в две группы – женские голоса и мужские голоса, что позволило улучшить результаты обучения. Всего таких пар текст-аудио для женских дикторов получилось 5131, а для мужских – 6997. Длительность аудиофайлов составляет от <1 секунды (одно слово) до 16 секунд (одно предложение). Общий объем данных – 1.2 GB для женских дикторов и 1.3 GB для мужских.

Для реализации платформы обучения нейросетевой моделиVits была использована библиотека coqui-ai/TTS[1], позволяющая обучать text-to-speech модели различных архитектур, в том числе поддерживает Vits.

Прежде всего стоит задача адаптации конфигурации модели для татарского языка. Для этого необходимо задать алфавит символов, используемых в текстах исходных данных, то есть татарский алфавит и символы пунктуации. Кроме того, необходимо настроить конфигурацию обучения, то есть задать такие параметры как: количество эпох, размер батча, размер батча для оценки в процессе обучения и т.п. После проведения экспериментов с обучением модели была выбрана наиболее эффективная конфигурация, основные параметры которой представлены далее:

```
{
 "model_param_stats": false,
 "run_eval": true,
 "mixed_precision": true,
 "epochs": 1000,
 "batch_size": 32,
 "eval_batch_size": 16,
 "lr": 0.001,
 "optimizer": "AdamW",
 "optimizer_params": {
 "betas": [0.8, 0.99],
 "eps": 1e-09,
 "weight_decay": 0.01
 },
 "cudnn_enable": true,
 "model": "vits",
 "use_noise_augment": false,
 "add_blank": true,
 "eval_split_max_size": 100,
 "eval_split_size": 10,
 "use_language_embedding": false
}
```

### 3.2. Оценочный тест результатов обучения

Во время обучения алгоритм библиотеки coqui-ai/TTS производит анализ, является ли модель лучшей, по сравнению с предыдущими моделями. Для этого при нахождении очередной лучшей модели алгоритм сохраняет ее. Такое сравнение происходит каждые 15 этапов в

[1] A deep learning toolkit for Text-to-Speech https://github.com/coqui-ai/TTS

эпохе. На выходе алгоритм выдает те параметры модели, которые были изменены в эпохе и вычисленные метрики. Основной вычисляемой оценкой лучшей модели является показатель функции потерь.

Наилучшая модель для исходных данных женских голосов была получена на эпохе 469 из 1000. Показатель функции потерь для нее составляет 29.4.

Наилучшая модель для исходных данных мужских голосов была получена на эпохе 370 из 1000. Показатель функции потерь для нее составляет 31.9.

Кроме вычисляемой метрики, основной, однако достаточно субъективной, метрикой для оценки моделей для синтеза речи является средняя экспертная оценка MOS, выражающаяся числовым десятичным значением от 1.0 до 5.0, вычисляемой как арифметическое среднее от субъективной оценки речи из синтезированного при помощи обученной модели аудиофайла, даваемой людьми.

Синтез речи с использованием обученной модели производится при помощи класса Synthesizer библиотеки coqui-ai/TTS, которому на вход при инициализации экземпляра подается обученная модель, конфигурационный файл и список дикторов. Далее, при помощи функции Synthesizer.tts(text, speaker), получающей на вход текст для синтеза и диктора, генерируется аудиофайл формата WAV, соответствующий поданному тексту.

Для оценки MOS сначала подавались на вход простые слова, впоследствии были использованы короткие фразы на татарском языке (не более 10 слов). Оценка MOS была дана авторами статьи и составила 3.2 для женского диктора и 2.8 для мужского диктора.

## 4. Заключение

Была реализована платформа для обучения и синтеза речи на основе модели Vits, получившей наивысшую оценку MOS среди наиболее актуальных моделей синтеза речи и обладающей при этом преимуществами низких требований к техническому обеспечению и к объему и форме данных для обучения. Модель Vits показала достаточно высокие результаты для татарского языка уже по результатам начальных экспериментов по обучению. Однако, требуется дальнейшая разработка и обучение на большем объеме данных, иные варианты их группирования по дикторам для доведения функции потерь как минимум до удовлетворительного в синтезе речи показателя 5.0. Кроме того, требуется уточнение оценки MOS от большего количества экспертов.

**Список литературы**

1. Киреев Н.С., Ильюшин Е.А. Обзор существующих алгоритмов преобразования текста в речь // International Journal of Open Information Technologies, Т.8, №7, 2020, с. 84-90.

2. Jaehyeon K., Jungil K., Juhee S. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech // ArXiv abs/2106.06103, 2021.

3. Loshchilov I., Hutter F. Decoupled Weight Decay Regularization // ArXiv abs/1711.05101, 2019.

# ЖАСАНДЫ ИНТЕЛЛЕКТ ЖҮЙЕЛЕРІНДЕ АҚПАРАТТЫ ҰСЫНУ МЕН ӨҢДЕУДІҢ СЕМИОТИКАЛЫҚ МОДЕЛДЕРІ

# СЕМИОТИЧЕСКИЕ МОДЕЛИ ПРЕДСТАВЛЕНИЯ И ОБРАБОТКИ ИНФОРМАЦИИ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

# SEMIOTIC MODELS OF INFORMATION REPRESENTATION AND PROCESSING IN ARTIFICIAL INTELLIGENCE SYSTEMS

*ӘОК 004.85*

***[1]Маңмұрын М. М., [2]Шәріпбай А.Ә.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]mmanmurynov@mail.ru, [2]sharalt@mail.ru*

## ШАБЛОН ҚҰЖАТ ҮЛГІЛЕРІНІҢ ФРЕЙМДІК МОДЕЛІ

**Аңдатпа.** Заман талабына сай қазіргі мемелекеттік және мемлекеттік емес мекемелерде электронды құжат айналым жүйесі енгізіліп, өзінің тиімділігін көрсетуде. Электрондық құжат айналым жүйелеріне жасанды интеллект әдістерін енгізу жұмыс өнімділігін арттырады. Білімдерді ұсынуды фреймдік құрылым арқылы бейнелеу әдісін құжаттарды ұқсастықтары мен айырмашылықтарына байланысты топтастырып жіктеуде, оларды іздеу, құжат деректемелерін біріктіруде қолдануға болады. Фреймдік жүйені тек қана құжатпен жұмыс жасауда ғана емес, басқа да кез келген объектілерді ұсынуда қолдануға болады.

**Түйін сөздер:** Фрейм, фреймдік модель, семантикалық желі, шаблон, құжат, электронды құжат айналымы, бұйрық құжат.

*УДК 004.85*

***[1]Манмурын М. М., [2]Шарипбай А. А.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]mmanmurynov@mail.ru, [2]sharalt@mail.ru*

## ФРЕЙМОВАЯ МОДЕЛЬ ШАБЛОНА ДОКУМЕНТА

**Аннотация.** В современных государственных и негосударственных

учреждениях внедрена система электронного документооборота. Внедрение методов искусственного интеллекта в системы электронного документооборота повышает производительность труда. Метод визуализации представления знаний с помощью фреймовой структуры можно использовать при группировке и классификации документов по сходству и различиям, их поиске, объединении реквизитов документа. Фреймовая система может использоваться не только при работе с документом, но и при представлении любых других объектов.

**Ключевые слова:** Фрейм, фреймовая модель, семантическая сеть, шаблон, документ, электронный документооборот, приказ документ.

*UDC 004.85*

***[1]Mangmuryn M., [2]Sharipbay A.***
*L. N. Gumilyov Eurasian National University*
*Nursultan, Kazakhstan*
*[1]mmanmurynov@mail.ru, [2]sharalt@mail.ru*

## THE FRAME MODEL OF THE DOCUMENT TEMPLATE

**Abstract.** An electronic document management system has been introduced in modern state and non-state institutions. The introduction of artificial intelligence methods into electronic document management systems increases labor productivity. The method of visualizing the representation of knowledge using a frame structure can be used when grouping and classifying documents by similarities and differences, searching for them, combining document details. The frame system can be used not only when working with a document, but also when presenting any other objects.

**Keywords:** Frame, frame model, semantic network, template, document, electronic document flow, order document.

### Кіріспе

Қазіргі ақпараттық технологияның қарыштап дамыған заманында әртүрлі өндіріс орындарында автомантандырылған жүйелер енгізіліп, адам қолымен атқарылатын жұмыстарды жеңілдететін жасанды интеллект әдістері кең етек алуда. Соның ішінде электронды құжат айналым жүйелерін автоматтандыру өзекті мәселе болып табылады. Қазіргі таңда көптеген ұйымдар қағаз түріндегі құжаттардан электронды құжаттар жүйесіне көшіп жатыр. Ол өз кезегінде экологиялық, материалдық тұрғыдан тиімді және құжаттармен жылдам жұмыс жасауды қамтамасыз етеді.

Электрондық құжат – о бұл Электрондық құжат айналымы жүйесін (ЭҚЖ) пайдалана отырып дайындалған, ЭҚЖ объектісі түрінде материалдық жеткізгіште тіркелген және деректемелермен жабдықталған, олардың көмегімен құжаттың орнын, жасалған уақытын және авторын сәйкестендіруге болатын құжат. ЭҚЖ электрондық құжаттармен жұмысты ұйымдастыруға мүмкіндік беретін программалық жасақтама болып табылады [Дымова, 2011, с. 21-25].

Білімдерді ұсынудың фреймдік моделі – кез келген ақпаратты абстрактілі-нақтылы түрде бейнелеуге мүмкіндік беретін жасанды интеллект ғылымының ажырамас бөлігі. Объектілерді фрейм және олардың жүйесі түрінде сипаттау адам ойлау жүйесінің айқындалған және компьютерлік жүйлерге түсінікті көрінісін береді.

**1. Шаблон құжат ұғымы**

Шаблон құжат – заңды құжаттың құрылымын, үлгісін және мазмұнын көрсетеді. Алдын ала дайындалған шаблон жоспарлау, құру, тексеру және сақтау процестерін қамтып өтеді. Шаблон арқылы дайын құжат форматын алу үшін кеңсе редакторлары, редакторларға арналған әзірлеуші кеңейтілімдері және сонымен қатар web-қосымшалар пен программалық қамтамалардың компоненттері қолданылуы мүмкін. Әрбір технологияның өзіндік мүмкіндіктері мен шектеулері болады. Кейбір шектік жағдайлар шаблон негізінде күткен нәтижедегідей құжат алуға мүмкіншілік жасай бермейді. Көбінесе құжатқа компания белгісін, колонтитул, қол қою операциялары орындалғанда нәтиже керек құжатқа сай келмей қалуы мүмкін. Шаблон құжаттар келесі мүмкіндіктер мен ерекшеліктерге ие:

– Әрбір заңды мағынаға ие болған кез келген құжат типінің астында бір ғана шаблон құжат жатады.

– Шаблондардың пакеттермен байланысын анықтауға, сондай-ақ жабық пакеттерге де қолдануға болады. Жабық пакет байланысындағы шаблондар бірге өңделеді және жіберу кезінде еш өзгертілместен жіберіледі.

– Шаблондар қолданылу ерекшелігіне қарай бір ретті және көп ретті болып бөлінеді [1]. Бір ретті шаблон ақырғы құжатта қажетті барлық құрылымдар мен мазмұнын қамтиды және оның негізінде бір ғана заіды құжат құруға болады. Ал көп ретті шаблоны арқылы бір типтегі көптеген құжат жасаса болады.

– Шаблонда құрылғалы жатқан нақты құжаттың мәндері мен өзгермелі деректерді сұрау баптаулары енгізіледі.

– Шаблонды құру және өңдеу үшін объектілерді құқықтары мен қатынау шектеулері қолданылады. Бұл шаблонды рұқсатсыз өзгертуден қорғайды.

**2. Білімді ұсынудың фреймдік моделі**

Білімді ұсыну – бұл адамдардың ақпаратты сақтау және өңдеу тәсілдерін жасанды интеллект көмегімен программамен ойлау арқылы білімді өңдеп, талдап, ұсыну әдістерін қамтиды.

Білімді ұсынудың негізгі мәселесі компьютерлік жүйелерді ақпаратты фомарьды қалыпта сақтау және оларды өңдеу болып табылады. Білімді ұсыну үшін фреймдік құрылымдар және семантикалық желілер секілді бейнелеу әдістері қолданылады. 60 жылдардан бастап ғылымда білім фреймі немесе жай фрейм деп аталатын түсініктер пайда бола бастады.

Фрейм түсінігін ең алғаш 1974 жылы американдық жасанды интеллект ғалымы М.Минским енгізген [3]. Фрейм дегеніміз адам ойлауының негізінде білімді ұсынудың объектілі-желілік құрылымы. Бұл кәдімгі программалау тілдеріндегі объектіге бағытталған программалау негіздемесіне ұқсас болып табылады. Яғни объектінің қасиеттері немесе айнымалылар және әдістері болатыны сияқты әрбір фреймнің өзіндік бірегей аты және қасиет жиынтықтары болады. Мысалы, машина фреймі түсі, жылдамдығы және т.б. атрибуттарды қамтуы мүмкін.

Фреймнің қайталанбайтын жеке аты және деректердің типін, мәнін, байланысын қамтитын біріне бірі тәуелсіз слоттардың ақырғы жиынынан тұратын ішкі құрылымы болады. Сондай-ақ, әр слот өзіндік деректер құрылымымен анықталады. Слоттың мәніне фреймнің осы қасиетін сипаттайтын мәлімет сәйкестендіріліп жазылады.

Бұл мәліметтер өз кезегінде сандар, жолдар, мәтіндер, функциялар мен формулалар, ұсыну ережелері, шарттар, басқа слоттарға сілтеме, программалар түрінде келуі мүмкін, яғни деректердің барлық түрін қамтиды. Сонымен қатар «матрешка принципіне» негізделіп слоттың мәні үшін пәс дәрежедегі слоттың жинағы түсуі мүмкін. Басқа фреймдердермен байланыс болуы үшін байланыс слотына әртүрлі байланыс ережелері мен шарттары жазылады.

Тұтас жағдайда, білімнің фреймдік құрылымы кең көлемде ақпарат қамтуы мүмкін, көбінесе келесідегідей атрибуттардан құралады.

Фрейм атауы. Ол фреймді жүйеде белгілеу үшін қайталанбайтын атау арқылы беріледі. Фрейм саны жағынан шектелмеген өзін сипаттайтын слоттардан құралады. Жобалаушы фреймдегі слоттардың саны өзі белгілейді немесе кей жағдайларда жүйелік функциялар слоттардың санын автоматты түрде анықтайды. Бұндай жүйелік слоттарға ата-ана слоты (IS-A), бала слоты, феймнің күнін белгілеу слоты, автор слоты және т.б. жатады.

Слот атауы. Ол фреймнің сипаттамасына сай түсінікті атқи ие және бір фрейм ішінде басқа слотармен бірдей болмауы керек. Жалпы жағдайда слоттың атауы белгілі бір семантикаға сай анықталады, сондай-ақ мәтін түрінде де берілуі мүмкін [3]. Мәселен, {слот атауы} = {«Абай жолы» романының басты кейіпкері}, {слот мәні} = {Абай}.

Мұрагерлік нұсқауыштар. Олар бас фреймнің слоттары туралы ақпараттың мұрегерлікпен берілетіндігі жөніндегі ақпаратты қамтиды. Мұрагерлік нұсқауыштар деректі-дерексіз иерархиялық фреймдік құрылымддарда қолданылады.

Нақты фреймдік жүйлерде мұрагерлік нұсқауыштарды келесідегідей белгілермен бейнелеуге болады: U (unique) – слоттың мәні мұрагерлікпен берілмейді, S (same) – слоттың мәні мұраға беріледі, R (range) – мұрагерлік слот мәндерінің белгілі бір интервалін қамтиды, O (override) – ағымдағы слот мәні жоқ болған жағдайда ғана жоғарғы деңгей слот мәнін мұраға алуы мүмкін, ал егер ағымдағы слот мәні бар болса ол бірегейлік сипатқа енеді. Яғни, O нұсқағышы U және S нұсқауыш қызметтерін бірге атқара алады.

Деректер типінің көрсеткіші. Ол слоттардың мәндерінің қандай деректер типіне жататындығын анықтайды.

Слот деректерінің типтеріне frame – фреймге сілтеме, real - нақты сандар жиыны, integer – бүтін сандар жиыны, boolean – логикалық өрнектер, text – мәтіндік жолдар, list - тізбектер, expression – өрнек-формулалар, lisp – байланысу ережелері, table – кестелер және т.б. жатқызуға болады.

Слот мәні. Ол келтірілген деректер типіне және мұрагерлік көрсеткішіне сәйкес келуі шарт.

Өздік процедуралар. Қандай да бір слотқа қатысты фугкциялар орындалған өздігінен орындалатын процедура болып табылады. Егер слоттың мәнін алу керек болған кезде if-needed процедурасы шақырылады. Ал егер слоттың мәні өзгертілетін болса if-added процедурасы іске асадаы. Сондай-ақ if-removed процедура түрі қандай да бір слот мәні жойылған кезде орындалады.

Бекітілген процедуралар. Объектіге бағытталған программалау тілдеріндегі әдістерге ұқсайтын Lisp тілінде слоттың мәніне қолданылатын арнайы процедуралар болады. Олар бір фреймнен басқа фреймге фунция нәтижелерін жібереді. Басқа фреймдерден мәлімет жіберілген кезде бекітілген процедуралар орындалады. Бекітілген және өздік процедуралар біртұтас жүйе ретінде қарастырылады.

**3. Бұйрық құжат және оның фреймдік моделі**

Бұйрық құжат – бұл өндіріс орны басшысы өз өкілеттіктері аясында шығарған және бағыныштылардың орындауы үшін міндетті әкімшілік сипаттағы құжат. Акционерлік қоғамдарда бұйрықтар шығаруға уәкілетті лауазымды тұлға бас директор болып саналады, медициналық мекемелерде – бас дәрігер, кәсіпорындарда – директор және т.б. Барлық қажетті деректемелерді қамтитын және тиісті тәртіппен куәландырылған құжаттар ғана заңды күшке ие болып табылады [4]. Бұйрық құжат бұйрық беруші мен сол бұйрықты орындаушы нысандар арасындағы процестерді заңды түрде бекітеді. Бұйрық ауызша немесе жазбаша, сондай-ақ қандай ді бір бағытта заң шығару арқылы іске асуы мүмкін.

Кез келген мекемелерде сол мекеме басшысы тарапынан жасалынған бұйрық құжатын атқаратын қызметіне қарай екі үлкен топқа бөліп қарастыруға болады: жеке құрам бойынша бұйықтар және атқару қызметі бойынша бұйрықтар.

Жеке құрам аясындағы бұйрықтар жұмыс беруші, яғни ұйым басшысы мен жұмысшы қызметкерлер арасындағы әртүрлі процестерді рәсімдеу үшін қолданылады. Бұндай бұйрық құжаттар арқылы жеке қызметкерге қатысты оны жұмысқа алу, қандай да бір қызметке тағайындау, жұмыстан шығару секілді мәселелер орындалады.

Негізгі атқарушы қызметіне байланысты бұйрық құжаттар ұйым басшысының қалауымен және ұйымның мақсаттары мен міндеттеріне байланысты шығарылуы мүмкін. Сонымен бірге белгілі бір ұйымға қатысты немесе жалпы жағдайларға байланысты жоғарғы деңгейдегі ұйымдар немесе мемелекеттік мекемелер шығарған бұйрықтарды да осы санаттағы құжаттарға жатқызса болады.

Кәсіпорынның ішкі жұмыстарына қатысты бұйрық құжаттар міндетті түрде көрсетілуі керек болған келесі реквизиттерден тұрады:

– Ұйымның немесе кәсіпорының толық атауы;

– Бұйрық аты (мысалы, жұмысқа қабылдау туралы);

– Жасалынған жері және күні;

– Бұйрық шығарушының аты-жөні, қолы;

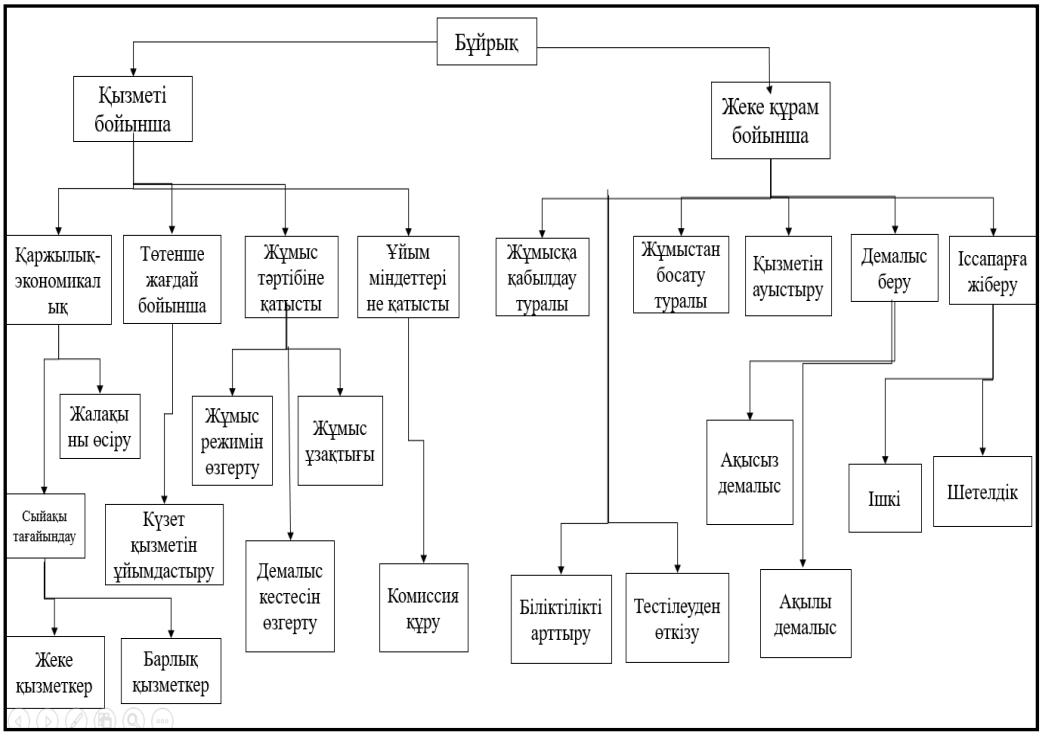
– Тіркеу нөмірі.

Аталған құжат реквизиттері күнделік жұмыс барысында кездесіп жүргендей бос орындармен беріліп, толтыру керектігін көрсетеді. Бұйрық құжаттың фреймдік құрылымын құру кезінде құжаттың міндетті реквизиттері фреймннің слоттарын береді. 1-кесте бұрық құжаттың фреймдік моделі көрсетілген.

Кесте 1. Бұйрық құжат фреймі

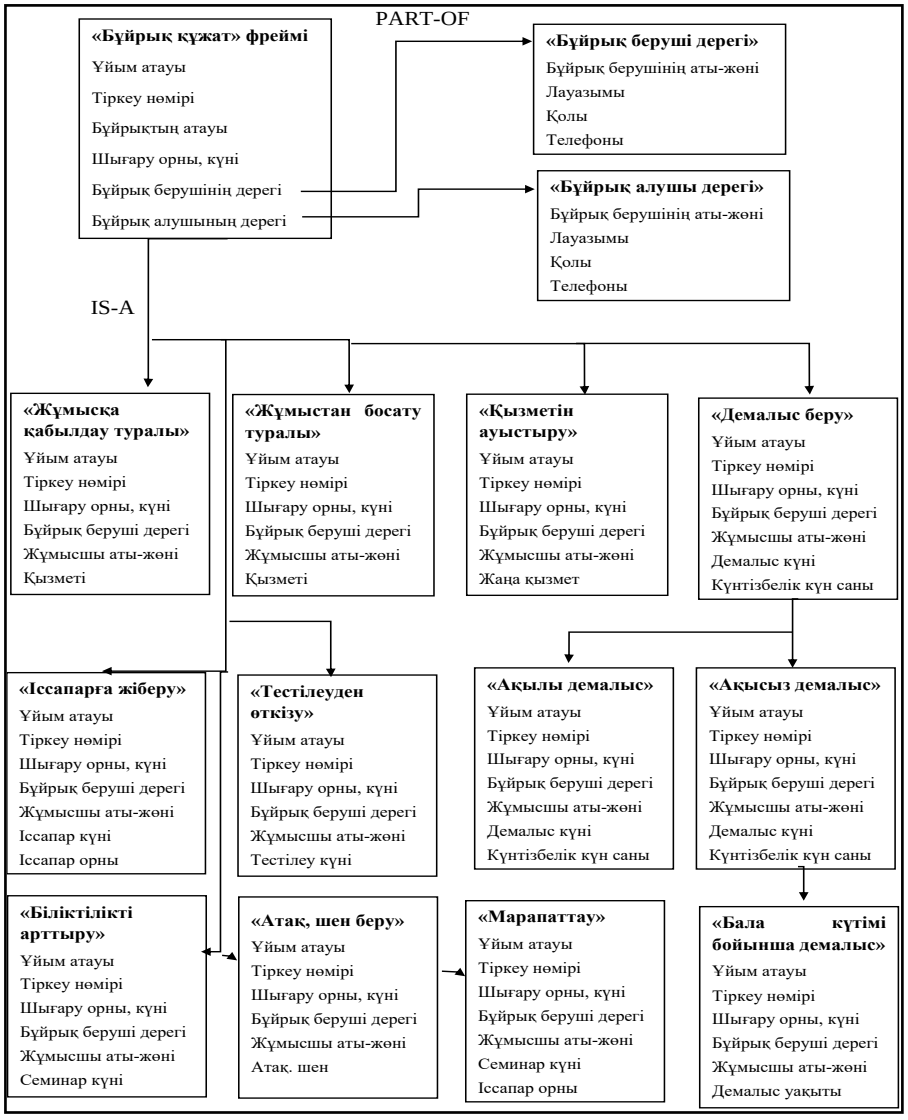
| Бұйрық құжат (Фрейм атауы) | |
|---|---|
| *Фрейм слоттары* | *Слоттардың мәні* |
| Ұйымның ресми толық атауы | |
| Бұйрықтың атауы | |
| Тіркеу нөмірі | |
| Шығарылған орны, күні | |
| Бұйрық беруші адамның аты-жөні, қолы | |
| Бұйрық берушінің қызметі | |
| Бұйрық алушының аты жөні | |

Жұмысшылардың жұмысқа орналасуынан бастап оның қызмет аясындағы еңбек қатынастары жеке құрамға қатысты бұйрықтарсыз шешімін таппайды. Бұйрықтардың семантикалық желісін құру олармен жұмыс жасауды, іздеу, өңдеу, ұқсастықтарына байланысты топтастыру процестерін жеңілдетеді. Бұйрықтардың бағытталған графтарға негізделген семантикалық желісін құру олардың фреймдік құрылымын жасаудағы бастама болып табылады. Бұйрықтардың семантикалық желісі 1-суретте көрсетілген.




Сурет 1 – Бұйрық құжаттың семантикалық желісі

Жеке құрам бойынша бұйрықтардың семантикалық желісі құру бұйрық құжаттардың ерекшеліктеріне байланысты жіктеу арқылы жүзе асыралады. Бұйрықтарды түрлеріне байланысты топтастырып жіктеу өз кезегінде олардың фреймдік құрылымын құруды оңтайландыруға септігін тигізеді. Белгілі бір топқа жататын бұйрықтардың өзіндік ортақ деректемелері болады. Осы ортақ деректемелер фрейм құру барысында бас фреймнің слоттары ретінде қолданылады. Әрбір фрейм арасында «болып табылады (IS-A)» байланысы қолданылады. Сонымен қатар, тұлғалардың жеке деректемелерін бөлек фрейм ретінде шығарып, оны бас фреймдегі осы слот атауына сәйкестендіріп, «бүтін бөлік» байланыс тәсілін қолданысқа енгізсе болады (1, 2-сурет).




Сурет 2 – Жеке құрам бойынша бұйрық құжат фреймдік желісі

Мысалы, 2-суретте көрсетілгендей «бұйрық құжат» бас фрейміндегі «Бұйрық берушінің дерегі» және «Бұйрық алушының дерегі» слоттары

PART_OF, яғни «бүтін бөлік» қатынасы арқылы слот атауларына сәйкесінше бөлек фреймдерде көрсетілген. Күрделі иерархиялық фреймдік желілерде бір фреймде тек сол фреймді нақты сипаттайтын қасиеттерді келтірген жөн болып табылады. Фрейм слоттарының қысқа әрі нұсқа болуы маңызды. Егер қолданушы туралы (аты-жөні, туған жері, туған күні, байланыс нөмірі, электронды поштасы және т.б.) ақпараты бар және тағы қосымша деректер қамтылған фреймде қолданушы туралы ақпаратты бөлек фреймге орналастырған тиімді. Өйткені қолданушы туралы ақпаратты өзгерту керек болған жағдайда тек бір фреймге ғана өзгеріс енгізіледі.

**Қорытынды**

Электронды құжат айналымы жүйелерінде фреймдік құрылымды қолдану өте тиімді. Құжаттарды олардың ерекшеліктеріне байланысты жіктеу көп істерді жеңілдетеді. Атап айтқанда, құжаттарды сұрыптау істерінде, оларды бірізділікке түсіріп, шаблон құжаттарды дайындау барысында, ұқсастықтары мен айырмашылықтарын айқындап топтастыру кезінде пайдасын тигізеді.

Қорыта келгенде, шаблон бойынша автоматты түрде құжат генерациялау жүйелерін құру және оны бизнеске ендіру қазіргі электронды-цифрлы дамудың жаңа талабы болып табылады.

**Әдебиеттер тізімі**

1 Документооборот шаблонов // https://api docs.diadoc.ru/ru/ latest/docflows/ TemplateDocflow.html. Қаралған күні: 13.03.2022 ж.

2 Дымова М. В. Обзор систем электронного документооборота. – 2011. – № 3. – С. 21-25

3 Приказ: виды и особенности оформления // https://www.kdelo.ru/art/ 385639-prikaz-vidy-i-osobennosti-oformleniya. Қаралған күні: 18.03.2022 ж.

4 Фреймовая модель представления знаний // https://itteach.ru/ predstavlenie- znaniy/ freymovaya-model-predstavleniya-znaniy. Қаралған күні: 15.03.2022 ж.

*UDC 004*

***[1]Aktaeva A., [2]Kubigenova A., [3] Esmagambetova G.***
*[1]Kokshetau University named after Sh.Ualikhanov*
*Kokshetau, Kazakhstan*
*[2]Kazakh Agrotechnical University named after S.Seifullin*
*Nur-Sultan, Kazakhstan*
*[3]Mongolian University of Science and Technology*
*Ulaanbaatar, Mongolia*
*[1]aaktaewa@list.ru, [2]akku_kubigenova@mail.ru, [3]esc.gal@mail.ru*

# SEMANTIC ASPECTS OF SECURITY IN BLOCKCHAIN TECHNOLOGIES: CRYPTOSEMANTICS

**Abstract.** We consider blockchain as a new cryptographic primitive, which is a special database in which records are linearly ordered, and permissible requests can be of only two types: requests to read records (anyone) and requests to add a record to the end of the database (user, fulfilling certain conditions).

Although being a part of general cryptology, cryptosemantics is characterized by a number of fundamental differences from classical cryptography and relies on its own axiomatic basis. The main attention in this work is focused on cryptosemantics—a new direction in ensuring the protection of information resources of intelligent systems from their unauthorized use.

At the same time, the blockchain must meet the requirements of survivability and inviolability, guaranteeing the addition of any record, as long as it is equipped with the necessary confirmation of the right to this action and the "eternal preservation" of the once added record.

Blockchain security and privacy analysis is a vast area of research, and for any blockchain, the key evaluation parameter is how well the security and privacy conditions meet the requirements of the blockchain. Privacy can be defined as data privacy and user privacy (anonymity).

Cryptosemantics is based on a class of formal reversible transformations of the semantics of classified information: data privacy and user privacy.

The article discusses the problem of justifying the existence of a blockchain in various models that allow the use of a cryptographic primitive to protect information. In this work, the main focus is on blockchains that meet the crypto-semantic paradigm, in the design of which cryptographic hash functions are significantly used.

***Keywords:*** *blockchain, cryptosemantics, semantic filters, information security, cryptographic primitive.*

*УДК 004*

***[1]Актаева А. У., [2]Кубигенова А. Т., [3]Есмагамбетова Г. К.***
*[1]Кокшетауский университет им. Ш. Уалиханова*
*Кокшетау, Казахстан*
*[2]Казахский агротехнический университет им. С. Сейфуллина*
*Нур-Султан, Казахстан*
*[3]Монгольский университет науки и технологии*
*Улан-батор, Монголия*
*[1]aaktaewa@list.ru, [2]akku_kubigenova@mail.ru, [3]esc.gal@mail.ru*

# СЕМАНТИЧЕСКИЕ АСПЕКТЫ БЕЗОПАСНОСТИ В ТЕХНОЛОГИЯХ БЛОКЧЕЙНА: КРИПТОСЕМАНТИКА

**Аннотация.** В работе блокчейн рассматривается как новый криптографический примитив, представляющий собой особую базу данных, записи в которой линейно упорядочены, а допустимые запросы к ней могут быть лишь двух видов — запросы на чтение записей (любым желающим) и на добавление записи в конец базы (пользователем, выполнившим определённые условия).

Являясь разделом общей криптологии, криптосемантика характеризуется рядом принципиальных отличий от классической криптографии и опирается на собственный аксиоматический базис. Основное внимание в настоящей работе сконцентрировано на криптосемантике – новом направлении обеспечения защищенности информационных ресурсов интеллектуальных систем от их несанкционированного использования.

***Ключевые слова*:** блокчейн, криптосемантика, семантические фильтры, защита информации, криптографический примитив.

## 1. Introduction

Blockchain is seen as a forward-looking and powerful technology, but it still faces many outstanding research challenges. A blockchain is a distributed register that supports an ever-growing list of data records confirmed by all participants. The blockchain structure is a special database consisting of linearly ordered records and allowing only two types of queries to:

1. Read any record-accessible to any user;

2. Add a record to the end of the database - the right of users to this action is determined by the method of implementing the blockchain and a specific application.

The data are recorded in the public registry as valid transaction blocks,and where the public registry is shared and accessible to all participants.The concept of blockchain from a mathematical point of view, most adequately fits in to a number of cryptographic primitives.

The gradual introduction of new cryptographical functions in to existing blockchain's is one of the main tasks to increase their safety: confidentiality, scalability,control of keys, and smart contracts,attack analysis, etc. These problems arise in connection with the structure of the network, and the basic mechanisms of Cryptologic schemes used in blockchain technologies. In order to clear up these problems and find improved solutions, many cryptological concepts must be carefully studied and applied, such as signature schemes,evidence of zero-knowledge, commitment protocols, as well as semantic and pragmatic aspects of ensuring information security: cryptosemantics.

The most general research in blockchain focuses on finding and identifying improvements to existing processes and procedures, mainly in industries that depend on intermediaries, including banking,finance,real estate,insurance,legal system procedures,and health care(seeTable1).

**Table 1**
**Blockchain classification**

| Blockchain Type | Scope of application | Anonymity | Scalability | Problems |
|---|---|---|---|---|
| Permissionless Public | Decentralized P2P Network | High | Low | Privacy, Scalability |
| Permissioned Public | Decentralized Organizations | High | Medium | Privacy, Centralization |
| Permissionless Private | Networks in-house | Medium | Medium | Consensus, Scalability |
| Permissioned Private | Restricted Access Organizational Registries | Low | High | Consensus, Centralisation |

The research on business innovation through blockchain presents some business applications using blockchain and their implementation [5]. To securely implement a blockchain business application, it is necessary to:

a)The correct selection of the appropriate cryptography model and algorithm used in their respective solution to meet business requirements;

b) A knowledge systematization that provides a comprehensive understanding of the existing cryptography techniques related to blockchain.

## 2. Fundamentals of cryptosemantics in the blockchain technology

Information security is defined as three components:

1. Confidentiality;
2. Integrity;
3. Accessibility.

In the general context, (1) privacy is a set of rules that restrict access to information; (2) integrity is a guarantee that information is valid and accurate; (3) availability is a guarantee of reliable access to information by authorized users.

Analysis of security and privacy issues is an extensive field of research for any blockchain. The key assessment parameter is how well the security and privacy conditions meet the requirements of the blockchain. Privacy can be defined as data privacy and user anonymity.

Cryptosemantics is based on a class of formal reversible transformations of the semantics of classified information such as a data privacy and user privacy. Semantic ciphers, in contrast to classical cryptographic ciphers according to K. Shannon, are determined by a structural-statistical model of a set of open messages and associated with transformations of their formal semiotic-syntactic representation in the model of J. von Neumann [1, 2].

The principal difference between cryptosemantics and cryptography is the use of phenomenology and models of communication information, fundamentally different from the classical interpretation by K. Shannon, defined on the a priori specified structural statistical model of the set of open messages and associated with the transformations of their formal semiotic-syntactic representation in the model by J. von Neumann. In the subject area of general cryptosemantics, it is included in the list of its essentially phenomenological distinguishable directions, such as cryptography, steganography, etc. (See Fig. 1).

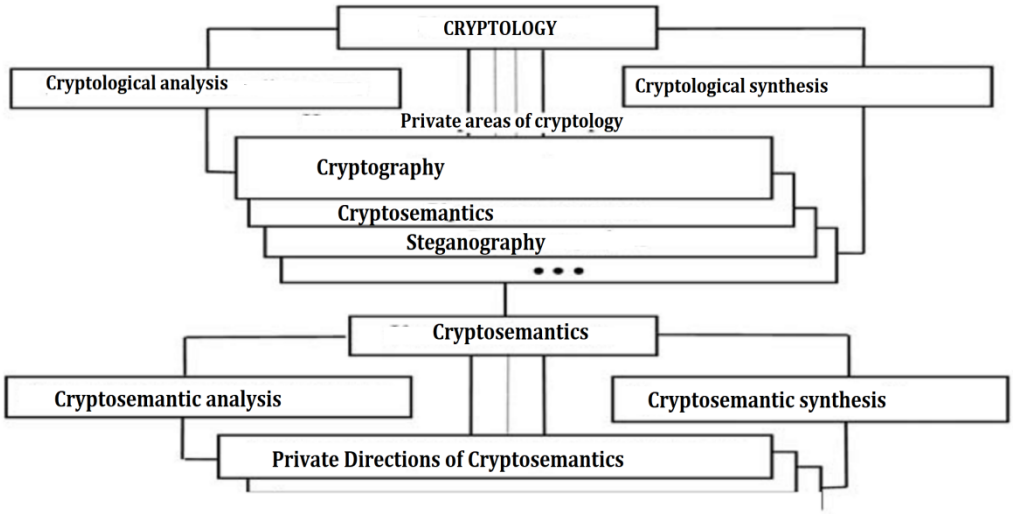



**Figure 1:** General structure of cryptology

The secondary, generated term "cryptosemantics" ("secret meaning, hidden meaning") is synthesized (by analogy with cryptography) from the primary word forms of the ancient Greek language:

1. κρυπτος - (when used in conjunction with the same - root κρυπτον, κρυπτοω) - a secret, hidden, secretive, etc.;
2. σημαντικος - denoting, designation, value (meaning), etc. [1.2].

The problems of the study of the semantic and axiological aspects of information, and the search for new cipher classes other than classical, traditional, belong to fundamental and methodological research in the field of ensuring the safety and privacy of information subordinate to various layers of the blockchain.

Some cryptographical mechanisms for security and privacy of information subordinated to different layers of the blockchain, as a cryptographic primitive is defined by a quadruple of the following objects:

1. E - the set of valid entries representing bit strings of finite length, $E \subseteq \{0,1\}*$;
2. P -an efficiently computable triplet predicate;
3. R - an efficient algorithm that performs read requests on records in blocks;
4. S - an efficient algorithm that performs queries to add records to the end of the database.

The **R**-record given in the query is added when the condition P = 1 (state; r; proof) , where the state-status of the blockchain at the time the request is received (in general, the state means the entire chain of blocks), proof - dependent on the type of blockchain additional information, which serves as a confirmation of the user's right to add this record.

Semantic analysis of existing cryptographic systems shows that the σ transformations used in them also provide a solution to the problem of secretly changing the semantics of messages (See Fig. 2).

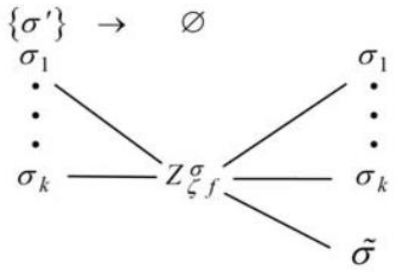



**Figure 2:** Structure of elements of the semantics model

An encryption scheme is the process of coding part of the information so that only the authorized parties can access it. The encryption scheme can be used to ensure the privacy of blockchain data by encryption.There are many encryption schemes that can be used on the blockchain. Searching and computing according to encrypted data is a big challenge.

According to the methodology of existing sets of blockchain data encryption schemes, the implementation of a perfect cipher in some sense of cryptosemantics can be reduced to the choice of a transformation $\boldsymbol{f : \sigma \to \sigma^{\wedge}}$, providing $Z_{\zeta f}^{\sigma} \equiv \varnothing$ in the collective knowledge model of the subject domain of communication. For any information systems $\boldsymbol{\zeta}$ except for some finite subset of the secret communication system $\boldsymbol{\{\xi\} \subset \{\zeta\}}$ . Here $\boldsymbol{\sigma}$ is the original semiotic structure (open message), and $\boldsymbol{\sigma^{\wedge}}$ is the resulting structure transformed by methods of cryptosemantics [1,2].

In particular, the representation of the transformed $\boldsymbol{\sigma^{\wedge}}$ as a random equivariant sequence of finite alphabetic characters of arbitrary length generates a completely "meaningless", by well-known, well-defined criteria, a message when the semantics cannot be identified.

Some of these methods, such as searchable encryption to search through encrypted data in the cloud, are already used in the allowed blockchain. Full homomorphic encryption and functional encryption can also be used to compute the encrypted data on the blockchain [5].

To ensure simultaneous confidentiality and authenticity, the blockchain can use authenticated encryption. In authenticated encryption, two peer-to-peer nodes establish a connection, both share their public keys and compute a shared secret that is used as a symmetric key for an authenticated encryption algorithm.

Broadcast encryption can be used in the blockchain to ensure the anonymity of recipient nodes,which enables the use of blockchain to ensure

the accessibility and accountability of the Internet of Things[5].This is because every user in the group receives an encrypted message, although only users with the appropriate permission or key can decrypt it.

Blockchain is supported by a peer-to -peer network (P2P). A P2P network is an overlay network created over the Internet. The P2P blockchain network can be modelled as structured, unstructured, or hybrid based on several parameters, such as consensus mechanism and blockchain type. Regardless of the network representation, the blockchain must quickly distribute the newly generated block, so that the overall blockchain representation remains unchanged. Therefore, a synchronization protocol is needed, and a routing protocol may or may not be needed.

A routing protocol is used in a traditional P2P network to route information through multiple hops. However, many blockchains (such as bit coin) do not require routing because a peer-to-peer node can only receive information through one jump,so the routing table is not supported. The P2P network can have a uniform or hierarchical structure to build a random graph between nodes.

This graph is not fully connected, butinor der tore cei veall message sand main taina registry; each node provides al is to address nodes.Thus,if any peer node propagates a message in the network,eventually all peer nodes receive it through their available connections.

In an unstructured network, techniques such as flooding and random roaming are used to establish new connections with nodes.In an unstructured network,nodes can leave and rejoin at any time.This can be used by an opponent who can join and see messages that are moved through the network,and can also perform source substitution, reassignment,or message embedding.

Blockchain can as well use an integrated P2P network in which nodes are organized according to a certain topology, and thus any resource/information search becomes easier. In this integrated P2Pnetwork,each host is assigned an identifier to route messages in a more accessible way.Each node also maintains a routing table.

Thus,if any peer node propagates a message in the network, eventually all peer nodes receive it through their available connections.In an unstructured network, techniques such as flooding and random roaming are used to establish new connections with nodes.In an unstructured network, nodes can leave and rejoin at any time.This can be used by an opponent who can join and see messages that are moved through the network,and can perform source substitution, reassignment, or message embedding.

Also, blockchain can use an integrated P2P network in which nodes are organized according to a certain topology, and thus any resource/information

search becomes easier. In this integrated P2P network,each host is assigned an identifier to route messages in a more accessible way. Also, each node maintains a routing table.

In addition to these problems,the opponent can carry out several attacks on the P2P network, where some of the main attacks areas follows:

1. Netsplit -Attack(Eclipse):The enemy monopolizes all node connections and separates that node from the entire network.In addition, the node cannot participate in the consensus protocol or authentication protocol, and this causes in consistency in the network.
2. Routing-Attack:The enemy isolates a number of members from the blockchain network, thus delaying the expansion of the block over the network.
3. DDOS-Attacks: the adversary drains the resources of the network, and targets truthful nodes so that honest nodes do not receive the services or information they should receive[5].

A structured P2P network maintains a distributed hash table(DHT) that's to repairs (key,value) corresponding to nodes that help in there source discovery process.

However, most blockchain networks are not structured. Moreover,if the blockchain is public and there are no restrictions on joining or leaving the network, many possible attacks can occur.

Thus,blockchainsecurityishighlydependentonnetworkarchitecture.Apropagationdelayorsynchronizationprobleminap2Pnetworkcanaffecttheblockchain'sconsensusprotocol,leadingtoaninconsistentglobalrepresentationoftheblockchain.

## 3. Conclusion

The introduction of the cipher model of cryptosemantics allows an on standard position on the basic objects used in classical cryptography.Given the abundance of possible forms of representation of semantics and focusing on the final physical capacity of communication channels,i.e., the representation of communication forms by a set of well-defined code semiotic structures{σ}.

The methodology of cryptosemantics of its achievement differs significantly from the methodology of solving classical cryptographic problems. The purpose of cryptosemantics is to restore the true content (the original meaning of the primary communication message) of incoming, possibly classified information. As a result, in cryptosemantics, its own axiomatic is formed-the terminological system of the basic concepts of the subject area is formed as necessary.

The study shows that cryptography models are based on the principle of independence of the message model from the characteristic individual properties of subscribers of the secret communication system, while message cryptosemantics models are fundamentally related to modelling subjective (semantic-pragmatic) characteristics of the blockchain data.

**References**

[1] A.E.Baranovich Semantic Aspects of Information Security: Concentration of Knowledge // Bulletin of RSGU, issue 13 (75), Ser. "Informatics. Information protection. Mathematics", 38-58 pp., 2011

[2] A.E. Baranovich On systematization of the axiomatic apparatus of the "Artificial Intelligence" subject area. // Intelligent systems, vol. 14, issue 1-4, 5-34 pp., 2010

[3] D. Boneh, J. Bonneau and et al. Verifiable delay functions // Advances in Cryptology - CRYPTO 2018: Springer International Publishing, 757-788 pp., 2018

[4] R. Henry, A. Herzberg, and A. Keith, Blockchain access privacy: challenges and directions // IEEE Security Privacy, vol.16, issue 4, 38-45 pp., 2018

[5] Iuon-Chang Lin, Tzu-Chun Liao A Survey of Blockchain Security Issues and Challenges // International Journal of Network Security, vol. 19, issue 5. - ISSN 1816-353X.— doi:10.6633/ijns.201709.19(5).01

[6] R. Pass, L.Siman, A.Shelat Analysis of blockchain protocol in asynchronous networks // Advances in Cryptology- EUROCRYPT'17 : Springer, 643-673 pp., 2017

[7] N.P.Varnovsky On definitions of cryptographic persistent hash functions. 1998. Manuscript [in Russian].

[8] I. Mironov Hash functions from Merkle-Damgard to Shoup // Advances in Cryptology - EUROCRYPT '01, vol. 2045 : Springer, 166-181 pp., 2001.

[9] C.E.Shannon, W.A.Weaver Mathematical theory of communication. University of Illinois Press, Urbana, 1949.

[10] Bitcoin to developers: https://bitcoin.org/ru/bitcoin-for-developers

[11] Blockchain and Bitcoin in Russia. Electronic Journal.: https://cryptorussia.ru/ zametki/ zakonoproekt-o-kriptovalyutah-glavnye-tezisy

[12] Pools for mining Bitcoin (BTC) cryptocurrency: https://bitmakler.com/ mining_Bitcoin-BTC_pools

# ТҮРКІТІЛДЕС ИНТЕРНЕТ-РЕСУРСТАР, ОНТОЛОГИЯЛАР, ТЕЗАУРУСТАР ЖӘНЕ СӨЗДІКТЕР

# ТЮРКОЯЗЫЧНЫЕ ИНТЕРНЕТ-РЕСУРСЫ, ОНТОЛОГИИ, ТЕЗАУРУСЫ И СЛОВАРИ

# TURKIC INTERNET RESOURCES, ONTOLOGIES, THESAURI AND DICTIONARIES

*УДК 004.891.2*

***Бурнашев Р.А., Галимов М.Р.***
*Институт прикладной семиотики*
*Академии Наук Республики Татарстан*
*Казань, Татарстан, Россия*
*r.burnashev@inbox.ru*

## ПОСТРОЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ ЯЗЫКОВЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ С ИСПОЛЬЗОВАНИЕМ ГЕОИНФОРМАЦИОННЫХ СИСТЕМ

**Аннотация.** В настоящее время диалектолог взаимодействует с большим набором языковых информационных ресурсов и баз данных. Из-за большого объема данных, специалисту сложно найти необходимую информацию в процессе профессиональной деятельности. В связи с этим, одними из основных первоочередных задач является создание современных программных приложений с элементами искусственного интеллекта, связанных геопространственными данными.

Нашим исследованием мы стремимся внести вклад в развитие интеллектуальных языковых информационных ресурсов с использованием геоинформационных систем и технологий.

В статье представлены результаты исследований по обработке знаний и последующией визуализации географических структур данных.

Для визауализации данных была использована программная библиотека Folium. Для анализа и обработке данных библиотека Pandas.

**Ключевые слова:** база знаний, геоинформационная система, нечёткая логика, folium

*UDC 004.891.2*
***Burnashev R.A., Galimov M.R.***
*Institute of Applied Semiotics of the*
*Academy of Sciences of Tatarstan Republic*
*Kazan, Russia*
*r.burnashev@inbox.ru*

# BUILDING INTELLIGENT LANGUAGE INFORMATION RESOURCES USING GEOINFORMATION SYSTEM

**Abstract.** Currently, a dialectologist interacts with a large set of language information resources and databases. Due to the large amount of data, it is difficult for a specialist to find the necessary information in the course of professional activity. In this regard, one of the main priorities is the creation of modern software applications with elements of artificial intelligence connected by geospatial data.

With our research, we strive to contribute to the development of intelligent language information resources using geoinformation systems and technologies.

The article presents the results of research on knowledge processing and subsequent visualization of geographical data structures.

The Folium software library was used to visualize the data. For data analysis and processing, the Pandas library.

**Keywords:** knowledge base, geoinformation system, fuzzy logic, folium

## 1. Введение

Из большого количества разнородной информации, поступающей к нам сегодня из разных источников, порой бывает трудно выделить главное и сделать правильное решение. Часто у диалектолога возникает сложность сбора и обработки входной информации, поступающих в режиме реального времени. Использование на практике современных средств геоинформационных технологий по считыванию и обработке различий диалектов народов с последующим картографированием является актуальность задачей. Использование лингвогеографического метода позволит повысить эффективность принятий решений при проведении научных экспедиций и исследований.

Графическая информация воспринимается в несколько раз быстрее (Жданова, Белых, 2014), нежели текстовая.

При разработке программного приложения были использованы программные библиотеки (Django, Pandas, Folium, os, NumPy и др.) языка программирования Python.

Для работы с интерактивными картами была использована библиотека Folium **(**Fedutinov, 2019**).**

**Folium** - библиотека Python, которая помогает разрабатывать карты Leaflet.js (язык программирования JavaScript). С помощью неё можно манипулировать данными в Python и визуализировать объекты их на карте.

Folium позволяет легко визуализировать данные, которые были обработаны в Python на интерактивной карте Leaflet. Он позволяет картографировать данные, а также передавать векторные/ растровые/ HTML элементы в качестве маркеров на карте.

В библиотеке имеется ряд встроенных наборов карт из OpenStreetMap, Mapbox и Stamen (Рис. 1.). Библиотека поддерживает обработку файлов формата представления различных структур географических данных (GeoJSON и TopoJSON), изображения и видео.


Рис. 1. Добавление интерактивных элементов на карту

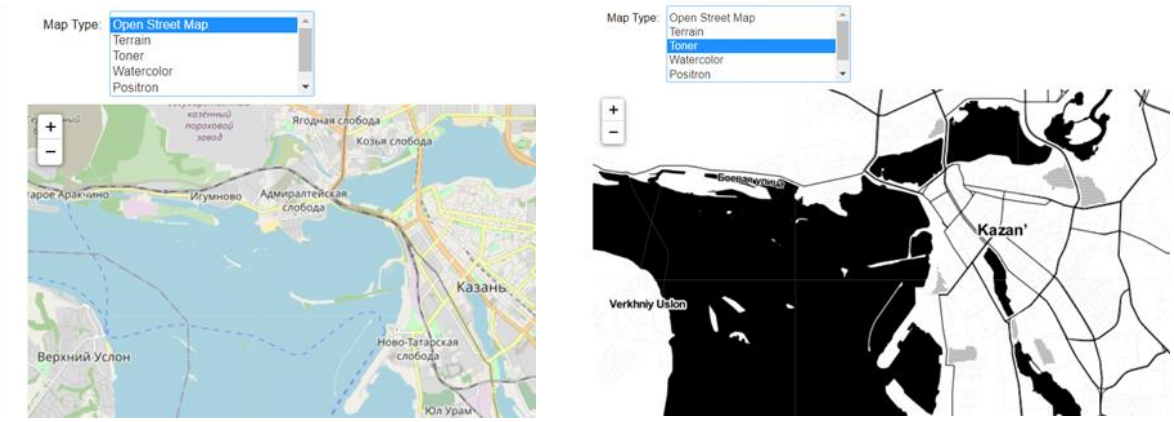



Рис. 2. Добавление интерактивных стилей(OpenStreet Map, Toner)

**2. Обработка и анализ геопространственных данных**

Для обработки и анализа данных была использована библиотека pandas (Ильичев & Юрик, 2020). Pandas — программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временными рядами.

Библиотека включает в себя объект DataFrame (двумерный массив), похожий на таблицу/лист Excel (данные из файла можно загрузить с помощью команды pandas.read_csv(‘наименование файла’)).

С помощью программной библиотеки можно проводить такие же манипуляции с данными: объединять список, сортировать по необходимому признаку, производить анализ, формировать отчёты и др.

Ниже (Рис. 3) приведен пример данных полученных с файла для последующей интеграции в геоинформационную систему диалектолога.

```
In [322]: df
```

Out[322]:

| | word | latitude | longitude | country | city | icon_num |
|---|---|---|---|---|---|---|
| 0 | җан | 55.78 | 49.13 | Russia | Kazan | 1 |
| 1 | җылы | 55.90 | 49.13 | Russia | Kazan | 2 |
| 2 | җирле | 55.95 | 49.13 | Russia | Kazan | 3 |
| 3 | җиһаз | 55.78 | 48.40 | Russia | Innopolis | 4 |
| 4 | бөжра | 56.40 | 49.13 | Russia | Kazan | 5 |
| 5 | җөмлә | 55.12 | 49.13 | Russia | Kazan | 6 |

Рис. 3. Фрагмент файла после чтения и обработки геопространственных данных

Ниже приведен программный код интеграции данных с носителя на карту Folium (Рис. 4.):

```
In [313]: # plot dialects locations
for (index, row) in df.iterrows():
    folium.Marker(location=[row.loc['latitude'], row.loc['longitude']],
                  popup=row.loc['word'] + ' ' + row.loc['city'] + ' ' + row.loc['country'],
                  tooltip='click').add_to(folium_map)

# display map
folium_map
```

Рис. 4. Фрагмент кода доблавления данных на карту

Для работы с графиками и отчётами была использована программная библиотека Matplotlib**.** Matplotlib — библиотека на языке

программирования Python для визуализации данных двумерной (2D) графикой. Обработанные графики могут быть использованы в качестве иллюстраций при формировании отчёта (статистики).

GeoJSON формат файла предназначен для хранения географических структур данных, основан на JSON.

GeoJSON поддерживает следующие типы объектов:

- Point (в том числе адреса и местоположения).
- Line string (в том числе, улицы, шоссе, границы).
- Polygon (в том числе страны, провинции и земельные участки).
- Составные объекты типов point, line string или polygon.

Ниже приведён листинг программного кода файла GeoJSON:

```
{"type":"FeatureCollection","features":
[{"type":"Feature","properties":{},
 "geometry":{"type":"Polygon",
 "coordinates":
 [[[46.120605,57.01192],
 [48.537598,57.569959],
 [51.61377,56.856072],
 [52.536621,54.782834],
 ……..,
 [44.736328,55.110943],
 [46.120605,57.01192]]]}}]}
```

Ниже представлен прототип системы (Рис. 4.) созданый в результате исследований по обработке знаний и последующией визуализации географических структур данных.

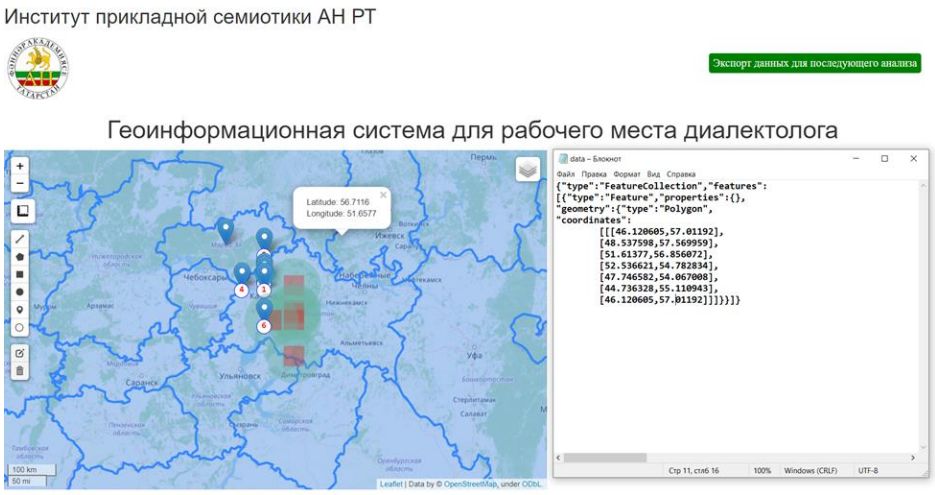



Рис. 4. Основная форма с интерактивными элементами на карте

## 3. Заключение

С использованием современных средств визуализации данных можно представить информацию максимально доступной для восприятия специалиста предметной области.

Для интеллектуализации системы планируется дальнейшая интеграция экспертной системы в оболочку геоинформационной системы.

Разработанный проект в дальнейшем планируется интегрировать в структуру портала "Тюркская Морфема" (http://modmorph.turklang.net/).

## Список литературы

1. Жданова Е. А., Белых А. А. Географические информационные системы в лингвистических исследованиях // Интеллектуальные системы в производстве. Ижевск, 2014. № 2 (24). С. 169–174.

2. Fedutinov K.A., Structuring environmental information using geographic information technologies. Modeling, optimization and information technology. 2019;7(4). Available from: https://moit.vivt.ru/wp-content/uploads/2019/11/Fedutinov_4_19_1.pdf DOI: 10.26102/2310- 6018/2019.27.4.044

3. Ильичев В.Ю., Юрик Е.А. Анализ массивов данных с использованием библиотеки Pandas для Python // Научное обозрение. Технические науки. – 2020. – № 4. – С. 41-45

***Nizomova Zuhra Komil qizi***
*Tashkent State University of Uzbek Language*
*and Literature named after Alisher Navoi*
*Tashkent, Uzbekistan*
*nizomovazuhra1995@gmail.com*

# THE IMPORTANCE OF THE THESAURUS OF CHEMICAL TERMS

**Annotation**. This article provides information on the importance of the thesaurus of chemical terms and their current needs, the establishment of semantic relationships between terms, what sources can be used for this, and so on. It should be noted that the principles of creating a thesaurus dictionary, which shows the semantic relationship of terms in the field of chemistry for the system of continuing education and researchers, are revealed.

**Keywords**: reaction chemistry explorer, list of chemical entities, chemical entity, atoms, isotopes, molecular substances & discrete molecules, photons, metals, alloys

The basis for the creation of educational and informational materials for the search and processing of large amounts of information is the development and improvement of modern information methods of teaching. An important and necessary condition for the implementation of these methods is the development and creation of innovative projects and educational information retrieval programs. The widespread development of teaching methods in this area has led to the study of innovative processes and the creation of new automated information retrieval systems. Any subject of education and information activity can be described by a hierarchical dictionary of concepts in this area - thesaurus. In addition, such a thesaurus has long been established for many subjects. The thesaurus is a tree of concepts on a particular topic, ending with the most general and inferior, the most specific, the narrowest concepts. Words (terms) in the thesaurus are usually associated with general-specific, whole-part, and so on. In the broadest sense, a thesaurus is interpreted as a description of a system of knowledge about reality that has an individual data carrier or group of carriers. This carrier can act as a receiver of additional information, as a result of which its thesaurus changes, and the original thesaurus determines the capabilities of the receiver when receiving semantic information. Thesaurus is a term widely used in computer science as an integral part of information retrieval systems. Information retrieval thesaurus is a glossary of terms and phrases created according to certain rules for a particular field of science, designed to

improve the quality of information retrieval in that field.

"The Chemical Thesaurus is a reaction chemistry information system that extends traditional references by providing hyper-links between related information. "The program goes a long way toward meeting its ambitious goal of creating a nonlinear reference for reaction information. With its built-in connections, organizing themes, and multiple ways to sort and view data, The Chemical Thesaurus is much greater than the sum of the data in its database. "The program does an excellent job of removing the artificial barriers between different subdisciplinary areas of chemistry by presenting a unified vision of inorganic and organic reaction chemistry." [1:35]

The word thesaurus means storehouse, and The Chemical Thesaurus a storehouse of information about chemical species [entities] and chemical reactions, interactions and processes.

Also, the application behaves rather like the thesaurus built into our word processors that allows us to jump from word-to-word by meaning:

The Chemical Thesaurus allows us to jump from chemical to chemical via the associated interactions, reactions and processes. For example, it is possible to click thru the industrial synthesis of nylon-6,6:

And, with the Chemical Thesaurus it is possible to *click back* to find out how nylon-6,6 is made.

***The Chemical Thesaurus is a reaction chemistry explorer. [1:47]***

Chemistry is often described as the study of matter and its changes. This is crucial because the relational database schema that under lies The Chemical Thesaurus – the very architecture of the application – is explicitly designed in terms of matter and the changes that occur to matter.

Matter is considered in terms of chemical entities.

Changes to matter are considered in terms of the interactions, reactions and/or processes of defined chemical entities.

The term chemical entity is used because it is inclusive and can be used to group together all objects of chemical interest including: atoms, isotopes, molecular substances & discrete molecules, photons, metals, alloys, ionic salts, network materials, electrons, ions, radicals, reactive intermediates, generic species such as nucleophile, and even specialist apparatus like the Dean & Stark trap.

No other term is so general:

The sodium ion, $Na^+$, is a chemical species but not a substance or a material.

Diamond is a material and it is a substance, but not a species. Aldehydes and nucleophiles are hypothetical, generic objects. The Dean & Stark trap is glassware.

A particular chemical entity may have one name or several synonyms. For example the compound CH3I is commonly called both methyl iodide and iodomethane, and both names appear in the synonyms database.

All chemical changes can be described by chemical equations:

- 2 H2 + O2 –> 2 H2O
- crude oil –> methane, propane, butane...
- A + B –> C

The reaction equation a powerful metaphor able to describe processes from elementary particle interactions to biochemistry.

Reaction equations can be balanced in terms of numbers of entities, mass, enthalpy, entropy and Gibbs free energy, or they may be unbalanced.

Hypothetical interactions and processes can be described.

Both physical changes and chemical changes can be modelled by chemical reaction equations. [2:4]

Actually, there is no theoretical or clear-cut separation between "physical" and "chemical" change, although the distinction may sometimes be useful with beginning science students. Technically, all material changes are changes in phase space.

Chemistry is commonly discussed in terms of hypothetical species with ideal behaviour, with real species assigned to these ideal, generic species. Consider the statement:

"Acetaldehyde and propanal are aldehydes."

Acetaldehyde and propanal are real chemical entities, while the hypothetical aldehyde is an idealised generic species.

The term 'Markush structure or group' is sometimes used for generic, particularly in the patent literature.

This logic is formalised and developed in The Chemical Thesaurus. This is possible because the reaction chemistry database can hold information about any type of chemical object:

chemical reagents

molecular ions

reactive intermediates

and generic species such as: aldehyde (generic)[5:5]

Moreover, the software allows the user to jump between real species and their associated generic species.

For example, acetic acid is a carboxylic acid and clicking on the Carboxylic acid (generic**)** link will jump to a page where all of the carboxylic

acids in the database are listed.

Don't worry, it is much easier to do with a click of the mouse than it is to to explain in words! But you may have been wondering what all the references to *"*generic*"* were. Generic species are *always* listed with **(generic)** after the name to avoid confusion.

A great deal of chemical education involves understanding the chemistry of generic species, and learning how to assign real species as generic species with each other. This approach is integral to how The Chemical Thesaurus is organised. Test your knowledge by going the Chemistry Tutorials & Drills web site.

Retro Synthetic Analysis (RSA) is a technique employed in advanced synthetic organic chemistry to help design the sequence of reactions to a large, multifunctional molecule entity, such as a natural product or pharmaceutical agent. The idea is to logically find the synthetic building blocks required for construction by "disconnection".[3:37]

This is achieved by looking for strategic bonds and the potential functional group inter-conversions in a molecule, and then to deducing the synthetic entities, or "synthons", required to construct the desired molecule in the lab.

For example, acetic anhydride can be disconnected onto an "acetyl cation synthon" and an "acetate ion synthon":

There is no actual reaction in which an acetyl cation reacts with an acetate anion, because both ions require counter ions, however, the RSA analysis is conceptually very useful.

RSA deconstruction logic has been extended in The Chemical Thesaurus to main group chemistry. For example, the trivial Na+ plus Cl– reaction to give sodium chloride is shown as a retro synthetic disconnection:

Please note that even simple chemistry can generate naming problems. For example:

The chemistry associated with elemental sulfur is commonly associated with S – as it is here – but the species S does not exist, at least not below 1000°C.[10:124]

Flowers of sulfur, the common yellow soft crystalline form of the element, is S8. If this species were to be used in the reaction chemistry database all stoichiometries would have to be multiplied by 8 and the numbers would become unnecessarily cumbersome.

The species S1 is invented for the sake of simplicity.

Likewise, there are two types of proton, $H^+$, in the database:

The proton of high energy physics: $H^+$(vacuum).

The proton associated with Bronsted acid reaction chemistry: $H^+$(solvated).

A decision has been made to have separate entries for these two types of proton.

A decision has also been made to have separate entries for minerals and reagent chemicals.[3:40]]

The reason is that few minerals are chemically pure and chemists like composition to be defined within 1% or better. [9:182]The decision to separate minerals from chemical reagents leads occasional double entries, such as two entries for gypsum: gypsum the mineral of variable composition and gypsum the pure chemical reagent.[9:184]

Another problem results from the usual conventions of writing chemical equations: reaction products and by-products are expressed as pure materials even though they seldom are.

For example, an aqueous industrial manufacturing process may produce sulfuric acid in water as a by-product. Clicking on the sulfuric acid icon will transport the user to the concentrated sulfuric acid data page, yet it is not possible (with any energetically efficiency, at least) to convert aqueous sulfuric acid into concentrated sulfuric acid.

Thus, some chemical intelligence is required when navigating through the relationally linked database tables.

In short, thesauruses are the most acceptable method of search engine in today’s modern information system. Through thesauruses, you can get more and more information you need. Nowadays, the need for thesauruses for every field of science is growing day by day. Therefore, the thesaurus dictionary of chemical terms, which is also intended to be created for users in the field of chemistry, is a useful resource.

**References**

1. K.R. Cousins, J.Am.Chem.Soc, pp 8645-6 (2001)

2. Aktaev, E. K. Introduction of modern methods of automated thesaurus in the process of learning and information / E. K. Aktaev. - Text: directly // Young scientist. - 2015. - No 6,2 (86,2). - P. 3-6. - URL: https://moluch.ru/archive/86/16526/)

3. Lukashevich N.V. "Tezaurusy v zadachax informatsionnogo poiska". 2011. C. 20-50.

4. L.V. Shcherba "Opyt teorii obshchey lexikografii". 1940. https://www.ruthenia.ru/apr/textes/sherba/sherba9.htm

5. "From redaktsii"; Dobrovolskiy V. Ot sostavitelya; ... S. Baranova. - M., 1965.C. 4-6.

6. Lukashevich N.V., Sali A.D., Demonstration of knowledge in the system of automatic text processing // NTI, 2-ser. 1997. № 3. S. 1-6.

7. Juravlev S.V., Yudina T.N., Information system RUSSIA // NTI, Ser. 1995. № 3. S. 18-20.

8. Winston M., Chaffin R., German D. Taxonomy otnosheniy v tselom // Cognitive science. 1987. Net. 11. S. 417-444.

9. Priss Yu.E., Formulation of WordNet s pomoshchyu metodov analiza rodstvennyx ponyatiy // WordNet. Electronic lexical base dannyx / Pod red. K. Fellbaum. Cambridge, Massachusetts, London, England .: MIT Press, 1998. p. 179–196.

10. Guarino N., Welti K., Официальная онтология признаков // Materials seminar ECAI-00 on the application of ontology and methods of solving tasks. Berlin: 2000. p. 121-128. (http://citeseer.nj.nec.com/guarino00formal.html).

*УДК 004.42*
***[1]Шарипбай А.А., [2]Омарбекова А.С.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]sharalt@mail.ru, [2]omarbekova_as@enu.kz*

# СТРУКТУРА БАЗЫ ДАННЫХ ТЕРМИНОВ ШКОЛЬНЫХ ПРЕДМЕТОВ И ВИДЫ ЗАПРОСОВ К НИМ

**Аннотация:** Статья посвящена разработке электронного словаря терминологии по школьным учебникам на казахском языке, описаны структура базы данных, запросы и пользовательский интерфейс на казахском языке. Терминологический словарь школьных учебников призван улучшить понимание учеников школьных предметов, повысить качество обучения. Данное исследование проводилось в рамках проекта, финансируемого Комитетом науки Министерства образования и науки Республики Казахстан (грант № BR11765535). Результаты проекта будут способствовать развитию казахского языка в области образовании и языкового капитала граждан Казахстана в соответствии с задачей “Расширение функций и повышение культуры использования казахского языка в области образования" Государственной программы по реализации языковой политики в Республике Казахстан на 2020 − 2025 годы.

**Ключевые слова:** Терминологический словарь, термин, школьный предмет.

*ӘОК 004.42*
***[1]Шәріпбай А. Ә., [2]Омарбекова А.С.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]sharalt@mail.ru, [2]omarbekova_as@enu.kz*

# МЕКТЕП ПӘНДЕРІ БОЙЫНША ДЕРЕКТЕР БАЗАСЫНЫҢ ҚҰРЫЛЫМЫ МЕН ОЛАРҒА ҚОЙЫЛАТЫН СҰРАТЫМДАР ТҮРІ

**Аңдатпа:** Мақала қазақ тіліндегі мектеп оқулықтарының электрондық терминологиялық сөздігін жасауға арналған, деректер базасының құрылымы, сұратымдары және қазақ тіліндегі пайдаланушы интерфейсі сипатталған. Мектеп оқулықтарының терминологиялық

сөздігі оқушылардың мектеп пәндерін түсінуін жақсарту мен білім сапасын арттыру мақсатында жасалған. Бұл зерттеу Қазақстан Республикасы Білім және ғылым министрлігі Ғылым комитеті (грант № BR11765535) қаржыландыратын жобаның бір бөлігі ретінде жүзеге асырылды. Жобаның нәтижелері Қазақстан Республикасындағы тіл саясатын іске асырудың 2020-2025 жылдарға арналған мемлекеттік бағдарламасының «Қазақ тілінің білім беру саласындағы функцияларын кеңейту және қолдану мәдениетін арттыру» міндетіне сәйкес Қазақстан азаматтарының білім беру саласындағы қазақ тілін дамытуға және тілдік байлығын арттыруға ықпал етеді.

**Түйін сөздер:** Терминологиялық сөздік, термин, мектеп пәні.

*UDC 004.42*

***[1]Sharipbay A., [2]Omarbekova A.***
*L. N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]sharalt@mail.ru, [2]omarbekova_as@enu.kz*

## THE STRUCTURE OF THE DATABASE OF TERMS OF SCHOOL SUBJECTS AND THE TYPES OF QUERIES TO THEM

**Abstract:** The article is designed to create an electronic terminological dictionary of school textbooks in Kazakh language, describes the structure of the database, queries and user interface on Kazakh language. The terminology dictionary of school textbooks is designed to improve students' understanding of school subjects and improve the quality of education. This study was implemented as part of a project funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (grant № BR11765535). The results of the project are in line with the state program for the implementation of language policy in the Republic of Kazakhstan for 2020-2025. In accordance with the task "Expanding the functions of the Kazakh language in the field of education and improving the culture of its use" promotes the development of the Kazakh language and increase the language wealth of the citizens of Kazakhstan in the field of education.

**Keywords:** Terminological dictionary, term, school subject.

В век цифровых технологий и прогресса применяется много различных новых терминов. Для облегчения изучения школьных предметов требуется создать электронный словарь терминологии школьных учебников, состоящий из исходного кода, базы данных

терминов школьных учебников, пользовательского интерфейса на казахском языке с функцией корректировки имеющихся и добавления новых терминов.

Сформирован список школьных предметов и определены классы:

| № | Пән ( Предметы) | Класс |
|---|---|---|
| 1 | Әліппе | 1 |
| 2 | Әдебиеттік оқу | 1,2,3,4 |
| 3 | Қазақ тілі | 2,3,4,5,6,7,8,9,10,11 |
| 4 | Қазақ әдебиеті | 5,6,7,8,9,10,11 |
| 5 | Орыс тілі | 1,2,3,4,5 |
| 6 | Орыс әдебиеті | 1,2,3,4 |
| 7 | Орыс тілі мен әдебиеті | 5,6,7,8,9,10,11 |
| 8 | Шетел тілі | 1,2,3,4,5,6,7,8,9,10,11 |
| 9 | Ақпараттық-коммуникациялық технологиялар | 3,4 |
| 10 | Математика | 1,2,3,4,5 |
| 11 | Жаратылыстану | 1,2,3,4,5,6 |
| 12 | Дене шынықтыру | 1,2,3,4,5,6,7,8,9,10,11 |
| 13 | Дүниетану | 1,2,3,4 |
| 14 | Көркем еңбек | 1,2,3,4,5,6,7,8,9 |
| 15 | Музыка | 1,2,3,4,5,6 |
| 16 | Өзін-өзі тану | 1,2,3,4,5,6,7,9,10.11 |
| 17 | Информатика | 5,6,7,8,9,10,11 |
| 18 | Қазақстан тарихы | 5,6,7,8,9,10,11 |
| 19 | Дүниежүзі тарихы | 5,6,7,8,9,10,11 |
| 20 | Алгебра | 7,8,9 |
| 21 | Геометрия | 7,8,9,10,11 |
| 22 | География | 7,8,9,10,11 |
| 23 | Биология | 7,8,9,10,11 |
| 24 | Физика | 7,8,9,10,11 |
| 25 | Химия | 7,8,9,10,11 |
| 26 | Алгебра және анализ бастамалары | 10,11 |
| 27 | Алғашқы әскери және технологиялық дайындық | 10,11 |
| 28 | Құқықтар негіздері | 10,11 |
| 29 | Кәсіпкерлік және бизнес негіздері | 10,11 |
| 30 | Инвариантық оқу жүктемесі | 10,11 |
| 31 | Графика және жобалау | 11 |
| 32 | Инвариантық оқу жүктемесі | 10,11 |
| 33 | Алғашқы әскери және технологиялық дайындық | 10,11 |

В настоящее время участниками проекта ведутся работы по формированию полной базы школьных терминов.


Рисунок 1. Структура базы данных «Terms»

Поле «Class» хранит информацию о номере класса, «kaz» - значение термина, «znach» - определение термина.

Для того, чтобы приложение находило не только конкретные слова, но и позволяло находить термины по первым буквам, в sql-запросе использован предикат LIKE.

Оператор SQL LIKE устанавливает соответствие символьной строки с шаблоном. В шаблоне разрешается использовать два трафаретных символа:

- символ подчеркивания (_), который можно применять вместо любого единичного символа в проверяемом значении;
- символ процента (%) заменяет последовательность любых символов (число символов в последовательности может быть от 0 и более) в проверяемом значении.

Используемый запрос имеет следующий вид:

$query = "select * from terms where kaz like '".$poiskk."%' order by kaz";

$result = $mysqli->query($query);

Здесь в переменной $poiskk хранится значение вводимых с клавиатуры первых букв термина.

Проект терминологического словаря размещен в интернете по ссылке https://e-zerde.kz/terms_class

В пустом поле необходимо ввести первые буквы термина и нажать кнопку «ОК», далее в таблице отобразятся класс и термины с определениями.

Терминологический словарь школьных учебников призван улучшить понимание учеников школьных предметов, повысить качество обучения.

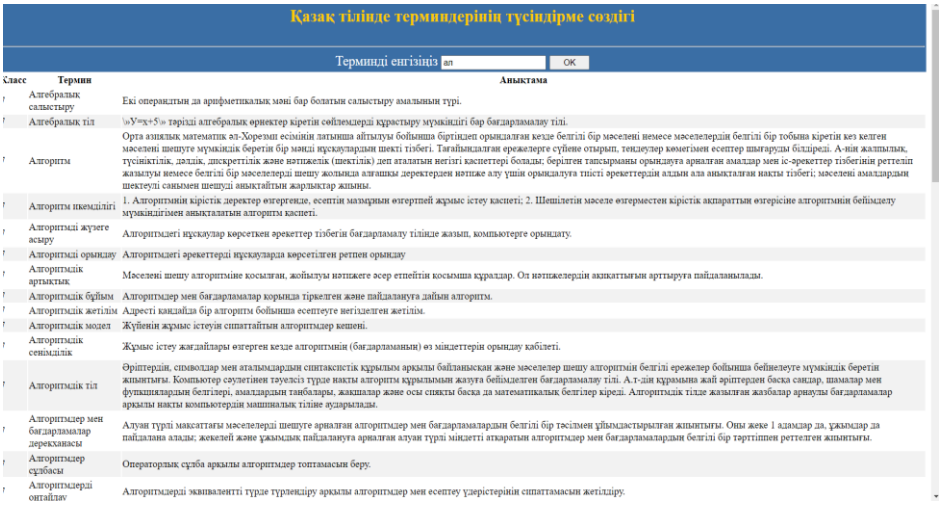



Рисунок 2. Приложение терминологического словаря

Проект соответствует ряду стратегически важных государственных задач, так он отвечает Седьмому вызову «Третья индустриальная революция» Стратегии «Казахстан-2050» от 14 декабря 2012 года [1] в части становления активными участниками процесса развития цифровых технологий. Согласно Национального Плана развития РК до 2025 года [2] соответствует общенациональному приоритету “Качественное образование”. Согласно задаче “Расширение функций и повышение культуры использования казахского языка в области образования” Государственной программы по реализации языковой политики в Республике Казахстан на 2020 − 2025 годы [3] проект способствует развитию казахского языка в области образовании и языкового капитала граждан Казахстана. Проект имеет очень высокую значимость в национальном масштабе, а эффект, вызванный повышением культуры казахского языка в образовании, качества обучения, развитие цифровых сервисов положительно скажется на международном имидже Казахстана.

Исследование, описанное в данной работе выполнена по программно-проекту «Разработка научно-лингвистических основ и IT-ресурсов по расширению функций и повышению культуры казахского языка» (грант № BR11765535), который финансируется Комитетом науки Министерства образования и науки Республики Казахстан в рамках программно-целевого финансирования на 2022–2023 годы.

**Список литературы**

1. Послание Президента Республики Казахстан - Лидера Нации Н.А.Назарбаева Народу Казахстана «Стратегия «Казахстан-2050»: Новый политический курс состоявшегося государства» (Астана, 14 декабря 2012 года).

2. Национальный план развития Республики Казахстан до 2025 года, утвержден Указом Президента Республики Казахстан от 15 февраля 2018 года № 636// https://adilet.zan.kz/rus/docs/U2100000521#z11

3. Государственной программе по реализации языковой политики в Республике Казахстан на 2020 − 2025 годы, https://adilet.zan.kz/rus/docs/P1900001045#z11

*ӘОК 81.35*
***Кадирхан А.К.***
*Ш. Шаяхметов атындағы «Тіл-Қазына»*
*ұлттық ғылыми-практикалық орталығы*
*Нұр-Сұлтан, Қазақстан*
*tilortalyk@mail.ru*

## «ТӘУЕЛСІЗ» СӨЗІНІҢ ОРФОГРАММАСЫ ЖӘНЕ ҚОЛДАНЫС ЕРЕКШЕЛІГІ

*(Қазақ тілінің публицистикалық кіші корпусы материалдарында)*

**Аңдатпа**. Дамыған елдерде барлық нәрсе тұрақтылыққа бейімделеді. Сол секілді орфографиялық сөздіктерде жиі жаңалана бермейді. Тіпті ғасырлар бойы сөздіктің өзгермеуі олар үшін қалыпты нәрсе. Өйткені дамушы елдерде ғалымдардан сөздіктегі нұсқаның тұрақтануын және нормаға айналуын талап етеді. Қазақстандағы ахуалға келер болсақ, бүгінгі жазарман кейбір сөздердің жазылуын орфографиялық сөздікке сүйенбей, өз бетінше жазуға дағдыланған. Өйткені сөздіктің өзінде бір модельді сөздер бірде бөлек, енді бірде бірге жазылған. Осы тұста тіліміздегі сөздердің әртүрлі вариацияға ие болуына жазарман кінәлі ме, әлде сөздік пе деген сұрақты қоя отырып, оның жауабын жүргізілген сауалнаманың қорытындысы арқылы анықтап, қазіргі қолданыстағы кейбір ала-құла жазылып жүрген лексикалық бірліктерді айқындадық.

**Түйін сөздер:** орфография, лексикография, корпус, модель, статистика.

*УДК 81.35*
***Кадирхан А. К.***
*Национальный научно-практический центр*
*«Тіл-Қазына» имени Ш. Шаяхметова*
*Нур-Султан, Казахстан*
*tilortalyk@mail.ru*

## ОРФОГРАММА И ОСОБЕННОСТИ УПОТРЕБЛЕНИЯ СЛОВА «НЕЗАВИСИМЫЙ»

*(На материалах публицистического подкорпуса казахского языка)*

**Аннотация**. В развитых странах все приспособлено к стабильности. Точно так же редко обновляются и орфографические словари. Более того, словари могут не меняться на протяжении веков,

так как в развивающихся странах от ученых требуется стабилизировать версию словаря и сделать ее нормой. Что касается ситуации в Казахстане, то современный писатель привык писать некоторые слова самостоятельно, не опираясь на орфографический словарь, поскольку даже в словаре слова одной модели пишутся как по отдельности, так и вместе. В связи с этим мы задались вопросом, в существовании различных вариантах слов в нашем языке виноват автор или это это происходит из-за словарей, и по результатам опроса выявили некоторые из существующих лексических единиц.

**Ключевые слова:** орфография, лексикография, корпус, модель, статистика.

*UDC 81.35*

***Kadirkhan A.***
*National Scientific and Practical Center*
*«Til-Kazyna» named after Sh. Shayakhmetova*
*Nursultan, Kazakhstan*
*tilortalyk@mail.ru*

## ORTHOGRAM AND FEATURES OF THE USE OF THE WORD «INDEPENDENT»

(Based on the materials of the journalistic sub-corpus of the Kazakh language)

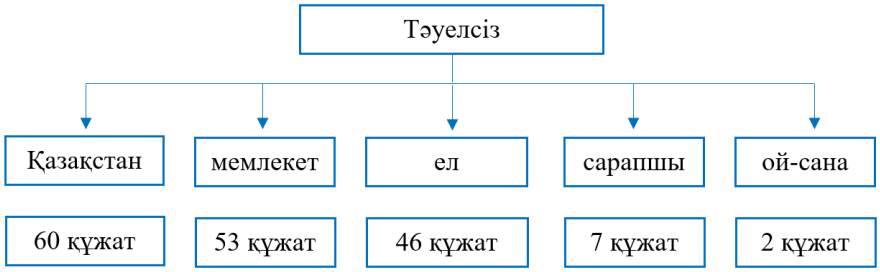
**Abstract**. In developed countries, everything is adapted to stability. Spelling dictionaries are also rarely updated. Moreover, dictionaries may not change for centuries, since in developing countries scientists are required to stabilize and normalize one version of the dictionary. With regard to the situation in Kazakhstan, modern writer is used to writing some words on his own, without relying on a spelling dictionary, since even in the dictionary the words of the same model are written both together and separately. In this regard, we wondered if the author is to blame for the existence of different variants of words in our language or if it is because of the dictionaries. According to the survey results, we identified some of the existing lexical units.

**Keywords**: spelling, lexicography, corpus, model, statistics.

Бүгінде мерзімді басылымдардан, БАҚ-тан, әлеуметтік желілерден бір түбірден өрбіген, морфологиялық құрамы ұқсас, жасалу жолдары бірдей сөздердің контекске келгенде әртүрлі жазылып жүргенін

баршамыз білеміз. Яғни қатенің көпшілігі мәтін ішінде бір модельді сөздердің бірде бөлек немесе бірге, енді бірде бас әріппен немесе кіші әріппен жазылуынан болып жатады. Осының салдарынан қазақ тілінде кейбір сөздердің жарыспа тұлғалары пайда болып, олар нормадан ауытқып, әртүрлі вариацияға ие болуда. Мәселен, *әртүрлі – әр түрлі, қонақ үй – қонақүй, байқұс – байғұс* т.т. Осы тектес сөздердің жазылуда дұрыс/ бұрыстығын анықтау үшін Орфографиялық сөздікке сүйену қажет екенін баршамыз білеміз. Алайда аталған сөздіктің құрылымына, оған енген сөздердің тізбегіне қарап, оның тек сөздерді бірге немесе бөлек жазылатынын анықтауға арналған секілді. Ал жазу практикасында әртүрлі себептерге байланысты жазылым нормасынан ауытқуға ұшыраған бірде бас әріппен немесе үнемі кіші әріппен жазылатын сөздердің тізбегі көрсетілмеген. Қазақ мәтіндерінде қатенің көпшілігі бір модельді сөздердің контекске келгенде, орын талғамай бірде бас әріппен, бірде кіші әріппен жазу салдарынан болып жүр. Оның қандай жағдайда бас әріппен, қандай жағдайда кіші әріппен жазылатынына қатысты ғылыми түсініктеме жоқтың қасы. Неге десеңіз, өйткені қазіргі қолданыстағы қазақ тілінің емле ережелеріндегі мысалдардың басым көпшілігі дау туғызбайтын, шатасуға алып келмейтін сөздермен берілген. Емле ережеде арнайы «Бас әріп» деген тарау бар, алайда бұл тарауда тек белгілі бір топтардың ғана бас әріппен жазылатындығы және соған сәйкес мысалдары дәйектеліп берілген. Ереже бойынша бас әріппен жазылатын сөздердің тобы мыналар – адамның аты-жөні, ғимарат, мекемелердің, құрылымдық бөлімшелердің атауы, жер-су атаулары, тарихи тұлғалар мен маңызды оқиғалардың атаулары және т.б. Бірақ бұл аталған топтарға жатпайтын, ережеленіп, нормаланбаған кейбір лексикалық бірліктер мәтін ішінде әртүрлі жазылып жүр. Мәселен, «тәуелсіз» деген сөзді алайық. Қазақ мәтіндерінен аңғарсаңыздар, бұл сөз бірде бас әріппен, енді бірде кіші әріппен жазылады. Неге бірізділік жоқ деген сұраққа жауапты бірге іздеп көрелік. Ең алдымен «тәуелсіз» сөзінің бас әріппен болмаса кіші әріппен жазылатын, жазылмайтындығын анықтау үшін аталған сөздің ұғымына, түсіндірмесіне мән берейік. 2015 жылы шыққан «Қазақ әдеби тілінің 15 томдық сөздігінде» тәуелсіз сөзіне екі түрлі анықтама келтірілген: с ы н. 1. Ешкімге, еш нәрсеге бағынышты емес; байланыссыз. 2. с а я с и-қ ұ қ ы қ. Өз саясатын, билігін өз еркімен жүргізуге қабілетті [1] Берілген анықтамалардан «тәуелсіз» сөзінің жалпы мәні түсінікті, тіпті оның қай сөз табынан екендігі де көрсетілген. Жалпы осы «тәуелсіз» сөзі мәтінде жеке дара қолданысынан гөрі, оның өзіне тән тіркестік қатар құрайтын негізгі бірліктерімен жиі қолданылады. Олар мыналар: ел, мемлекет, сарапшы,

ой-сана, Қазақстан т.т. Сонымен қатар бұл сөздерден бөлек «тәуелсіз» сөзі республика, саясат, әрекет, журналист, комиссия, кеңесші деген секілді сөздермен де тіркеседі. Енді осы келтірілген мысалдардың ішінде «тәуелсіз» сөзі қандай сөзбен жиі тіркеседі және оның қалай жазылатынына тоқталайық. Ол үшін «Қазақ тілінің ұлттық корпусының кіші корпустар» базасына енгізілген публицистикалық мәтіндердегі қолданысын негізге алдық. Аталған корпуста барлығы 5304 құжат, 5 141 233 сөз бар. Осы базаны негізге ала отырып, біз «тәуелсіз» сөзінің жалпы қолданысын және тіркесін анықтадық. Яғни «тәуелсіз» сөзі аталған базада жалпы 531 құжаттың құрамында кездеседі және ол қолданылу жиілігі жағынан 1000-сөзді құрайды. Төмендегі кестеде «тәуелсіз» сөзімен ең жиі тіркесетін сөздердің қатары және мәтіндердегі статистикалық қолданысы көрсетілген:




Берілген мысалдардан публицистикалық мәтіндер бойынша «тәуелсіз» сөзі «Қазақстан» сөзімен қолданысы жағынан жиілігі жоғары күрделі сөз тіркесі екенін аңғаруға болады. Қазақ мәтіндерінде «тәуелсіз» сөзі әсіресе осы «Қазақстан» сөзімен тіркескенде позиция талғамай үнемі бас әріппен жазылады. Мейлі ол тіркес сөз ортасында болсын, сөз соңында болсын үнемі бас әріп сақталады. Сөзімізге дәлел ретінде, «Қазақ тілінің ұлттық корпусының кіші корпустар» базасына енгізілген публицистикалық мәтіндер бойынша мысал келтірейік: *1)* ***Тәуелсіз Қазақстан*** *тарихында тұңғыш рет басшылардың қоластындағы қызметкерлері үшін жауапкершілігі заңнамалық деңгейде бекітілді. 2) Бұл батыр апаларымыздың есімін* ***Тәуелсіз Қазақстан*** *жастары ешқашан ұмытпауы керек. 3) Қазақ хандығы дәуірінде зор бедел, абыройға ие болған кемел ойлы, текті тұлға Нияз батырдың бүгінгі* ***Тәуелсіз Қазақстан*** *тарихынан лайықты орнын, әділ бағасын алуы заңдылық деп білеміз* [2] Қарап отырсаңыздар, бірінші сөйлемде «Тәуелсіз Қазақстан» сөз тіркесі сөйлемнің басында, екінші және үшінші сөйлемде сөз ортасында кездесіп тұр. Егер алғашқы сөйлемді «Әрбір сөйлем бас әріппен жазылады» деген ережеге сәйкестендіретін болсақ, логикалық құрылымы дұрыс. Сөз басында тұр, демек бас әріппен жазылады. Ал енді сөз ортасында болмаса сөз

соңында жазылған жағдайды ережемен де, сөздікпен де дәлелдеу мүмкін емес. Өйткені бұл тіркестің жазылымына байланысты арнайы емле соның негізінде оның тұрақты орфографиясы қалыптаспаған. 2013 жылы шыққан Орфографиялық сөздікте «тәуелсіз» сөзі кіші әріппен жазылғанымен, оның дұрыс-бұрыстығын да негіздей алмаймыз. Өйткені сөздіктегі барлық сөздер кіші әріппен таңбаланған, ал осы тектес бас әріппен жазылатын сөздердің тізбегі мүлдем көрсетілмеген. Ал орыс тілінің орфографиясында «тәуелсіз» сөзі қандай сөзбен тіркессе де, ол позиция талғамай бірыңғай кіші әріппен жазылады. Орыс ғалымы С.И. Ожегованың «Түсіндірме сөздігінде» тәуелсіз сөзінің екі лексикалық мағынасы көрсетілген: «1. см. независимый.2. Политическая самостоятельность, отсутствие подчинённости, суверенитет» [3]. Қазақ және орыс тілдерінде «тәуелсіз» сөзінің беретін түпкі мағынасы бір болғанымен, олардың басты айырмашылығы жұмсалым орнында және жазылымында. Мәселен, 1) *За период национальной* ***независимости Казахстан*** *не раз демонстрировал особое понимание кавказских проблем, не совпадающее как с официальной позицией Российской Федерации, так и со взглядами геополитических конкурентов России. 2) Развивается недавно введенная в эксплуатацию новая подземная шахта «10 лет* ***независимости Казахстана****», общая мощность которой рассчитана на 4 млн труды* [4]. Бұл келтірілген мысалдар «Национальной корпус русского языка» (Орыс тілінің ұлттық корпусы) базасынан алынды. Яғни бұдан байқағанымыздай кез келген мәтінде орыс орфографиясы бойынша анықтауыш сөз (*независимость*) кіші әріппен, ал анықталушы сөз (*Казахстан*) бас әріппен жазылады. Мұндай сөздердің жазылуында бірізділіктің сақталуы сөздердің жарыспалы нұсқасын тудырмауға жол береді. Ал қазақ орфографиясы бойынша кейбір сөздер ережеге бағынбай, қоғамның қабылдауы бойынша әртүрлі жазылып, қалыптасады. Сонымен «Тәуелсіз Қазақстан» тіркесіндегі анықтауыш сөз неге кіші әріппен жазылмайды дегенде оны орфография жағынан емес, лингвомәдениеттану тұрғысынан түсіндіруге болатын секілді.

«Тәуелсіз» сөзі қазақ халқының танымында сакральдық мәнге ие болған сөздердің бірі. Олай дейтін себебіміз қазақ халқы «тәуелсіз» деген сөздің мән-мағынасын жіті түсінген және оны жоғары бағалаған. «Тәуелсіз», «тәуелсіздік» сөздері бұл – бүкіл әлемге қазақ деген ұлттың, халықтың өмір сүретінін, оның өзге ұлттардан ерекшелігін танып-білуге мүмкіндік беретін, ұлтымызды рухтандыратын тілдік бірлік. Тәуелсіздікке қол жеткізу жолында жанын берген, терін төккен ата-баба аманатын баршаға жеткізу, әрі еліміздің дербес мемлекет екендігін көрсету мақсатында қоғамның санасында «Тәуелсіз» деген сөз әсіресе

«Қазақстан» сөзімен тіркескенде үнемі бас әріппен жазылуы тиіс деген дағды қалыптасқан. Яғни «тәуелсіз» дегенде оның астарында «Қазақстан» деген мемлекетіміздің атауы тұрғанын, сол арқылы елге, жерімізге деген құрметтің белгісі ретінде оны бас әріппен таңбалау дұрыс деген графикалық сауаттылық қалыптасқан.

Біз қоғам арасында жалпы «тәуелсіз» сөзінің өзіне, сондай-ақ оның «Қазақстан» сөзімен тіркескендегі қолданысына байланысты кішігірім электронды нұсқада сауалнама жүргізген болатынбыз. Сауалнаманы жүргізудегі негізгі мақсат орфографиялық олқылықтарды анықтау, ала-құлалықты біріздендіру және ұсыныс-пікірлерді саралау. Жалпы сауалнамаға 18 жастан 50 жасқа дейінгі әртүрлі әлеуметтік топ өкілдері қатысты. Сауалнама 5 сұрақтан және әр сұрақта 2 немесе 3 жауап көрсетілді. Қатысушылардың көзқарасын білу мақсатында кейбір сұрақтардың жауабын арнайы көрсетпей, өздеріне қалдыруды жөн санадық. Соның негізінде ала-құла жазылып жүрген қазіргі қолданыстағы сөздердің қатарын анықтадық. Сонымен сауалнамада берілген сұрақтардың мазмұны төмендегідей:

1. «Тәуелсіз» сөзі сізге қандай ассоциациялық ұғым қалыптастырады?
   - Қазақстан
   - Сарапшы
   - Адам
2. «Тәуелсіз Қазақстан» сөз тіркесі қандай әріппен жазылғаны дұрыс деп ойлайсыз?
   - бас әріп;
   - кіші әріп
   - бір сыңары бас әріп, бір сыңары кіші әріп
3. «Тәуелсіз» сөзімен тіркесетін сөздердің тізбегін сөздікке енгізу қажет пе?
   - иә;
   - жоқ;
   - тек қиындық тудыратын тіркестерін қана
4. «Тәуелсіз Қазақстан» сөзі мәтін ішінде бірде бас әріп, енді бірде кіші әріппен жазылады. Бұдан бөлек тағы қандай әртүрлі жазылатын сөздерді немесе сөз тіркестерін білесіз?
5. Жазылуда қиындық келтіретін сөздердің дұрыс/бұрыстығын тексеру үшін неге сүйенесіз?
   - сөздікке;
   - ережеге;
   - өз бетімше жазамын.

Енді әр сұрақтың қорытындысын берер болсақ, бірінші сұрақ бойынша қатысушылардың басым көпшілігі «тәуелсіз» сөзінің негізгі ассоциациялық мән тудыратын тіркесі деп 71,4% – Қазақстан деген сөзді көрсетсе, ал қатысушылардың қалғаны, яғни 14,3 % – сарапшы, 14,3% – адам дегенді нұсқаны таңдаған. Екінші сұрақтың қорытындысы бойынша қатысушылардың 83,2% – «Тәуелсіз Қазақстан» сөз тіркесін бас әріппен жазылуы тиіс деген нұсқаны қолдаса, қалған 9% – кіші әріппен жазылуы тиіс деген нұсқаны, қалған 7,8 % – бір сыңарын бас әріп, бір сыңарын кіші әріппен жазылғаны дұрыс деген нұсқаны белгілеген. Қатысушылардың 100% – Орфографиялық сөздікке «тәуелсіз» сөзімен тіркесетін сөздердің тізбегі көрсетілуі қажет деген нұсқаны таңдаған. Демек қатысушылардың бұл сұраққа берген жауаптарына қарап, «тәуелсіз» сөзінің тіркесімділігін жазуда қиналатындарын аңғаруға болады. Төртінші сұрақ бойынша қатысушылар қазіргі қолданыста қиындық тудырып жүрген, ала-құла жазылып жүрген сөздердің қатарын көрсетті. Олар мыналар – *Атамекен, Отан, Заң, Ұлы, Егемен, Президент, Туған жер, әр түрлі, Әл-Фараби* және т.б. Ең соңғы сұраққа қатысушылардың 51,7% сөздердің жазуда дұрыс/бұрыстығын анықтау үшін сөздікке сүйенетін көрсетсе, қалған 35,8% – ережеге, 12,5% – өз бетінше жазатындықтарын көрсеткен. Қарап отырсақ, қатысушылардың басым көпшілігі «тәуелсіз» сөзінің әсіресе «Қазақстан» сөзімен тіркескенде екі сыңары да бас әріппен жазылуын және сөздікке сол күйінде енгенін қалайды. Сонымен бұл жүргізілген сауалнама бойынша біз нені анықтадық десек, ең алдымен бүгінгі жазарманның берген жауаптары арқылы Орфографиялық сөздікті әлі де жетілдіру қажеттігін, сөздердің ала-құлалығын біріздендіру керектігін, әртүрлі әлеуметтік топпен бірге жазу мәдениетін, сауаттылықты қалыптастыру үшін түрлі тест, сауалнама жүргізу қажеттігін түсіндік. Қазақ мәтіндерінде «Тәуелсіз Қазақстан» деген тіркестің бас әріппен жазылатынын болмаса анықтаушы сыңары кіші әріп, анықталушы сыңары бас әріп екенін нақты көрсету үшін оларды ережелеп, сөздікке енгізу қажет деп ойлаймыз. Бұл тектес тіркестер, сөздер қаншама *(атамекен, Отан)*. Солардың барлығын жинақталып, жүйеленіп бір негізге түссе нұр үстіне нұр болар еді. Орфографиялық сөздік – тіліміздегі сөздердің дұрыс жазылуын реттейтін құрал десек, онда әрбір сөздің жазылуы ғылыми дәйектеліп, қазықтай қағылуы тиіс деп ойлаймыз. Сондықтан алдағы шығатын Орфографиялық сөздікте «Тәуелсіз Қазақстан» сынды осыған ұқсас тілдік бірліктердің жазылуына қатысты мәселелер қайта қаралып, бірізденсе деген тілегіміз бар.

**Әдебиеттер тізімі**

1. Қазақ әдеби тілінің сөздігі. Он бес томдық. 14-том. / Құраст.: М.Малмақов, Қ.Есенова, Б.Хинаят және т.б. – Алматы, 2011. 800 б.
2. https://test.qazcorpora.kz/result?q=0&dt[]=тәуелсіз&dt[]=қазақстан&dv[]=2
3. https://dic.academic.ru/dic.nsf/ogegova/123498
4. https://ruscorpora.ru/new/search-main.html

*ӘОК 004.89*
***[1]Жұмаш Б., [2]Муканова А. С.***
*[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*[2]Астана Халықаралық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]zh.batyrzhan@bk.ru, [2]asiserikovna@gmail.com*

# ОНТОЛОГИЯЛЫҚ МОДЕЛЬ НЕГІЗІНДЕ КӨЛІК САЛЫҒЫ БОЙЫНША АҚПАРАТТЫҚ ЖҮЙЕНІ ӘЗІРЛЕУ

**Аңдатпа:** Мақалада «Көлік салығы» пәндік облысы бойынша білім базасы негізінде ақпараттық жүйені жобалауға байланысты мәселелер көрсетілген. Жүйені әзірлеудің әдіснамалық тәсілін қарастыра отырып, зерттеу саласы анықталып, пәндік облыстың онтологиясын құрудың теориялық аспектілері көрсетілген. «Көлік салығы» пәндік облысы бойынша білімдерді құрылымдау және формализациялау жүргізілген. Онтологиялық моделдің негізгі класстар мен ішкі класстар жіктеліп, оларды бейнелейтін сипаттамалар анықталған. «Көлік салығы» пәндік облысының онтологиялық моделі құрылып, білім базасы негізінде ақпараттық жүйе әзірленген.

**Түйін сөздер:** онтологиялық модел, білімдер базасы, пәндік облыс, SPARQL.

*УДК 004.89*
***[1]Жұмаш Б., [2]Муканова А. С.***
*[1]Евразийский национальный университет имени Л.Н. Гумилева,*
*[2]Международный университет Астана*
*Нур-Султан, Казахстан*
*[1]zh.batyrzhan@bk.ru, [2]asiserikovna@gmail.com*

# РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ПО ТРАНСПОРТНОМУ НАЛОГУ НА ОСНОВЕ ОНТОЛОГИЧЕСКОЙ МОДЕЛИ

**Аннотация:** В статье отражены вопросы, связанные с проектированием информационной системы на основе базы знаний по предметной области «Транспортный налог». рассмотрен методологический подход к разработке системы, определена область исследования, выявлены теоретические аспекты построения онтологии предметной области. Произведена структуризация и формализация

знаний по предметной области «Транспортный налог». Определены характеристики элементов онтологической модели и описаны их значения. Были классифицированы основные классы, подклассы и выявлены характеристики, описывающие данные понятия. Создана онтологическая модель предметной области «Транспортный налог» и разработана информационная система на основе базы знаний.

**Ключевые слова:** онтологическая модель, база знаний, предметная область, SPARQL

*UDC 004.89*

***[1]Zhumash B., [2]Mukanova A.***
*[1]L. N. Gumilyov Eurasian National University*
*[2]Astana International University*
*Nur-Sultan, Kazakhstan*
*[1]zh.batyrzhan@bk.ru, [2]asiserikovna@gmail.com*

## DEVELOPMENT OF A TRANSPORT TAX INFORMATION SYSTEM BASED ON AN ONTOLOGICAL MODEL

**Abstract:** The article presents issues related to the design of an information system based on the knowledge base in the subject area "Transport tax". Considering the methodological approach to the development of the system, the field of research is determined and the theoretical aspects of building the ontology of the subject area are indicated. Structuring and formalization of knowledge in the subject area" transport tax " was carried out. In the ontological model, the main classes and subclasses are classified and the characteristics that represent them are determined. An ontological model of the subject area" transport tax " has been created, and an information system based on the knowledge base has been developed.

**Keywords:** ontological model, knowledge base, subject area, SPARQL

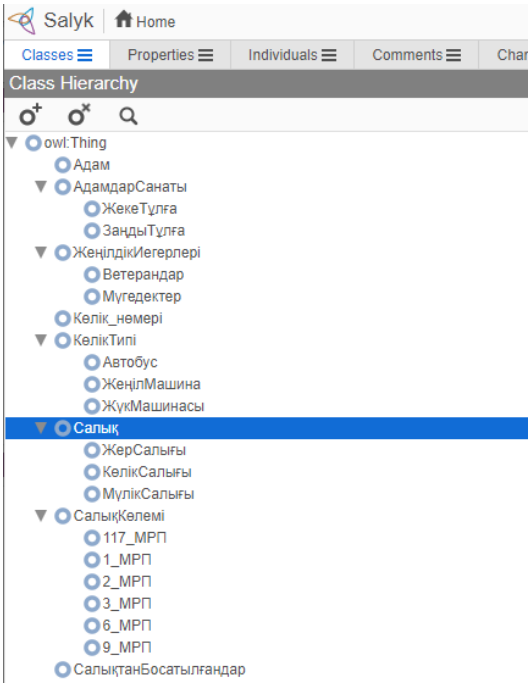
Онтология кез келген пәндік облыс бойынша білімдерді сипаттау үшін қолданылатын ұғымдарды жинақтайды. Пәндік облыс туралы ақпаратты бірлесіп пайдалануы қажет қосымшалар, мәліметтер базасы және адамдар онтологияны пайдалана алады. Онтологияға пәндік аймақтың негізгі ұғымдарының машинамен оқылатын, яғни формальды түрде бейнеленген анықтамалары және олардың арасындағы қатынастар кіреді. Олар пәндік облыстан және онымен байланысты басқа салалардан берілген білімдерді қайта пайдалана отырып жаңа білімдер алуға мүмкіндік береді [1].

Әдетте онтология егжей-тегжейлі, дәл, дәйекті, негізделген және класстарды, қасиеттер мен қатынастарды нақты бөле алатын логикаға негізделген тілде жасалады. Онтологиямен жұмыс істеудің кейбір құралдары онтологияны қолдана отырып, автоматты ойлауды жүзеге асыра алады, осылайша тұжырымдамалық, семантикалық іздеу, программалық қамтама агенттері, шешімдерді қолдау, сөйлеу және табиғи тілді тану, білімді басқару, смарт мәліметтер базасы және электрондық коммерция сияқты смарт қосымшалар үшін жетілдірілген қызметтерді ұсынады [2].

Онтология семантикалық желінің пайда болуында маңызды рөл атқарады, өйткені құжаттар семантикасының көрінісі және осы семантиканы желілік қосымшалар мен интеллектуалды агенттер қолдана алады. Онтология қазіргі уақытта жинақталып, стандартталатын метадеректер терминдерінің мағынасын құрылымдау және сипаттау тәсілі ретінде қоғам үшін пайдалы болуы мүмкін. Онтологияны қолдана отырып, болашақтың қосымшалары "интеллектуалды" бола алады, яғни олар адамның ойлау деңгейінде және тілінде дәлірек жұмыс істей алады. Сондықтан да онтология табиғи тілде деректерді өңдеумен айналысатын барлық салаларда кеңінен қолданылады. Онтологияны әртүрлі қосымшаларда қолдануға байланысты оларды ұсынудың стандартталған тәсілдерін құру қажеттілігі туындады. Барлық жүйелерде қолдануға болатын әртүрлі тілдердің дамуы басталды, олардың ішіндегі ең танымалдары -RDF және OWL.

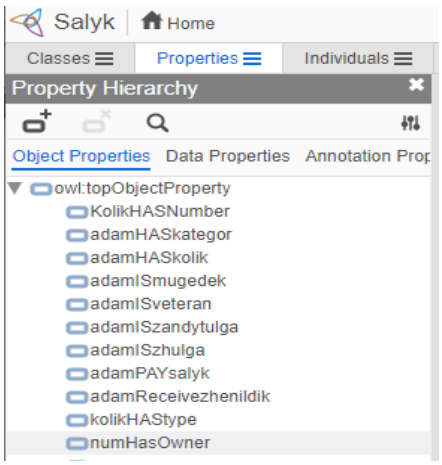
Веб-онтология тілі (OWL) – веб-онтологияларды анықтауға және жасауға арналған тіл. OWL онтологиясы класстардың, қасиеттердің және олардың даналарының сипаттамасын қамтиды. OWL сөздіктердегі терминдердің мағынасын және сол терминдер арасындағы қатынастарды айқын көрсету үшін қолданылады [3].

Бұл жұмыста «Көлік салығы» саласы бойынша онтологиялық модель құрылып, білімдер базасы жасалды, SparQL сұратымдары арқылы табиғи тілде берілген сұрауларға жауап беретін жүйе құрылды. 1 суретте «Көлік салығы» онтологиясындағы негізгі класстар көрсетілген.

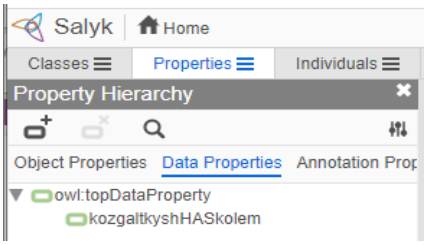



Сурет 1. Класстар иерархиясы

Суретте көрсетілгендей, көлік салығын төлеуші адамдарды «Жеке тұлға» және «Заңды тұлға» деп қарастырылған. Сонымен қатар онтологиялық модельде салық түрлері, транспорт түрлері мен салық төлеуден босатылған немесе жеңілдік түрлері көрсетілген. Пәндік саланың негізгі түсініктерінен басқа, олардың қасиеттері мен қатынастары анықталған. Сурет 2 және Сурет 3 те объектілер қасиеттері мен деректер қасиеттері көрсетілген.

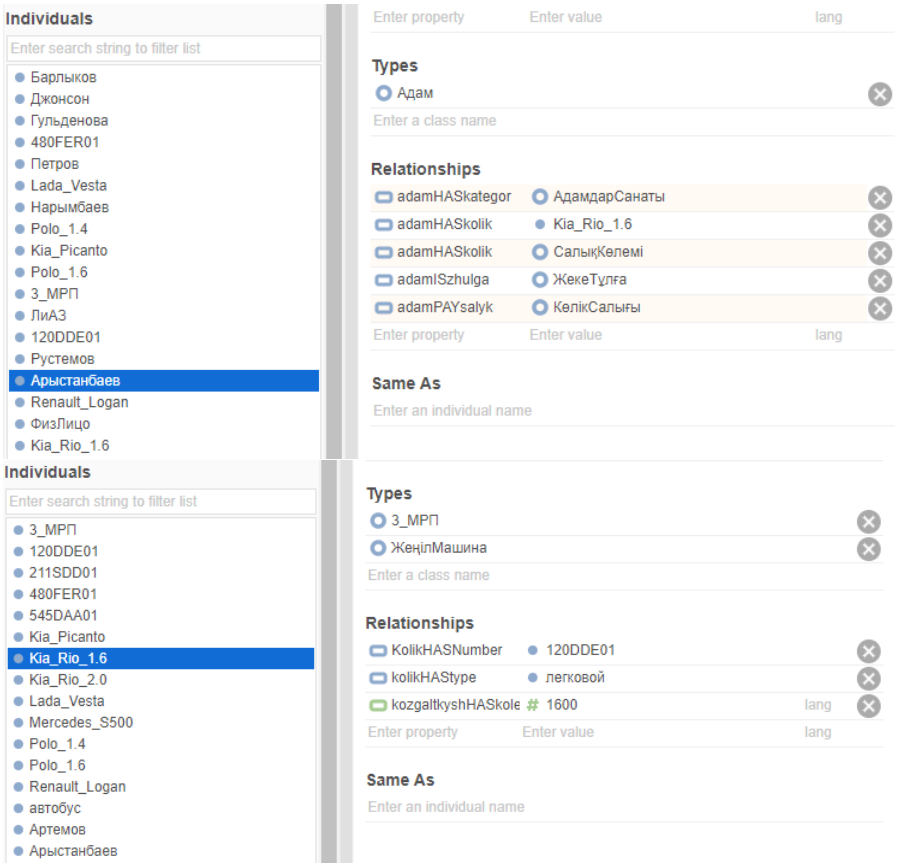



Сурет 2. Объектілер қасиеттері




Сурет 3. Деректер қасиеттері

Пәндік облыс класстарына сәйкес индивидтер қосылып, олардың арасындағы қатынастар көрсетілді (Сурет 4).




Сурет 4. Индивидтер беті

«Көлік салығы» пәндік облысы бойынша құрылған онтология негізінде табиғи тілде қажетті ақпарат алуға мүмкіндік беретін жүйе жасалды. Ол үшін Java программалау тілі қолданылды. Java – тілінде онтологиялармен жұмыс істеуге арналған «Jena ontology API» таңдалды. Ақпараттық жүйе Apache Netbeans IDE программалау ортасында жазылды. Онтологиядан деректерді алу үшін SPARQL сұралым тілі қолданылды.

SPARQL (ағылшын тілінен рекурсивті акроним. SPARQL Protocol and RDF Query Language) - RDF моделі бойынша ұсынылған деректерге сұрау салу тілі, сондай-ақ осы сұрау салулар мен оларға жауап беруге арналған хаттама. SPARQL-бүкіләлемдік ғаламтор консорциумының (W3C) ұсынысы және семантикалық веб технологиясының бірі[4]. SPARQL семантикалық вебті көрудің маңызды элементтерінің бірі болып табылады - жобаның бастамашысы және бастамашысы Тим Бернерс-Ли 2006 жылдың мамыр айында сұхбатында "SPARQL үлкен өзгеріс әкеледі"деп атап өтті [5].

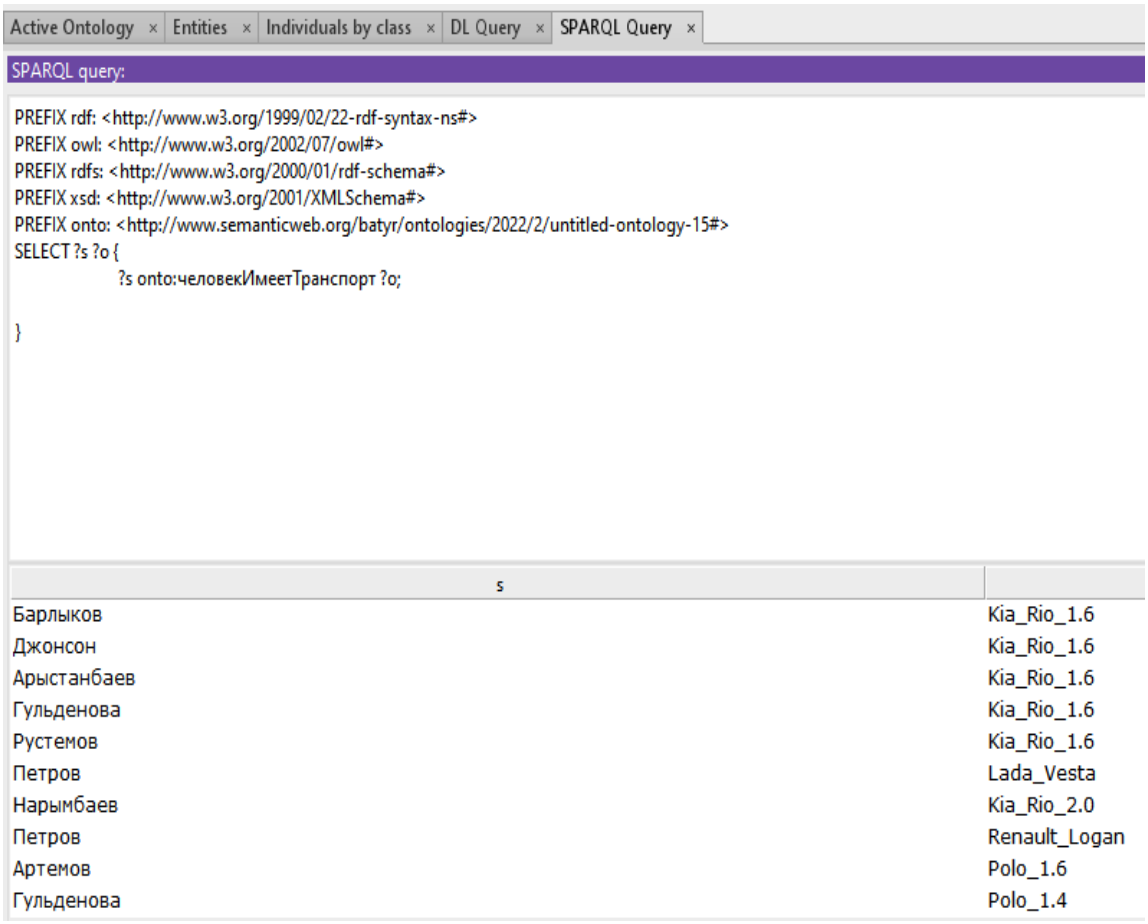
Жұмыста көрсетілген онтология үшін сұратылым префикстері төмендегідей жазылды:

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX owl: <http://www.w3.org/2002/07/owl#>

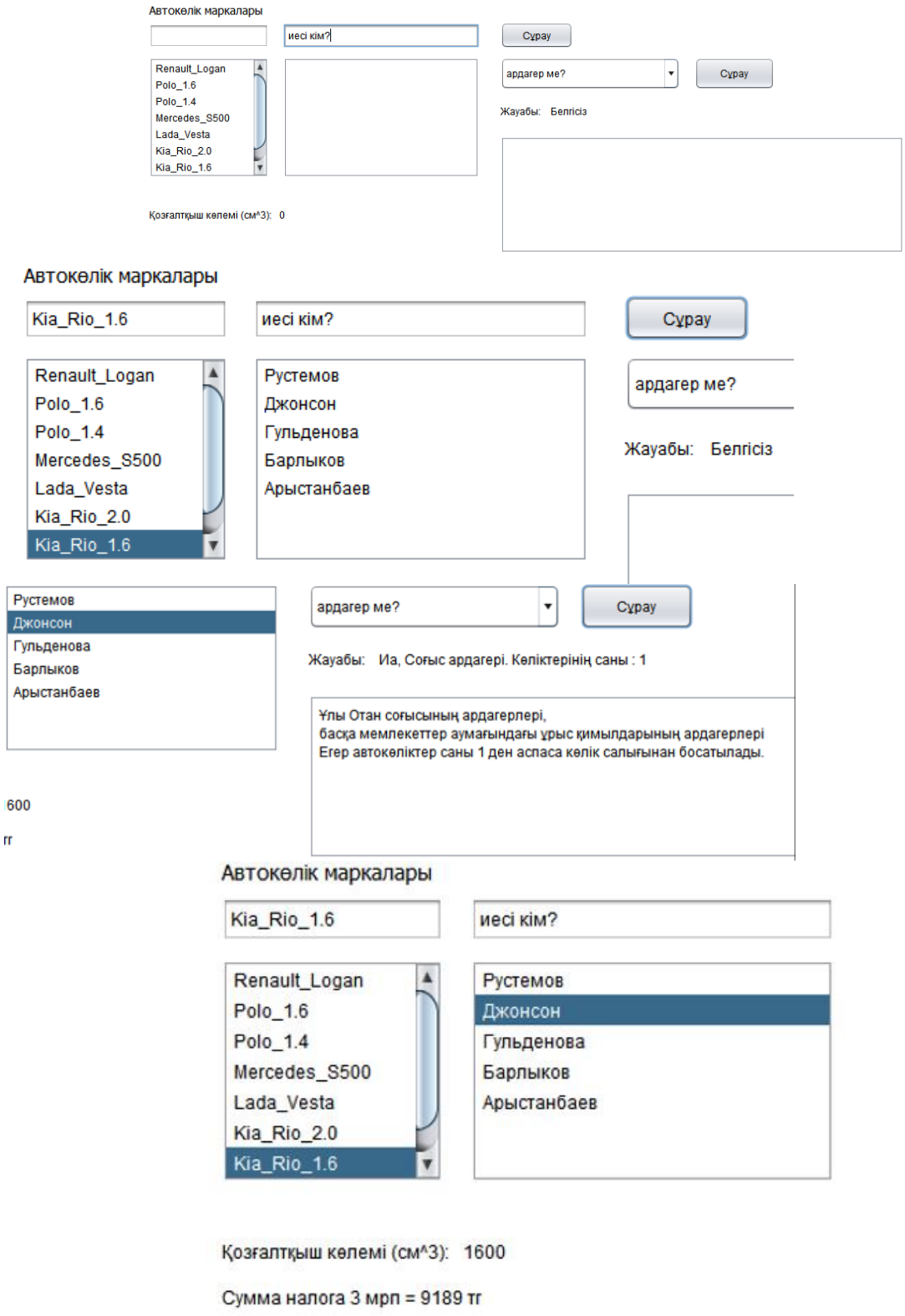
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX onto: http://www.semanticweb.org/batyr/ontologies/2022/2/untitled-ontology-15#

Мысалы, белгілі бір адамға тиесілі көліктер тізімін алу үшін келесі сұратылым жасалынады (Сурет 5).




Сурет 5. SPARQL сұратымдарын қолдану мысалы

Netbeans ортасында java тілінде сұратылымдар жасалып, онтологиядан деректер алу үшін функция құрылды. Пәндік онтология негізінде жасалған «Көлік салығы» бойынша табиғи тілде сұратымдар жасауға мүмкіндік беретін жүйенің фрагменттері (Сурет 6) да көрсетілген.




Сурет 6. «Көлік салығы» бойынша табиғи тілде сұратымдар жасауға мүмкіндік беретін жүйенің фрагменттері

Көрсетілген жұмыста «Көлік салығы» пәндік саласы бойынша негізгі түсініктер мен олардың арасындағы қарым-қатынас және соларға сәйкес ережелер анықталып, солардың негізінде пәндік облыстың

онтологиялық моделі құрылды, білімдер базасы жасалды. Білімдер базасына сұратымдар жасау арқылы табиғи тілде машинаның түрі, көлемі, адамдардың санатына қарай көлік салығы бойынша ақпарат алуға мүмкіндік беретін жүйе жасалды.

**Әдебиеттер тізімі**

1 OWL Язык Сетевых онтологий: Варианты использования и требования. — Текст : электронный // https://www.w3.org/2006/04/OWL_ UseCases-ru.html : [сайт]. — URL: (дата обращения: 23.05.2022).

2 Date C.J. Deit K.D. Vvedenie v sistemy baz dannykh // M.: Izd. Dom «Viliams», 2001. – 72 s.

3 OWL Web Ontology Language. – URL: https://www.w3.org/TR/2004/REC-owlfeatures-20040210/

4 W3C Semantic Web Activity Publications - www.w3.org/2001/sw/Specs.html (англ.). W3C. — Перечень публикаций W3C по проекту семантической паутины.

5 Berners-Lee looks for Web's big leap web.archive.org/web/20070930221904/http://news.zdnet.co.uk/internet/0,1000000097,39270671,00.htm (англ.). — Интервью Тима Бернерса-Ли.

*ОӘК 81’373.2*
***[1]Орынбай Л. О., [2]Сайранбекова А. Д.,***
***[3]Елибаева Г. К., [4]Бекманова Г. Т.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]laura.aktobe.kz@gmail.com, [2]sairanbekova98@gmail.com,*
*[3]gaziza_y@mail.ru, [4]gulmira-r@yandex.kz*

## ҚАЗАҚ ЕСІМДЕРІНІҢ СЕМАНТИКАЛЫҚ БАЗАСЫНЫҢ ҚҰРЫЛЫМЫН АНЫҚТАУ ЖОЛДАРЫ

**Аңдатпа.** Кез-келген адам үшін қызықты болып табылатын тақырыптардың бірі ол өз есімінің шығу тарихын білу немесе ғылыми тілде айтатын болсақ, антропонимия. Аталған мақалада ономастиканың үлкен бөлімі ретінде қарастыралатын антропонимияға тоқталады. Зерттеу жұмыстарының нәтижесінде қазақ есімдерінің семантикалық базасының құрылымы жасалды. Құрылым бойынша есімдер шығу тегі қай елдің тілдерінен кіріккен екендігіне байланысты, сонымен қатар лексикалық мағынасына, өзіндік қасиеттеріне қарай топтастырылды.

**Түйін сөздер:** ономастика, антропонимия, есімдер, семантикалық талдау, семантикалық база.

УДК 81’373.2
***[1]Орынбай Л. О., [2]Сайранбекова А. Д.,***
***[3]Елибаева Г. К., [4]Бекманова Г. Т.***
*Евразийский национальный университет им. Л. Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]laura.aktobe.kz@gmail.com, [2]sairanbekova98@gmail.com,*
*[3]gaziza_y@mail.ru, [4]gulmira-r@yandex.kz*

## СПОСОБЫ ОПРЕДЕЛЕНИЯ СТРУКТУРЫ СЕМАНТИЧЕСКОЙ БАЗЫ КАЗАХСКИХ ИМЕН

**Аннотация.** Одной из тем, которая интересна любому человеку, является изучение истории происхождения его имени или антропонимия, если говорить на научном языке. В данной статье речь идет об антропонимии, которая рассматривается как большой раздел ономастики. В результате исследовательской работы была разработана структура семантической базы казахских имен. По структуре имена группировались в зависимости от того, из каких языков страны они были интегрированы по происхождению, а также по лексическому значению, по собственным качествам.

**Ключевые слова:** ономастика, антропонимия, имена, семантический анализ, семантическая база.

*UDC 81’373.2*
***[1]Orynbay L., [2]Sairanbekova A.,***
***[3]Yelibayeva G., [4]Bekmanova G.***
*L. N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]laura.aktobe.kz@gmail.com, [2]sairanbekova98@gmail.com,*
*[3]gaziza_y@mail.ru, [4]gulmira-r@yandex.kz*

## METHODS OF DETERMINING THE STRUCTURE OF THE SEMANTIC DATABASE OF KAZAKH NAMES

**Abstract**. One of the topics that is interesting to any person is the study of the history of the origin of his name or anthroponymy, if we speak in scientific language. In this article we are talking about anthroponymy, which is considered as a large section of onomastics. As a result of the research work, the structure of the semantic database of Kazakh names was developed. According to the structure, the names were grouped depending on which languages of the country they were integrated from by origin, as well as by lexical meaning, by their own qualities.

**Keywords:** onomastics, anthroponymy, names, semantic analysis, semantic base.

### Кіріспе

Кез-келген тілдің лексикалық қорының маңызды бөлігінің бірі – жалқы есімдер. Оларсыз адамның қарым-қатынасы мүмкін емес. Қоршаған ортадағы табиғаты жағынан бірдей болмыстарды бір-бірінен ажырату үшін, ажырату функциясын атқаратын ерекше сөздер ойлап табу керек.

Ономастика – кез-келген жанды немесе жансыз затты, құбылысты жеке белгілеу үшін қызмет ететін тиісті атауларды зерттейтін тіл білімінің саласы. Ономастиканың объектісі – тілдің жалпы лексикалық қорының өзіндік ішкі жүйесін құрайтын барлық категориялардың жалқы есімдері [Мадиева, 2003, 9 б]. Өз атымен аталған объектіні басқа объектілер арасында бөліп көрсетуге, оны дараландыруға және сәйкестендіруге қызмет ететін барлық сөздер, сөз тіркестері немесе сөйлемдер жалқы есім мағынасында түсініледі. Жалқы есімдер ономастикалық тұрғыда әмбебап болып табылады, өйткені оларсыз бірде-бір тіл немесе мәдениет болмайды [Жанұзақов, 1971].

Ономастика пәні – табиғи және жасанды атауларды қалыптастыру үшін қолданылатын тілдік құралдар жүйесі, олардың құрылымы, семантикалық және сөзжасамдық модельдер, тиісті атаулардың өзара байланысы, септеу ерекшеліктері, тиісті атаулардың синтаксисі, орындалатын функциялар [Мадиева, 2003, 9 б].

Аталған объектіге байланысты жалқы есімдер әртүрлі санаттарға бөлінеді [Керимбаев, 1995]. Оның ішінде, антропоним – бұл адамның немесе адамдар тобының жеке атын, әкесінің атын, тегін, лақап атын, криптонимін (құпия аты), андронимді (әйелінің аты, лақап аты немесе күйеуінің тегі), гинеконимді (ер адамның аты, анасының аты немесе лақап аты), патронимді (әкесінің немесе ата-бабаларының атынан (лақап атымен) құрылған адамның аты) білдіретін ұғым [Мадиева, 2003, 10 б].

Қазақ антропонимиясы – қазақ тіліндегі антропонимдердің, яғни адамды қазақ тілінде атау үшін қолданылатын жалқы есімдер жиынтығы. Қазақ антропонимиясы алуан түрлілігімен ерекшеленеді және байырғы қазақ және кірме, басқа тілден енген есімдерді де (негізінен араб немесе парсы тілінен) қамтиды. Қазақ есімдерінің басым тобы халқымыздың өткені мен бүгінінің арасындағы мәдени-тарихи өмірінің көрінісін анық көрсетеді [Жанұзақ, 2006].

**Қазақ есімдерінің семантикалық базасын әзірлеу**

Қазақ есімдерін семантикалық талдау барысында, авторлар антропонимдердің түрлі тұрғыдан жіктелуін қарастырды, шығу тектері мен есімдердің мағынасын аша отырып, соған сәйкес сөздердің қасиеттері анықталып, семантикалық базаның алғашқы құрылымдары жасалды. Зерттеу жұмысы барысында әзірленіп жатқан семантикалық базаның фрагменті 1-суретте көрсетілген. Бұл базада қазақ есімдерінің мүмкін болатын 100-ге жуық қасиеттері белгіленді [Жанұзақ, 2004, 6-10 бб].

| Есімдер | Ер адамның аты | Әйел адамның аты | Мағынасы | Өзгеріске ұшыраған | Негізгі түбірі | Араб тілінен | Табиғатқа байланысты | Аспан денелеріне (Ай, жұлдыз) қатысты есімдер | Өсімдік атауы | Үй-жиһаздарымен, мүліктерге байланысты | Сұлу болсын деп |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Айдария | | 1 | айдай сұлу, теңіз, үлкен өзен | | | | 1 | 1 | | | 1 |
| Айжамал | | 1 | айдай сұлу, әдемі | | | 1 | | 1 | | | 1 |
| Айжан | | 1 | айдай сұлу, әдемі | | | | | 1 | | | 1 |
| *Айжания* | | 1 | айдай сұлу, әдемі | 1 | Айжан | | | 1 | | | 1 |
| Айзада | | 1 | Айдай перзент | | | | | 1 | | | 1 |
| Айзиба | | 1 | Ай сияқты сұлу | | | | | 1 | | | 1 |
| Айзия | | 1 | Ай сәулелі сұлу | | | 1 | 1 | 1 | | | 1 |
| Айкүміс | | 1 | Күміс сияқты | | | | | 1 | | | 1 |
| Айман | | 1 | белгілі, әйгілі, даңқты; ар, ант, серт | | | 1 | | | | | |
| Айманай | | 1 | белгілі, әйгілі, | | | 1 | | 1 | | | 1 |
| Аймангүл | | 1 | белгілі, әйгілі, | | | 1 | | | 1 | | 1 |

**Сурет 1.** Қазақ есімдерінің семантикалық базасының көрінісі

Қазақ тіліне кірігіп кеткен шығу тегі әртүрлі тілдерден болатын есімдер және көне қазақ тіліндегі есімдерді 1-кестеден көруге болады.

Кесте 1. Әлем тілдерінен қазақ тіліне кіріктірілген есімдер

| **Шығу тегіне байланысты топ** | **Мысалы** |
|---|---|
| Көне қазақ тілі | Шажабек, т.б. |
| Көне түркі тілі | Алдар, Алтай, Баян, Руслан(а), т.б. |
| Көне герман тілі | Берта, Эмма, Эдуард, Эрик, т.б. |
| Көне еврей тілі | Адам, Айша (Ғайша, Қайша), Анна, т.б. |
| Батыс Еуропадан енген | Долорес, Дина, Артур, Марат, т.б. |
| Ағылшын тілі | Айзақ, Жастина, Хамер, т.б. |
| Латын тілі | Адриан, Агнесса, Белла, Флора, т.б. |
| Испан тілі | Эльвира, Диас, т.б. |
| Неміс тілі | Грета, Верт, Ада, т.б. |
| Скандинавия тілі | Ольга, Инга, Елизавета, т.б. |
| Кельт тілі | Артур, т.б. |
| Грек тілі | Арсен, Ағлия, Арина, Лариса, Мирон, т.б. |
| Иран тілі | Бағдат, Құбан, Заида, Фируза және т.б. |
| Араб тілі | Аббас (Ғаббас, Қаппас, Қапбас), Даниял, Жәмила, Нағима және т.б. |
| Алтай тілі | Мөңке, т.б. |
| Қалмақ тілі | Адық, Бату, Мерген, т.б. |
| Қырғыз тілі | Жолай, Манас, т.б. |
| Монғол тілі | Құрал, Ойрат, Ноян, Жирен, т.б. |
| Тибет тілі | Тайшық, Шойбек, т.б. |
| Манчжурия тілі | Құралай, Тана, т.б. |
| Түрік тілі | Бекеш, Севиль, Эсмира, т.б. |

| | |
|---|---|
| Түркмен тілі | Темір, Дулат, Аттила, Дамира, т.б. |
| Азербайжан тілі | Бейқан, т.б. |
| Тәжік тілі | Абира, Аза, Майса, Нияз, т.б. |
| Татар тілі | Ишмурат, Алсу, т.б. |
| Грузин тілі | Дила, т.б. |
| Орыс тілі | Замир, Владлен, Бәтес, т.б. |
| Украин тілі | Оксана, т.б. |

Қазақ тіліндегі есімдердің арасында бірнеше түбірден құралған біріккен сөздер де бар, мұндай есімдерді авторлар жасап жатқан кестеде толық мағынасын ашып, толықтыру үстінде (2-сурет). Олардың әр түбірі әр тілден енген болуы да мүмкін. Мысалы, Аймира есімі “ай” – қазақ тіліндегі аспан денесі және “мир” – орыс тіліндегі бейбітшілік, тыныштық деген мағынаны білдіретін екі сөзден құралған. Біріккен сөз болып табылатын есімдердің басқа мысалдары ретінде келесілерді көрсетуге болады: Айнажан, Ақбота, Анаргүл, Бибігүл, Абылғазы, Абдолла және т.б.

| Есімдер | Бірнеше түбірлі (күрделі сөз) | 1-түбір | тіл | мағынасы | сөз табы | 2-түбір | тіл | мағынасы | сөз табы |
|---|---|---|---|---|---|---|---|---|---|
| Амангүл | 1 | аман | қазақ | сау, есең, саламат | с.е. | гүл | тәжік | роза | з.е |
| Аманғайша | 1 | аман | қазақ | сау, есең, саламат | с.е. | ғайша (айша) | көне еврей, араб | сүруші, өмірге | з.е |
| Аманбала | 1 | аман | қазақ | сау, есең, саламат | с.е. | бала | қазақ | | з.е. |
| Анар | | | | | | | | | |
| Анаргүл | 1 | анар | иран | дәні бар алма тектес жеміс | з.е. | гүл | тәжік | роза | з.е |
| Анархан | 1 | анар | иран | дәні бар алма тектес жеміс | з.е. | хан | қазақ | лауазым | з.е |
| Анаржан | 1 | анар | иран | дәні бар алма тектес жеміс | з.е. | жан | иран | тірінің бойындағы | з.е. |
| Гүлнәр | 1 | гүл | тәжік | роза | з.е | анар | иран | дәні бар алма тектес | з.е. |

**Сурет 2.** Бірнеше түбірлі есімдердің мысалы

Лексикалық тұрғыдан және есімдердің шығу тарихын зерттей отыра, келесідей қасиеттер анықталды:

1. Тарихпен, ұлтпен байланысты есімдер:

a. Ру-тайпа атауларына (этноантропоним): Адай, Алшынбай, Найман және т.б.

b. Тарихи оқиғамен байланысты есімдер: Жеңіс, Сайлау және т.б.

c. Діни есімдер: Абыз, Ажар, Айша және т.б.

d. Салт-дәстүрге қатысты: Айдар, Кекілбай, Тұлымбек, Ақмарал және т.б.

2. Әдеби шығармалармен байланысты есімдер:

e. Аңыз-ертегі кейіпкерлері (миф) есімдері: Алдар, Алпамыс, Арина және т.б.

f. Әдеби есімдер: Біржан, Әйгерім, Балқадиша, Баян және т.б.

3. Табиғатпен байланысты есімдер:

a. Жер-су атауларына байланысты есімдер: Алтай, Бағдат, Еділ және т.б.

b. Табиғатқа, табиғат құбылыстарына байланысты есімдер: Айдария, Айзия, Айнұр, Айсәуле және т.б.

c. Аспан денелеріне (Ай, жұлдыз) қатысты есімдер: Айбас, Алтынай, Базарайым, Бәдер және т.б.

d. Ауа-райына байланысты есімдер: Боран, Дауылбай және т.б.

4. Бағалы заттармен байланысты есімдер:

a. Асыл тас аттары: Алмас, Маржан және т.б.

b. Әшекейлік бұйымдар: Айзере, Айна, Бәтес және т.б.

c. Металл (темір) атауы: Болат, Айкүміс, Алтын және т.б.

d. Валюта атауы: Динар, Динара және т.б.

5. Аңдар және құстармен байланысты есімдер:

a. Төрт түлік мал және оның төлдеріне қатысты: Ақбота, Аққозы, Дөненбай, Жанбота және т.б.

b. Жыртқыш жануарлар: Арыстан, Қасқырбай, Байбөрі және т.б.

c. Аң аты: Бұғыбай, Ақмарал және т.б.

d. Құс аты: Бүркіт, Бұлбұл, Лашын және т.б.

6. Өсімдік атауымен байланысты есімдер: Гүлім, Қызғалдақ, Раушан, Роза және т.б.

7. Жеміс-жидек атауларымен байланысты есімдер: Алма, Анар, Қарақат және т.б.

8. Түрлі тағам аттарына байланысты есімдер: Алуа, Әсел, Шекер, Шыран және т.б.

9. Үй-жиһаздарымен, мүліктерге байланысты есімдер: Айнагүл, Табақбай және т.б.

10. Түс атаулары бар есімдер: Ақбота, Сары және т.б.

11. Дене мүшесі, адам ағзасына байланысты қойылған есімдер: Аққал, Алтыншаш, Ботагөз және т.б.

12. Отбасына байланысты есімдерді келесідей анықтауға болады:

a. Ер бала тұрақтамағаннан қойылатын есімдер: Тоқтар, Тұрар, Өтеген, Төлеміс және т.б.

b. Әке-шешесінің неше жасқа келгенде туғанына байланысты: Алпыс, Жетпісбай, Қырықгүл

c. Үйде бала көп болсын деп қойылатын есімдер: Байжан және т.б.

d. Отбасында ер бала көп болсын деп қойылатын есімдер: Байқошқар, Қошқарбай және т.б.

e. Үйдің тұңғышына қойылатын есімдер: Әбубәкір, Әуел және т.б.

f. Соңынан еріп туған балаға қойылатын есімдер: Ергеш, Жалғас және т.б.

g. Үйдің кенжесіне қойылатын есімдер: Кенжебек, Балбала, Балшекер және т.б.

h. Аңсап дүниеге келген балаға қойылатын есімдер: Аңсаған, Аңсар және т.б.

i. Жалқы туған бала қойылатын есімдер: Біржан, Балқадиша және т.б.

j. Егіз туған балаларға қойылатын есімдер: Егізбай, Қосжан, Жақып және т.б.

k. Баланың туған күні, айы, жылына байланысты қойылған есімдер: Әшір, Бейсенбі, Әдита және т.б.

l. Туыстық атаулармен байланысты: Абира, Жиенбай, Туғанбай және т.б.

Есім беруді түрлі тілектермен, мақсаттармен байланысты сөздерді қолдану да көптеп кездеседі [Жанұзақов, 1965]. Ата-ананың сәбиіне батырлық (Айбар, Айтемір, Айбын, Қайсар, Жігер, т.с.с.), ақылдылық пен білімділік (Ақылбек, Әбутәліп, Ағила, Бекдана, т.с.с.), сұлулық (Ажар, Айбике, Әдемі, т.с.с.), ұзақ өмір (Амантұр, Әбілқайым, Айша, Амангүл, Жанұзақ, т.с.с.), құрмет (Әшім, Ардақ, Әзиза, Ғалия, т.с.с.), байлық (Байахмет, Аида, Әлиша, т.с.с.), көп дос (Байдос, Аниса, т.с.с.), басшылық (Бижан, Дегдар, Әмира, т.с.с.), әділділік пен адалдық (Агнесса, Арна, Әдила, т.с.с.), бақыт (Бағымша, Бақтыгүл, Бақытбек, т.с.с.), жомарттық (Ақжібек, Анна, Аршат. т.с.с.), мейірімділік (Мейірім, Әмина, Мейіргүл, т.с.с.), еңбекқорлық (Әмила, Бәдиға. т.с.с.), халқының баласы болсын деп (Еламан, Елдар, Елжас, т.с.с.), өнерпаздық (Бұлбұл, Өнербай, т.с.с.) секілді жақсы қасиеттерді тілеу ниетінен туындаған.

**Қорытынды**

Зерттеу жұмыстарының нәтижесінде, ономастика, қазақ антропонимиясы терминдеріне анықтама берілді. Жүргізілген зерттеулер барысында жасалған қазақ есімдерінің семантикалық базасының құрылымы сипатталды. Бұл база есімдердің келесі

қасиеттері бойынша жіктелді: шыққан тегімен (тілі), тарихпен, ұлтпен, әдеби шығармалармен, табиғатпен, бағалы заттармен, аңдар және құстармен, өсімдік атауымен, жеміс-жидек атауларымен, түрлі тағам аттарына, үй-жиһаздарына, мүліктерге, түс атауларына, дене мүшесіне, адам ағзасына, отбасына байланысты есімдер. Сонымен қатар семантикалық базада ата-ананың баласына деген ізгі тілектері негізінде қалыптасқан есімдер де қамтылды.

Бұл зерттеуді Қазақстан Республикасы Білім және ғылым министрлігінің Ғылым комитеті қаржыландырады (№BR11765535 грант).

**Әдебиеттер тізімі**

1. Мадиева Г. Б., Супрун В. Теория и практика ономастики //Алматы: Қазақ университеті. – 2003.
2. Жанұзақов Т.Ж. Қазақ есімдерінің тарихы. Алматы, Ғылым, 1971 – 218 бет. (Лингвистикалық және тарихи этнографиялық талдау).
3. Керимбаев Е.А. Казахская ономастика в этнокультурном, номинативном функциональном аспектах. – Алматы, 1995. – 248 с.
4. Жанұзақ Т.Ж. Қазақ ономастикасы – Казахская ономастика. I том. Астана, IC-Сервис ЖШС, 2006 ж. – 400 бет.
5. Жанұзақ Т.Ж. Есімдер сыры – Тайны имен. Алматы, 2004. – 208 бет.
6. Жанұзақов Т. Қазақ тіліндегі жалқы есімдер. – Алматы: ХОЗУ СМ КазССР, 1965. – 145 б.

# КОМПЬЮТЕРЛІК ЖҮЙЕЛЕРДІҢ ҰЛТТЫҚ ЛОКАЛИЗАЦИЯСЫ МЕН ТЕРМИНОЛОГИЯ

# НАЦИОНАЛЬНАЯ ЛОКАЛИЗАЦИЯ КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕРМИНОЛОГИЯ

# NATIONAL LOCALIZATION OF COMPUTER SYSTEMS AND TERMINOLOGY

*УДК 811.512*

***[1]Сеилов Ш. Ж., [2]Ахметова Ж. Ж., [3]Зұлпыхар Ж.Е.***
*Евразийский национальный университет им. Л.Н.Гумилева*
*Нур-Султан, Казахстан*
*[1]seilov_shzh@enu.kz, [2]zaigura@mail.ru, [3]zulpykhar_zhye@enu.kz*

## МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ ПЕРЕВОДА ТЕРМИНОВ НА КАЗАХСКИЙ ЯЗЫК И ИХ РАЗВИТИЕ В СФЕРЕ ЦИФРОВОЙ ЭКОНОМИКИ

**Аннотация.** В статье рассматриваются вопросы перевода на казахский язык и развития терминов цифровой экономики в настоящее время. В современных условиях формирование международной терминологии в казахском языке является одной из противоречивых проблем. После перехода к рыночной экономике на казахском языке появилось множество иностранных (английских) и международных терминов, некоторые из которых не переводятся и не подлежат переводу, в том числе те, которые полностью соответствуют казахскому языку.

Другой проблемой терминологии в настоящее время является разделение ее на общенаучную, общетехническую и отраслевую терминологию. В связи с появлением новых наук и открытием новых явлений, возникает необходимость дать название новым понятиям, явлениям, а многие слова имеют определенную качественную специфику, смысл которых преобразуются в отраслевые.

В данной статье предлагаются пути рассмотрения и решения таких проблем, приводятся примеры некоторых проблем. Дается определение общенаучным и общетехническим, отраслевым терминам.

С целью рассмотрения и представления путей решения таких проблем в данной статье описывается научная работа, которая реализуется в виде проекта «Словарь терминов цифровой экономики" на факультете информационных технологий Евразийского национального университета им. Л.Н.Гумилева под руководством декана факультета, профессора Сеилова Ш.Ж.

Основной целью проекта является разработка словаря наиболее часто используемых терминов по цифровым технологиям в сфере экономики и повседневной жизни.

**Ключевые слова:** терминология, термины, технические термины, научные термины.

*UDC 811.512*

***[1]Seilov Sh., [2]Akhmetova Zh., [3]Zulpykhar Zh.***
*L.N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]seilov_shzh@enu.kz, [2]zaigura@mail.ru, [3]zulpykhar_zhye@enu.kz*

## METHODOLOGICAL PROBLEMS OF TRANSLATING TERMS INTO KAZAKH AND THEIR DEVELOPMENT IN THE FIELD OF DIGITAL ECONOMY

**Abstract.** The article deals with the issues of translation into the Kazakh language and the development of digital economy terms at the present time. In modern conditions, the formation of international terminology in the Kazakh language is one of the controversial problems. After the transition to a market economy, many foreign (English) and international terms appeared in the Kazakh language, some of which are not translated and cannot be translated, including those that fully correspond to the Kazakh language.

Another problem of terminology at present is its division into general scientific, general technical and industry terminology. In connection with the emergence of new sciences and the discovery of new phenomena, there is a need to give a name to new concepts, phenomena, and many words have a certain qualitative specificity, the meaning of which is transformed into industry.

This article suggests ways to consider and solve such problems, and provides examples of some problems. The definition of general scientific and general technical, industry terms is given.

In order to consider and present ways to solve such problems, this article describes the scientific work that is being implemented in the form of the project "Dictionary of Terms of Digital Economy" at the Faculty of

Information Technology of the L.N. Gumilyov Eurasian National University under the guidance of the Dean of the Faculty, Professor Seilov Sh.Zh.

The main goal of the project is to develop a dictionary of the most commonly used terms on digital technologies in the field of economics and everyday life.

**Keywords:** terminology, terms, technical terms, scientific terms.

В настоящее время наблюдается тенденция широкого проникновения в нашу повседневную жизнь цифровых технологий. Наблюдается усиление информационного воздействия на человека и общество. Все это не может не сказаться на терминологию, которая естественным образом влияет на развитие науки и общества в целом. Такой эффект приводит к изменениям в терминологии, в результате которых обновляется научное представление. Эти процессы отражают развитие теории и практики различных отраслей науки.

В современных условиях формирование международной терминологии в казахском языке является одной из противоречивых проблем. После перехода к рыночной экономике на казахском языке появилось множество иностранных (английских) и международных терминов, некоторые из которых не переводятся и не подлежат переводу, в том числе те, которые полностью соответствуют казахскому языку. Такие конкурирующие синонимы часто встречаются в языке средств массовой информации. Поэтому проблема создания и распространения национальной экономической цифровой терминологии является проблемой не только экономистов, но и всех специалистов в области терминологии.

Вопросы терминологии были актуальны еще в начале XX века. Об этом свидетельствует статья «Мемлекет термин комиссиясы және оның жұмысы» *(пер. на рус. Государственная комиссия и ее работа).*, которая была опубликована в 1934г. в журнале «Ауыл мұғалімі» *(пер. на рус. Сельский учитель).*

31 декабря 2019 г. Правительством Республики Казахстан принята «Государственная программа реализации языковой политики в Республике Казахстан на 2020-2025 гг.» - для проведения гармоничной языковой политики, направленной на модернизацию казахского языка, дальнейшее повышение языковой культуры и развитие языкового капитала на основе латинского графического алфавита с обеспечением полноценной деятельности казахского языка, как государственного. В данной программе важнейшей проблемой является соблюдение принципов включения в национальную терминологическую систему активно используемых отраслевых терминов, определение путей и конкретных закономерностей создания национального термина,

унификация национальной терминологической системы путем утверждения терминов и пропаганды утвержденных терминов и обеспечение их массового доступа [1].

Другой проблемой терминологии в настоящее время является разделение ее на общенаучную, общетехническую и отраслевую терминологию. В связи с появлением новых наук и открытием новых явлений, возникает необходимость дать название новым понятиям, явлениям, а многие слова имеют определенную качественную специфику, смысл которых преобразуются в отраслевые.

Общенаучные и общетехнические термины - это термины, используемые в нескольких областях науки и техники. Отраслевые термины - это термины, присущие только определенной области знаний.

Филологи и специалисты языкознания при переводе международных терминов на казахский язык должны руководствоваться их восприятием отраслевыми экспертами и мнением общественности. Необходимо проводить широкое обсуждение терминов среди населения.

С целью рассмотрения и представления путей решения таких проблем на факультете информационных технологий Евразийского национального университета им. Л.Н.Гумилева под руководством декана факультета, профессора Сеилова Ш.Ж. реализуется научная работа в виде проекта «Словарь терминов цифровой экономики".

Основной целью проекта является разработка словаря наиболее часто используемых терминов по цифровым технологиям в сфере экономики и повседневной жизни.

Задачи проекта:

- привлечение международных экспертов из университетов России и стран тюркского мира для экспертизы и разработки терминов;
- исследование совместимости терминов цифровой экономики в странах тюркского мира;
- разработка терминов цифровой экономики на казахском языке;
- разработка онлайн платформы, предназначенной для обсуждения и принятия обществом разработанных терминов;
- представление терминов для утверждения в терминологической комиссии Республики Казахстан.

Анализ работ Е.Омарова, М.Жумабаева, А.Байтурсынова и других ученых, внесших вклад в развитие казахской терминологии, указывает на необходимость изучения истории становления и этапов внедрения международных терминов в казахский язык .

В становление и развитие казахской терминологии в начале XX века внесли огромный вклад труды ученых, представителей партии

«Алаш», принципы, предложенные ими, и в наши дни до сих пор не утратили своей значимости. Причины этого определены следующим образом:

1) Время, в которое они жили и работали, является для языка переходным периодом. В связи с этим, решения принятые ими были актуальны и результативны чему свидетельствуют исторические материалы исследования современных ученых.

2) Эпоха, в которой они жили, был периодом характеризующийся сложностями в развитии казахской истории и культуры. Особенностью того периода являются натиски политической идеологии. Несмотря на существующие проблемы, этот период был периодом решения проблем языка в интересах казахского народа, периодом становления научных идей и решений.

3. Впервые в этот период произошли противостояния в развитии языка. Основной причиной которых являются стремление к сохранению национального языка в чистом виде, то есть в национальном стиле, призыв к правильному озвучиванию специфических звуков казахского языка, образование терминов и слов, соответствующих природе казахского языка. Таким образом, благодаря научным воззрениям представителей партии «Алаш», международные термины были приняты и эффективно использовались в языке [2].

На наш взгляд, участники проекта выработали правильную методологию перевода и технологию согласования терминов, которая была определена еще основоположником казахской лингвистики и литературоведения, и профессиональной журналистики Алиханом Байтурсыновым, которая имеет следующий принцип: «Переводить то, что переводится, не придумывать всем иностранным терминам новые казахские слова».

В ходе выполнения проекта будут проанализированы переводы терминов на языки стран Турции, Азербайджана, Узбекистана, Кыргызстана, Туркменистана, а также на русский язык. Предполагается привлечение экспертов из ведущих университетов указанных государств.

В настоящее время при создании терминологических систем на казахском языке используются языковые средства из трех источников:

1) из литературного языка (и диалектов),

2) с европейских языков (в основном, с английского),

3) с русского языка.

Поскольку XXI век - это век информационных технологий и цифровизации, вопросы терминологии нуждаются в постоянном развитии и совершенствовании.

В наши дни основными проблемами перевода терминов на казахский язык являются:

Первая проблема – это последовательность, несоблюдение последовательности, что можно наглядно продемонстрировать в следующем реальном примере, который показан на рисунках 1 и 2.

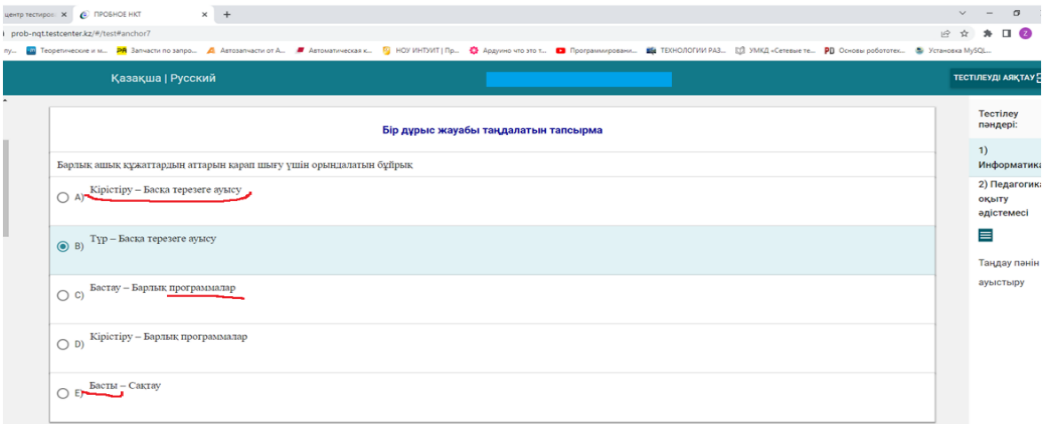



Рис.1-Тест по дисциплине "Информатика" Национального квалификационного тестирования (https://prob-nqt.testcenter.kz/)
(Test in the discipline "Informatics" of the National qualification Testing)

На рис.1 в экзаменационных вопросах по «Информатике» команда ***Вставка- Переход на другое окно*** на казахском языке, а на рис. 2 другая команда ***Файл-Переименовать*** написана на русском языке, что свидетельсвует о несоблюдении системности в терминологии на казахском языке.

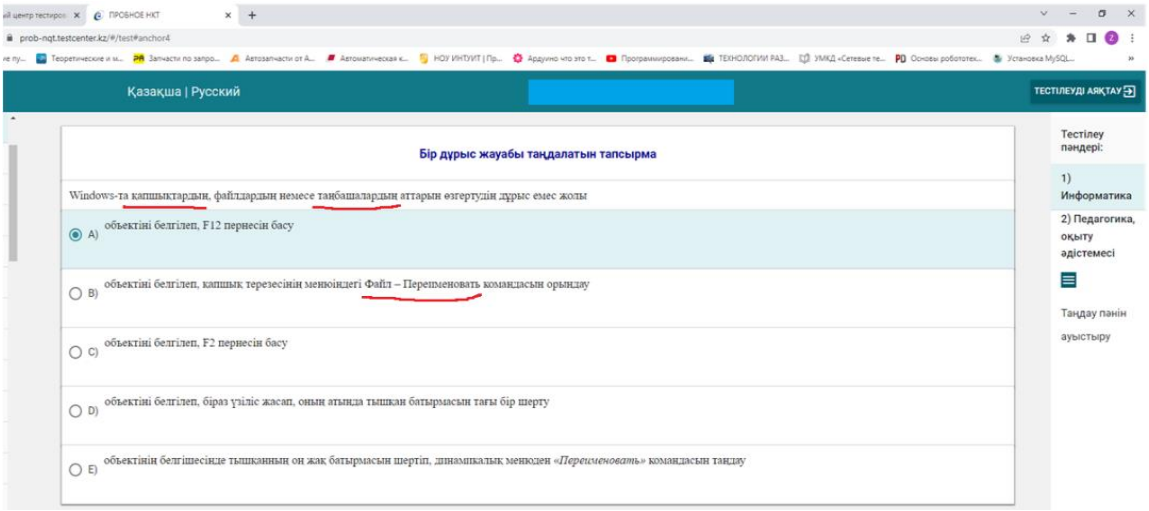



Рис. 2 - Тест по дисциплине "Информатика" Национального квалификационного тестирования (https://prob-nqt.testcenter.kz/)
(Test in the discipline "Informatics" of the National qualification Testing)

Данный пример является одним из многих примеров, которые встречаются как в науке и образовании, так и в повсевдневной жизни.

Вторая проблема - перевод некоторых международных терминов на казахский язык и его неприменение среди населения.

Третья проблема - не унифицированность национальной терминологической системы.

Четвертая проблема - создание республиканских терминологических комиссий по каждой отрасли и осуществление надзора за полным использованием терминов, утвержденных этой комиссией [3].

Несоблюдение системности, последовательности в изложении терминов, унификация национальной терминологической системы могут создавать дополнительные проблемы для обучающихся. В связи с этим, есть необходимость унифицировать национальную терминологическую систему. Для решения данной проьблемы необходимо создать республиканские терминологические комиссии по каждой отрасли и осуществлять надзор за полным использованием терминов, утвержденных этой комиссией.

В заключении, следует отметить, что принципы, принятые учеными еще в начале XX века, по сей день не утратили своей актуальности, чему могут свидетельствовать слова ученого Ж.Курманбайұлы: «Принципы, установленные учеными XX века, представителями партий «Алаш», несмотря на то, что со временем были дополнены, бесспорно остаются основой для формирования новых принципов перевода».

**Список литературы**

1. Қазақстан Республикасындағы тіл саясатын іске асырудың 2020 - 2025 жылдарға арналған мемлекеттік бағдарламасы (https://adilet.zan.kz/kaz/docs/P1900001045).

2. Ө. Айтбайұлы. Основы казахской терминологий. – Алматы. Издательство «Абзал-Ай», 2014. – 384 стр.

3. Helena Pociechina, Aida Musagułowa. Экономическая терминология в современном казахском языке: феномены постсоветского языкового пространства.

4. Скрины из сайта: https://prob-nqt.testcenter.kz/

*УДК 811.512.122*
***Илиуф Хаджи-Мурат Шаяхметович***
*Государственный университет им. Шакарима*
*Семей, Казахстан*
*murat_20@mail.ru*

# О СОВЕРШЕНСТВОВАНИИ ТЕРМИНОЛОГИИ В КАЗАХСКОМ ЯЗЫКЕ

**Аннотация.** Для того, чтобы язык мог успешно применяться в какой-либо области знания или деятельности для формулирования той или иной теории, обозначения основных понятий и связи между ними, необходима соответствующая терминология. По мнению автора статьи, возвращение казахов к латинской графической системе позволяет не только использовать оптимальный вариант алфавита, в котором реализуется принцип соответствия одной графемы одной фонеме, но и разработать терминологию, отвечающую потребностям современной научной и профессиональной деятельности. Успешное развитие любого языка обеспечивается за счет увеличения словарного фонда, пополнения лексикой, релевантной современным потребностям. Например, предлагается ввести в речевую практику употребление известного в ряде тюркских языков слова-арабизма äyile в качестве социологического термина для обозначения понятия семья.

Литературному языку присуща нормативность, благодаря которой язык сохраняет свою целостность и понятность для всех говорящих на нем. Обращаясь к словам-арабизмам, нередко представленным в речи существенно различающимися фонетическими вариантами, автор считает целесообразным определить в качестве лексической нормы вариант, наиболее близкий по звучанию к оригиналу, что способствует осознанию связи между однокоренными словами-заимствованиями.

Грамматический строй тюркских языков позволяет заимствовать прилагательные, не осложняя их аффиксами прилагательных, но, несмотря на это, некоторые слова европейского происхождения в казахском языке имеют избыточную морфемную структуру. По этой причине в статье приведено мнение об использовании таких заимствований в оригинальной форме по примеру других письменных тюркских языков.

**Ключевые слова:** термин, семья, слова-заимствования, архаизмы, прилагательные.

*ӘОК 811.512.122*
***Илиуф Хажы-Мұрат Шаяхмет-ұлы***
*Шәкәрім атындағы мемлекеттік университеті*
*Семей, Қазақстан*
*murat_20@mail.ru*

# ҚАЗАҚ ТІЛІНДЕГІ ТЕРМИНОЛОГИЯНЫ ЖЕТІЛДІРУ ТУРАЛЫ

**Аңдатпа.** Тіл белгілі бір теорияны тұжырымдау, негізгі ұғымдарды белгілеу және олардың арасындағы байланыс үшін білімнің немесе қызметтің кез-келген саласында сәтті қолданылуы үшін тиісті терминология қажет. Мақала авторының пікірінше, қазақтардың латын графикасына оралуы әліпбидің бір графеманың бір фонемаға сәйкес келу принципі іске асырылатын оңтайлы нұсқасын пайдалануға ғана емес, сонымен қатар қазіргі ғылыми және кәсіби қызметтің қажеттіліктеріне жауап беретін терминологияны әзірлеуге де мүмкіндік береді. Кез-келген тілдің сәтті дамуы сөздік қорын ұлғайту, қазіргі заманғы қажеттіліктерге сәйкес келетін лексикамен толықтыру арқылы қамтамасыз етіледі. Мысалы, бірқатар түркі тілдерінде белгілі арабизм сөзін отбасы ұғымын білдіретін социологиялық термин ретінде қолдануды сөйлеу практикасына енгізу ұсынылады.

Норматив әдеби тілге тән, соның арқасында тіл барлық спикерлер үшін тұтастығы мен түсінігін сақтайды. Көбінесе сөйлеуде айтарлықтай ерекшеленетін фонетикалық нұсқалармен ұсынылған арабизм сөздеріне сілтеме жасай отырып, автор лексикалық норма ретінде түпнұсқаға ең жақын нұсқаны анықтаған дұрыс деп санайды, бұл түбір сөздер-қарыз алу арасындағы байланысты түсінуге ықпал етеді.

Түркі тілдерінің грамматикалық құрылымы сын есімдерді сын есімдердің аффикстерімен қиындатпай қарыз алуға мүмкіндік береді, бірақ соған қарамастан, қазақ тіліндегі еуропалық шыққан кейбір сөздер артық морфемалық құрылымға ие. Осы себепті мақалада мұндай қарыз алуды басқа жазбаша түркі тілдерінің мысалы бойынша түпнұсқа түрінде пайдалану туралы пікір келтірілген.

**Түйін сөздер:** термин, отбасы, кірме сөздер, архаизмдер, сын есімдер.

*UDC 811.512.122*
***Iliuf Khazhi-Murat***
*Shakarim State University*
*Semey, Kazakhstan*
*murat_20@mail.ru*

# ON THE IMPROVEMENT OF TERMINOLOGY IN THE KAZAKH LANGUAGE

**Abstract.** In order for the language to be successfully applied in any field of knowledge or activity for the formulation of a theory, the designation of basic concepts and the relationship between them, appropriate terminology is necessary. According to the author of the article, the return of Kazakhs to the Latin graphic system allows not only to use the optimal variant of the alphabet, which implements the principle of matching one grapheme to one phoneme, but also to develop terminology that meets the needs of modern scientific and professional activities. The successful development of any language is ensured by increasing the vocabulary, replenishing vocabulary relevant to modern needs. For example, it is proposed to introduce into speech practice the use of the well-known word in a number of Turkic languages - the Arabic word äyile as a sociological term for the concept of family.

The literary language is characterized by normativity, thanks to which the language retains its integrity and comprehensibility for all speakers of it. Referring to the words-Arabisms, often represented in speech by significantly different phonetic variants, the author considers it expedient to define as a lexical norm the variant closest in sound to the original, which contributes to the awareness of the connection between the same-root borrowed words.

The grammatical structure of the Turkic languages allows you to borrow adjectives without complicating them with adjective affixes, but despite this, some words of European origin in the Kazakh language have an excessive morphemic structure. For this reason, the article provides an opinion on the use of such borrowings in the original form, following the example of other written Turkic languages.

**Keywords:** term, family, loan words, archaisms, adjectives.

Известно, что различные предметы и явления порождают определенные представления и понятия о них. Потребность оперирования этими понятиями вызывает необходимость их обозначения в языке, как важнейшем средстве человеческого общения. Одним из путей обогащения словарного фонда предстает полисемия, в

результате которой отдельные слова, отражающие наши представления о каких-либо проявлениях объективной реальности, становятся многозначными, называя новые предметы, явления и понятия на метафорической, метонимической или функциональной основе. Однако во многих случаях полисемия слов и наличие синонимов, отличающихся смысловыми и стилистическими оттенками, приходят в противоречие с требованиями строгой определенности понятий, особенно в разных отраслях науки, в юриспруденции, в сферах экономической и производственно-технической деятельности. Как известно, этим требованиям отвечает особая категория слов – *термины*, которые направлены на специальные понятия в разных областях знаний [1, с. 206].

Принятие в нашем государстве Закона «О языках в Республике Казахстан» явилось одним из стимулов деятельности по развитию и совершенствованию казахской терминологической лексики, особенно в сфере права и экономики. Здесь можно отметить, что традиционными источниками пополнения фонда специальных слов и составных наименований в казахском языке предстают исконная лексика, слова арабского происхождения и интернациональные термины. Например, устаревшее слово beren ‘сталь высокого качества’, встречавшееся в казахском фольклоре в словосочетаниях aq beren ‘доспех из металлических пластинок, скрепленных кольцами’, kök beren ‘ружье, ствол которого изготовлен из прочной стали’ (beren mıltıq), в настоящее время перешло в разряд используемой лексики, получив новое терминологическое значение ‘бронежилет’ [2, с. 49].

Учитывая значительное влияние военного искусства средневековых тюркских народов на организацию ратного дела восточнославянских племен (в результате которого последними заимствовались предметы вооружения, структура войск и тактика ведения боевых действий), можно утверждать, что наряду с многими словами-тюркизмами, обозначающими соответствующие реалии, слово beren было освоено русским языком в формах *бронь* в значении ‘кольчуга, ратные доспехи’ и *броня* ‘металлическая одежда, защищающая туловище воина’. В последствие это слово расширило свою семантику, используясь в значении ‘прочная защитная облицовка из специальных стальных плит на военных судах, танках и т.п.’

Однокоренные арабские слова كامل kamil ‘полный, законченный, совершенный’, كمال kämal ‘совершенство, законченность, полнота’ и производное от него كماليت kämaliyyät (с тем же значением) заимствованы казахским языком, соответственно, в виде kämil, kemel и kämelet. Последняя лексема приобрела со временем специальное

значение ‘совершеннолетие, зрелость’ и закрепилась в устойчивом словосочетании kämelettik attestat (аттестат зрелости – свидетельство об окончании средней школы); в форме дательного падежа сочетаясь с глаголами jetiw ‘достичь’ и tolıw ‘исполниться’, переводится как ‘достижение совершеннолетия’.

У казахов издавна существует ряд слов и словосочетаний, ядерной семой в которых является значение ‘семья’, понимаемое в прежние времена чаще всего как единица хозяйствования. В первую очередь, укажем на древнетюркское слово tütün, букв. ‘дым’. Очевидно, что в условиях кочевого скотоводства коллективные поселения и отдельные хозяйства не были стационарно связаны с конкретными земельными наделами, дворами, усадьбами; поэтому за счетную единицу принималось жилище, в которой проживала группа близких родственников.

Центральное место в жилом помещении занимал очаг, на котором разводили огонь для приготовления пищи и обогрева внутреннего пространства дома, кибитки. Поговорка *Нет дыма без огня* объясняет, почему дым как явление стал визуальным сигналом обитаемого жилья, а у слова tütün появилось новое значение. Оно было воспринято некоторыми земледельческими народами, переведено на их языки и усвоено вместе с системой исчисления подданных и налогового сбора, вероятно, еще со времен Великого переселения народов, господства гуннов в Средней Азии (эфталиты-хиониты и кидариты), в Восточной и Центральной Европе.

Возможно, в период существования в центре средневековой Европы могущественного Аварского каганата либо позже – во время появления на севере Балканского полуострова Болгарского ханства – данники-славяне заимствовали этот социальный термин в виде лексической кальки, т.е. буквально перевели его на свой язык, но с закреплением за ним его нового, переносного значения.

О порядке сбора правителями Хазарского каганата с подвластных восточнославянских племен полян, северов и вятичей дани “от дыма”, т.е. от каждого жилья упоминается в “Повести временных лет” [3, с. 7]. Вплоть до XIX века единица податного обложения в Польше называлась *дымом*, через понятие *очаг* прямо обозначая дом, крестьянскую усадьбу (семью) [4, с. 799]. Добавим, что в Имеретии (историческая область Грузии) со времен средневековья численность населения определялась также словом, имевшем прямое значение ‘дым’. Хроники сообщают о переселении на территорию Восточной Грузии в XII веке 40 000 “дымов”, или кибиток половцев-кипчаков.

Валиханов Ч., повествуя о нападении джунгар на кочевья киргиз-калмыков, употребляет слово *дым* в значении ‘семья’ [5, с. 285]. Персидское دودمان dud(e)man имеет значения ‘семья, семейство; род, племя’. Производящая основа этой лексемы – корень دود dud ‘дым’. В казахском языке слово tütin, имеющее прямое значение ‘дым’, на основе метонимии прежде использовалось в переносном смысле ‘семья’. Фразеологизм tütin tütetiw можно перевести как буквально – ‘зажечь очаг’ (букв. ‘задымить’), так и в переносном значении ‘жениться, создать семью, жить самостоятельно’. Выражение otı janbağan (tutanbağan), букв. ‘тот, у кого не горит (не зажжен) огонь’, характеризует человека, не имеющего своего очага, т.е. дома и семьи [6, с. 145]. Словосочетанием tütin salıq называли государственный сбор, налагаемый на каждое отдельное хозяйство (ärbir jeke üy bası’na salınatın salıq). Переносное значение слова tütin выражено в казахской рифмованной поговорке Yeki jartı – bir bütin, / Yerli-qatın – bir tütin ‘Две половинки – одно целое, / Муж с женой – одна семья’.

Развитие новых лексических значений произошло через ряд последовательных метонимических переносов, осуществляемых на основе постоянной связи во времени и пространстве двух явлений или предметов, в данном случае *группы жильцов* и *зажженного очага,* почитаемого святым местом жилища (oşaq); затем очага, где готовили пищу и огонь которого обогревал помещение в осенние холода и зимнюю стужу, и *очажного дыма* (tütin); впоследствии – дыма и отверстия-дымохода в своде, ограниченного размерами *круглого навершия* юрты (şangıraq); и, наконец, последнего звена в этой логической цепи – отверстия в куполе и его *войлочного покрытия* (tündik).

В результате смыслового развития, слова tütün и tündik (tündük) оказались синонимами во вторичных значениях, определяя семью как субъект хозяйственной деятельности. Отметим, что у киргизов один из старых видов налога назывался tündük-zäket, т.е. подымный сбор с каждой юрты (взимался с местного населения в пользу кокандских ханов) [7, с. 259]. Кроме того, в казахском языке встречалось изафетное словосочетание tündik bası (tütin tütetetin ärbir üy), которое по грамматической структуре и композиционной семантике аналогично выражению ot bası, букв. ‘глава, начало огня’ (современное значение у этой синтагмы – ‘семья’). Фразеологизм ot bası, oşaq qası, букв. ‘близ огня, возле очага’, отражает понятие *домашний очаг,* т.е. *семья* [6, с. 145]. Казахское слово oşaq имеет следующие лексические значения ‘очаг; печь; семья’.

Русское слово *очаг* было заимствовано из древнетюркского наречия (ср. *турецк.* ocak, *узбекск.* uçoq, *казахск.* oşaq) в прямом и переносном значениях: ‘устройство для разведения и поддержания огня, печь; родной дом, семья’ (*Вернуться к родному очагу*). Второе производное значение лексемы *очаг* ‘место, откуда что-нибудь распространяется’, видимо, появилось уже в русском языке.

Понятие *семья* как элементарный социально-экономический комплекс (люди, хозяйство, земельное владение) в казахском языке выражается такими парными словами, как mal-jan, букв. ‘скот, имущество + души’, üy-jay, букв. ‘дом + место, местожительство’. Кроме того, отделение взрослых женатых сыновей от отцовского очага и создание новых семей обозначалось словами üy bolıw, букв. ‘стать домом’, в результате чего у лексемы üy ‘дом’ появилось переносное значение ‘семья’. Производный от тюркского слова ev ~ üy ‘дом’ глагол evlen- ~ üylen-, букв. ‘обзавестись домом’, приобрел значение ‘жениться’, а противопоставленные ему по признаку *активный – пассивный залог* глаголы *evger > ever- и üylendir- – значение ‘женить’. В Турции, где издавна развито земледелие, слова ev и hane, имеющие прямое значение ‘дом, помещение’, в переносном значении используются для обозначения двора, отдельного крестьянского хозяйства и служат единицей счета [8, с. 67].

Часто в речи казахов для обозначения только домочадцев, т.е. жены, детей и иных проживающих вместе близких родственников, применяется словосочетание üy işi, которое имеет следующую семантику: ‘внутренняя обстановка (в противоположность üy sırtı ‘внешнее пространство дома’); семья’. В родственном узбекском языке одним из обозначений семьи также является устойчивое сочетание слов üy içi ‘дом, семья’. Здесь следует добавить, что для перевода слова *семья* с русского на туркменский и узбекский языки, помимо других слов, используется лексический дуплет bala-çaga ~ bolä-çaqa, букв. ‘дети’, где обе части сочетания синонимичны (ср. казахск. balalı-şağalı ‘семейный’).

Новая семейная ячейка, получившая свой надел (скот, угодья, постройки, иное имущество), называлась otaw ‘юрта, дом для молодоженов; семья молодых’ [9, с. 271]. Рассматриваемое слово в прямом значении обозначало небольшую свадебную юрту, которая обычно изготовлялась в приданное дочери (aq otaw). Русским эквивалентом устойчивого словосочетания otaw qur- предстает выражение *создать семью*. Следует отметить, что во многих тюркских языках аналогичным словом обозначали как разновидность жилья, так и отдельное помещение в нем.

Валиханов Ч. в своей статье “Очерки Джунгарии” сообщает, что в жаргоне купцов походная палатка именовалась словом *огонь*: “К 25 числу сентября на этом месте собралось до 60 палаток, или, как принято в караванном языке, до 60 огней… Соединенный наш караван состоял из десяти огней и число людей увеличилось до 60” [5, с. 291]. Основываясь на данном сообщении, а также на наличии в лексике казахского языка глагола otas- ‘жить совместно (о супругах и их родных, букв. ‘вместе зажигать огонь’)’, можно предположить о том, что слово otaw ‘жилище’ образовано от глагола ota- ‘разводить огонь’, в свою очередь, производного от существительного ot ‘огонь’.

Киргизы словом otaw называют облегченный тип юрты, балкарское otaw – пристроенная к жилью дополнительная комната для молодоженов, азербайджанское otag и турецкое oda переводится уже просто словом *комната* (в Западной Грузии заимствованным словом oda называют дом, жилую часть хлева). Отмеченные изменения в семантике характерны и для общего родового слова üy ‘дом’, одним из значений которого является ‘отдельное помещение в жилище’ (awız üy – передняя, tör üy – гостиная, demalıs üy – спальная, qorjın üy или qarsı üy – противоположные комнаты) [9, с. 132].

Изменения социально-экономических условий обусловили потребность в слове, обозначающем семью как новую социальную формацию, которое бы имело определенное, отвечающее юридическим отношениям, содержание – ‘группа живущих вместе родственников (муж и жена, родители и дети)’ и было свободно от дополнительных смысловых нюансов.

Если члены семьи проживают в многоквартирном доме, включены в совершенно иную, отличную от прежних времен, правовую, в т.ч. налоговую, систему, работают в разных местах, не имеют единого крестьянского хозяйства либо частного предприятия, то для обозначения подобной семьи, имеющей определенные законом имущественные права и обязательства, такие слова, как tütin, tündik, ot bası, oşaq, mal-jan, üy-jay не вполне приемлемы. У всех вышеприведенных слов, используемых в казахском языке для обозначения понятия ‘семья’, семантическая структура сохраняет излишние дополнительные значения, смысловые и стилистические коннотации. Кроме того, переносные значения лексем tütin и tündik оказались архаичными, так как семантическое поле (логическая связь обозначенных предметов, явлений и понятий, тематическая группировка соответствующих слов) и исторический фон, связанный со временем возникновения переносного значения, существенно изменились.

Следовательно, у многих людей возникает потребность в использовании слова-термина, имеющего строго определенную семантику ‘совокупность близких родственников, проживающих вместе и ведущих совместное домашнее хозяйство’. Иными словами, современные условия жизни в нашей республике диктуют необходимость употребления в официальной сфере социологического термина, обозначающего понятие ‘сообщество, основанное на браке супругов, помимо которых включает их холостых детей, связанных духовно, общностью быта, взаимной моральной ответственностью, имущественными и правовыми отношениями’.

Ответом на указанную потребность было появление в нашем языке лексического двухкомпонентного образования jan + uya, а также употребление в речи русского слова *семья*. Оба названия нельзя признать литературной нормой по следующим причинам.

Первое выражение januya представляет собой метафорическое выражение (буквально ‘гнездо души’) и, по сути, подразумевает место проживания людей; слово jan ‘душа’ здесь используется в значении ‘человек’, ср. русское выражение *Не видно ни души*. Употребление этого неологизма более оправдано в художественном стиле речи, допускающем использование поэтических образов, нежели, например, в строго детерминированном тексте Кодекса о браке и семье.

Встречающееся не только в устной речи, но и зафиксированное в некоторых казахских словарях слово *семья* в сложившихся условиях развития и взаимоотношения культур может рассматриваться как факт неправомерного смешивания двух языков и относится, на наш взгляд, к разряду варваризмов.

Случается так, что казахи, недостаточно хорошо владеющие родным языком, в процессе общения вносят в свою речь чуждые элементы и даже целые лексико-грамматические конструкции (проявление лингвистической интерференции). Через их смешанную речь подобные нарушения языковых норм становятся распространенными. Например, вместо общетюркской модели пожелания **…** (quttı, mübaräk, hayırlı, qabul) bolsın! все чаще встречаются русифицированные грамматические построения с постоянным компонентом **…**-men (-ben) quttıqta-.

Взамен традиционного новогоднего поздравления Jaña jıl(ıñız) quttı bolsın! в наше время обычно говорят и пишут: Janga jılıñız’ben quttıqtaymız! или даже Janga jılıñız’ben! ~ Meyramlarıñız’ben ~ Merekeleriñiz’ben!, что представляет собой грамматическую кальку русских выражений *С Новым годом* (*поздравляем*)! ~ *С праздником*!, в

которых имена нарицательные оформлены окончаниями творительного падежа.

По шариатским нормам поведения и речевому этикету чихнувший мусульманин должен обратиться к Всевышнему и восхвалить его за милость словами الحمدالله Alhamdulillah ‘Хвала Аллаху’. Услышав это, другой человек должен сказать ему يرحموكالله Yarhamukallah ‘Да помилует тебя Аллах’ (просторечный вариант: Jarahmalla). Иногда произносят доброе пожелание чихнувшему по-казахски: Ber täñir (bes jüz jılqı) ‘Дай Боже (пятьсот лошадей)’. Однако часто в своей речи представители молодежи в таких случаях ошибочно используют фразу Saw bol, букв. ‘Будь здоров’, несмотря на то, что она традиционно произносится в качестве пожелания при расставании и соответствует русскому ‘Бывай здоров, до свидания’.

Приведенные факты относятся к категории варваризмов, т.е. к оборотам речи, построенным по чуждому образцу и нарушающим чистоту речи носителей казахского языка.

То или иное слово, представляющее собой определенный звуковой комплекс и рассматриваемое в фонетическом аспекте, может оцениваться формально, с учетом количества и качества фонем, т.е. размера слова и его благозвучия (эвфонии), с учетом артикуляционной сложности, т.е. соответствия звуков слова фонологии родного языка. Но в этом плане нельзя делать выводы о целесообразности использования какого-либо слова, будь то исконное или заимствованное. Однако на лексико-семантическом уровне, при определенных историко-культурных условиях положение меняется.

Дело в том, что в современных условиях развития национальной культуры и роста этнического самосознания, некоторые лексемы, в некоторых случаях – фонографические варианты слов (при их наличии), могут восприниматься как более или менее предпочтительные либо отвергаться. Например, трудно оправдать существование в казахском языке таких словосочетаний, как *балалар творчествосы* ‘творчество детей’ или *творчестволық бірлестігі* ‘творческое объединение’. Создатели этих лексических композит, видимо, не утруждали себя поиском эквивалента русского слова *творчество*.

Обратившись к лексическому фонду и лингвистическому опыту иных тюркских народов, находим, например, в турецком языке обозначение понятия *творчества* (как деятельности) словами yaratma kuvveti, yaratıcılık, а русский глагол *творить* переводится словами yaratmak, vücud’e (форма дат. падежа) getirmek. Укажем на наличие в казахском языке слова jaratıw, имеющего значение ‘сотворить, создать что-нибудь’, которое генетически связано с турецким глаголом

yaratmak. От арабского заимствования ايجاد idjad ‘создание, творение’ образовано узбекское глагольно-именное сочетание ījod qılış ‘творение (как процесс)’. Производное слово ījodiyot < ايجادية idjadiyat и синонимичное ему собственно-тюркское yaratış < yarat- ‘создавать’ со значением ‘творчество’ имеются в лексическом фонде узбекского языка.

Изредко в казахской литературе встречаются указанные арабизмы: лексема ījad в значении ‘творение; издание’ и однокоренное ей слово ujud (ср. турецк. vücud) ‘существование, бытие; наличие, присутствие’ [10, с. 208]. Следовательно, есть в языке собственные лексические возможности для обозначения понятия *творчество*, а также иных понятий, явлений и предметов, и надо их использовать, не создавая серию неудачных словесных гибридов.

Вслед за кириллическим алфавитом, который был введен в конце 30-х годов прошлого века решением политического руководства страны вместо латиницы, в казахский язык пришли слова, среди которых были не только те, что обозначали новые для народа понятия. С определенного времени стали “правильно” употреблять и писать слова: *Москва* вместо используемой исстари адаптированной формы *Мәскеу* (Mäskew), *Россия* – вместо названия *Ресей* (Resey), видимо, заимствованного в прежние времена у яицких казаков, говоривших *Расея*. Появилось слово *ағылшын,* которое оказалось существенным искажением и при этом выглядит гибридом старорусского *аглицкий* и нового *англичане.* Форма *ағылшын* сосуществует в казахском языке с однокоренным словом – хоронимом *Англия*.

Слово *кеңес* в словосочетании *Кеңес республикасы* стало заменяться на заимствованное *совет*, см. также синтагму *отряд советі* ‘совет отряда’, *Совет союзының батыры Мәншүк Мәметова* (в тексте фронтовой листовки). Возникли словосочетания *қазақ искусствосы* ‘казахское искусство’ (например, в номинативном предложении *Қазақ искусствосы мен әдебиетінің Москвадағы декадасы* – названии декады национальной культуры, проводившейся в 1958 году), *халық творчествосы* ‘народное творчество’, *тең праволы* ‘равноправный’, *тең праволылық* ‘равноправие’. К большому сожалению, подобные “французско-нижегородские” изобретения выхолащивают родной язык, внушают мысль об отсутствии в нем внутренних возможностей для обозначения понятий, не связанных с какими-то совершенно новыми идеями, явлениями или предметами.

Аналогичные явления наблюдались в языках других народов, проживающих на территории бывшего СССР, о чем свидетельствуют следующие примеры: алтайск. *марксистский философский*

*материализм* ‘марксистский философский материализм’, бурятск. *советскэ общественнэ байгуллат* ‘советский общественный строй’, удмуртск. *советской власть* ‘советская власть’, *советской калык* ‘советский народ’, хакасск. *социал-демократическай* ‘социал-демократический’, чеченск. *революционно-демократически* ‘революционно-демократический’ и т.д.

Негативная тенденция эрозии лексики, наблюдавшаяся в советские времена как в общественно-политической терминологии, так и в географических названиях, с возникновением суверенного государства была преодолена, и в современной казахской печати (например, в газете “Zaman-Қазақстан”) стали появляться исконные слова, прежние названия, сознательно “забытые” в предшествующую эпоху. Например, самоназвание венгров magyar издавна в тюркских языках имела звучание majar (под этим именем было известно одно из средневековых тюркских племен), соответственно Венгрия называлась Majaristan (турецк. Macaristan), Грузия была известна как Gürjistan (турецк. Gürcistan; русское слово *грузин* произошло от тюркского gürji, пройдя стадию в виде старорусской формы *гурзи*), столица Грузии Тбилиси по-казахски адаптируется в форме Tiblis (турецк. Tiflis).

На мусульманском Востоке западных европейцев называли fereñg (от имени германского племени франков), выделяя среди них англичан – iñgiliz (от *английск*. English). Поэтому в мусульманских странах Западная Европа в старину называлась فرانكستان Ferengistan ‘Страна франков’, современная Англия – انگلستان Ingilistan. Учитывая сказанное, правильным будет введение в лексику казахского языка слова iñgiliz вместо употребляемого ныне *ағылшын*.

Слово *Алмания* в современных публикациях на казахском языке соответствует русскому названию страны Германия; так она называется в турецком языке – Almanya от французского Allemagne ‘Германия’, в свою очередь ведущего происхождение от названия германского племени аллеманов; в языке казахов, проживающих в Турции, *алман* ~ *албан* ‘немец’ (от турецкого alman ‘немец’).

Законы фонетики тюркских языков не допускают сочетания в слове трех согласных звуков. Однако вместе с принятием кириллицы появилось новое написание названия нашей республики в форме *Қазақстан* (в котором оказались соединенными согласные звуки -qst-), хотя в период использования латинизированного алфавита писали Qazaqʰstan в соответствии с народным произношением, т.е. без нарушения орфоэпической нормы. Более того, скопление четырех согласных [-rkst-] допущено в топониме *Түркстан*, встречающемся в

монографии А. Абдрахманова “Географические названия Казахстана” [11, с. 23, 133].

Этимологически верная графическая передача названия нашей республики – Qazaqıstan: слова qazaq и stan объединены изафетной связью по правилам грамматики персидского языка (ср. турецк. Özbekistan, Kazakistan, Kırğızistan, Moğolistan, Tacikistan, Türkmenistan и т.д.), следовательно, записанная кириллицей синтагма *Егемен Қазақстан* должна передаваться в виде Yegemen Qazaqıstan.

Если обратиться к тюркским языкам, имеющим развитую письменность и богатый лексический фонд, то можно отметить логический подход к формированию терминологии, учитывающий многовековую традицию усвоения через религиозную, научную и художественную литературу и устную речь слов арабского происхождения.

На примере некоторых слов отметим тенденцию, определившуюся в современном казахском языке. Вместо ранее широко употреблявшихся (но в тоже время явно осознаваемых как инородные) слов *директор, текст, процент*, появились известные в ряде других тюркских языков новые лексемы müdir, mätin, payız, которые воспринимаются в речи органично, соответствуя фонетике (звуковому строю) языка.

Будучи общими для многих тюркоязычных этносов, слова-арабизмы и фарсизмы, обозначающие социальные и природные явления, абстрактные понятия и предметы материальной культуры, подобно исконно-тюркской лексике, являются одним из факторов, объединяющих представителей этих народов и облегчающих их взаимопонимание.

Для номинации понятия *семья* в ряде тюркских языках наряду с другими синонимами активно применяется арабское слово عايله ‘ayilä (например, татарск. ğayilä, узбекск. oilä, турецк. aile). В персидском языке указанная лексема в фонетических вариантах ayele ~ aele имеет значения ‘семья, семейство; жена, супруга’. Казахи широко используют однокоренное слово äyel в значении ‘женщина; жена’, заменив исконное qatın, которое характеризуется стилистически сниженной окраской, получив в речи оттенок вульгарности. В фарси ему соответствуют варианты äyal ~ eyal, в узбекском – äyöl’ ‘женщина’.

Использование в казахском литературном языке для обозначения понятия ‘семья’ – с учетом современного статуса этой общественной формации – арабского слова عايله ‘ayilä в адаптированной форме äyile не противоречит культурно-исторической традиции; оно будет лишь способствовать обогащению словарного состава, сохранению близости

литературных языков тюркских народов. Это слово свободно от излишних смысловых нюансов, имеет строго ограниченное значение, т.е. обладает признаками термина, генетически и семантически соотносится с ранее освоенным словом äyel и может использоваться в текстах юридического, научно-философского характера, общественно-политической направленности, в деловых и иных документах.

Употребление таких слов, как üy, üy işi, üy-jay, tütin, tündik bası, ot bası, mal-jan, jan uya, otaw, oşaq и *семья* в официально-деловом стиле, в научно-публицистической и иных сферах, требующих строго определенного значения ‘совокупность живущих вместе близких родственников (супруги, родители с детьми)’, представляется неприемлемым ввиду приведенных выше доводов.

Важным представляется то, что в процессе возврата к латинизированному алфавиту можно успешно решать проблемы орфографии, опираясь на морфологический принцип. Иными словами, из всех вариантов написания слов (особенно заимствованных из других языков), отражающих разное произношение и встречающихся ныне в письменной речи, выбрать те фонографические виды, которые соответствуют принципу единообразного написания морфем.

Например, литературными окажутся такие однокоренные слова, как *рахмет* – rahmet ‘спасибо, благодарю’, *рахыйм* – rahıym ‘милость; милосердие’, *мархұм* – marhum ‘покойный, умерший’, *мархамат* – marhamat ‘милость, милосердие; пожалуйста; добро пожаловать!’ (корень слов-арабизмов رحم r-h-m), а их варианты *рақмет*, *рақым*, *марққм*, *марқабат* – просторечными формами слова, так как их гомогенность и семантическая связь завуалирована.

В последнее время вместо используемого ранее и отраженного в словарях слова *хұқық* ‘право’ (корень حقّ ~ حقق h-q-q) получило преобладание в официальных документах его просторечное видоизменение *құқық*, утратившее всякую фонетическую близость с такими однокоренными словами, как *хақ* ‘правда, истина; право’, *хақиқат* ‘истина, правда; истинный, правый’ и, к тому же, менее благозвучное. Из вариантов *харекет ~ қарекет ~ әрекет* ‘действие, деяние; поступок’ (арабск. حركت harakat) правомерно выбрать первый, как более близкий к арабскому оригиналу.

Наличие в литературном языке просторечных вариантов *ақпар* ‘сводка; итоговые данные’ и *ақпарат* ‘информация’ не позволяет распознать в них формы множественного числа от лексемы *хабар* ‘весть, известие, сообщение; осведомленность’ (корень خبر h-b-r), поэтому в целях отражения их корневого единства следует писать *ахбар* и *ахбарат*.

Вместо вариантов *мубарак*, *мұбарак*, *мүбәрак*, *мүбәрәк* и даже *мұбарақ* слово-арабизм следует писать в форме *мүбарәк* ~ mübaräk, соответствующей исходной лексеме مبارك mubārak ‘благословенный; благополучный; счастливый’.

Термин تَسْبِيح tasbīḥ, используемый в религиозной сфере общественной коммуникации для обозначения формул прославления Бога в исламе, служит также названием для четок, перебирая которые повторяют по 33 раза соответствующие выражения. Этот арабизм насчитывает в казахском языке 25 фонетических вариантов, но в литературном языке должно использоваться в качестве нормативного написание, наиболее близкое к оригиналу, т.е. в виде *тәсбих* ~ täsbiyh.

Также правомерным представляется возвращение в активный словарь казахского языка терминов-архаизмов, связанных с административно-территориальным делением страны. Например, вместо *облыс* следует использовать слово wälayat (ولاية) ‘область’, *округ* – dayire (دايره) ‘округ’ и т.д.

Приобретение нашей республикой государственного суверенитета, широко распространившаяся практика использования электронно-вычислительной техники во многих сферах жизнедеятельности человека, возврат к национальному варианту латинской графики, представляющему собой систему из 34 символов в соответствии с принципом “одна буква – одна фонема”, – все эти обстоятельства позволяют вывести наш родной язык из прежнего состояния стагнации. И важным направлением в творческом процессе совершенствования казахского языка предстает разработка научной терминологии в разных областях знаний и сфер человеческой деятельности (пополнение словарного фонда философской, общественно-политической и научно-технической лексикой). Особое значение приобретает концепция совершенствования компьютерной терминологии в казахском языке.

Стихийно сложившаяся во многих языках мира практика графической передачи названий электронных программных продуктов, компьютерных технологий, так называемой “компьютерной терминологии” латинскими буквами свидетельствует о значительном влиянии английского языка, роль которого приоритетна в международной коммуникации научного, экономического, политического, социального и культурного характера – многие термины, имена собственные и аббревиатуры, связанные с информатикой и компьютерными технологиями, пишутся по-английски и понимаются верно.

Использование казахами национального варианта латиницы позволяет сохранять на письме идентичную или близкую к ней форму иноязычных имен собственных, если в оригинале они передаются с

помощью латинских букв. Иными словами, следует избегать применения метода транскрипции. Так, например, в целях адекватного восприятия элементов текста название электронного приложения-мессенджера WatsApp нецелесообразно транслитерировать в виде ýatsap, название многоязычной универсальной интернет-энциклопедии Wikipedia – как ýıkıpedıa и т.д.

Касаясь проблемы совершенствования терминологии в казахском языке, следует обратить внимание на важную особенность: грамматический строй тюркских языков позволяет заимствовать прилагательные, не осложняя их аффиксами прилагательных и окончаниями, как это происходит, например, в русском языке: actual > *актуальный*, academic > *академический*, radioactive > *радиоактивный*. Приведем примеры слов, относящихся к разряду интернационализмов, усвоенных многими тюркскими языками: азербайджанск. alternativ ‘альтернатива; альтернативный’, analitik ‘аналитик; аналитический’, mineral ‘минерал; минеральный’, program ‘программа; программный’, profilaktik ‘профилактический’, башкирск. monopolistik ‘монополистический’, турецк. aktif ‘актив; активный, действенный’, akustik ‘акустика; акустический’, metodik ‘методический’, komik ‘смешной’, paradoksal ‘парадоксальный’, татарск. absolüt ‘абсолютный’, logīk ‘логический’, pedagogīk ‘педагогический’, stomatologīk ‘стоматологический’, normal′ ‘нормальный’, узбекск. agressīv ‘агрессивный’, operatīv ‘оперативный’, revolutsion ‘революционный’, fīzīk ‘физический’, klassīk ‘классический’, туркменск. abstrakt ‘абстрактный’, aktual ‘актуальный’, koalicion ‘коалиционный’, чувашск. venerik ‘венерический’ и т.д.

Вследствие этой особенности органично вошли в казахскую лексику следующие слова: *актив* ‘активный’, *натурал* ‘натуральный’, *поляр* ‘полярный’, *пропорционал* ‘пропорциональный’, *радикал* ‘радикальный’.

Но вместе с этими лексемами, к большому сожалению, в нашем языке появились заимствования, оформленные явно избыточными аффиксами прилагательных:

*актуальды* ‘актуальный’,
*астралдық* ‘астральный’,
*классикалық* ‘классический’,
*лунарлық* ‘лунарный’,
*минералды* ‘минеральный’,
*модальдық* ‘модальный’,
*радиоактивті ~ радиоактивтік* ‘радиоактивный’,
*революциялық* ‘революционный’,
*солярлық* ‘солярный’,

*химиялық* ‘химический’,

*физикалық* ‘физический’ – в этом случае мы слепо следуем примеру русского языка, несмотря на типологическое различие тюркских и славянских языков.

Нельзя не воспользоваться грамматическим потенциалом тюркских языков в отношении усвоенных слов-интернационализмов. Например, словосочетания *перинаталдық орталық* ‘перинатальный центр’ следует писать на латинице в виде perinatal ortalıq, *патрульдік полиция* ‘патрульная полиция’ – patrol policia, *командалық бекет* ‘командный пост’ – komanda beketi (postı).

Прилагательные, имеющие германское или романское происхождение, предпочтительно усваивать в форме, близкой к оригиналу, в результате чего они будут краткими при записи, более легкими в произношении и не противоречащими правилам тюркской грамматики, например: *американдық* – ämerikän, *индиялық* – indiyän, *евразиялық* – yevräziyän.

Мы должны отказаться от просторечных слов-искажений *кәріс* (кореец), *неміс* (немец), *үндіс* (индеец), заменив их вариантами köreän, alman, indiyän; вместо прилагательного *ядролық* ‘ядерный’ использовать слово *nükleär* (ср. турецк. *nükleer*).

Необоснованным представляется перевод прилагательного *оперативный* на казахский язык словами *жедел* и *қауырт*, передающим соответственно значения ‘быстрый, скорый’ и ‘спешный; авральный’ (см. примеры: *оперативная работа* – жедел жұмыс; *оперативная группа* – жедел топ; *оперативное реагирование* – жедел жауап қату; жедел жауап қайтару.

Лексема *оперативный* – оформленное по правилам русского языка слово-интернационализм, образованное от английского operative ‘действующий’, восходящего, в свою очередь, к латинскому существительному operatio ‘действие’ (исходное слово opus ‘работа’).

В русских толковых словарях приведены следующие значения для рассматриваемого слова:

1. Связанный с хирургическим вмешательством, операцией.

2. Относящийся к военным операциям, действиям силовых структур.

3. Непосредственно, практически осуществляющий что-либо.

4. Предназначенный для выполнения операций.

5. Действующий быстро; способный быстро, вовремя исправить или направить ход дела.

Для передачи всего семантического спектра рассматриваемого прилагательного в казахском языке стали использовать слово-кальку

jedel, букв. ‘скорый, быстрый; срочный, экстренный’, связанное лишь с его пятым, переносным значением.

По примеру других тюркских языков, правомерно введение в лексику казахского языка слов-синонимов operativ (вместо *оперативті*, *оперативтік*), hareket -(s)ı (например, hareket tapsırmaları ‘оперативные задачи’) и amaliy < amal ‘дело; действие’ (ср. турецк. operatif , amelî ‘оперативный’, ameliyat ‘операция хирургическая’, harekât, operasyon ‘военная операция’, amelye, muamele ‘финансовая операция’).

Приведем характерные факты: в одном из телеинтервью женщина, по всей видимости, не имеющая филологического образования, но интуитивно осознавшая суть родного языка, в своей речи использовала прилагательные moral, materiyal, чувствуя неоправданную сложность литературных форм *моральдық*, *материалдық*; в другом интервью на казахском языке прозвучало прилагательное klassik вместо нормативного *классикалық*.

Аффиксы -lı ~ -lıq (с вариантами) образуют прилагательные от именных основ (say ‘овраг’ > saylı ‘овражистый’, kir ‘грязь’ > kirli ‘грязный’, muz ‘лед’ > muzdı ‘ледяной’, yen ‘ширина’ > yendi ‘широкий’, dañq ‘слава’ > dañqtı ‘знатный’, mädeniyet ‘культура’ > mädeniyetti ‘культурный’ и т.д.), поэтому нелогично и даже абсурдно выглядит стремление присоединять к заимствованным прилагательным аффиксы, которые образуют слова, относящиеся к этой же части речи.

Необходимо использовать возможности, возникшие при переходе казахского языка на латинскую графику, и интернациональные лексические заимствования писать в виде äktüäl, modal, radioäktiv, revolücion, geömetrik, hiymik, fiyzik, klassik (во всех примерах ударение падает на последний слог) и т.д., не осложняя их служебными морфемами, несущими излишнюю информацию. Кроме того, препозиция таких слов по отношению существительным ясно указывает на их главную функцию определения, характерную в первую очередь для прилагательных.

**Список литературы**

1. Кодухов В.И. Введение в языкознание. – М.: Просвещение, 1979.
2. Словарь русского языка / Сост. С.И. Ожегов. – М.: ГИИНС, 1950.
3. Повесть временных лет: Хрестоматия по древней русской литературе. – М.: Просвещение, 1973.
4. Жеромский С. Пепел. Роман-хроника к. XVIII – н. XIX вв. – М.: Худ. лит-ра, 1967.
5. Валиханов Ч. Избранные произведения. – М.: Наука, Гл. ред. вост. лит-ры, 1986.

6. Казахско-русский фразеологический словарь / Кожахметов Х.К. и др. – Алма-Ата: Мектеп, 1988.

7. Советская историческая энциклопедия. – Том 7. – М.: Советская энциклопедия, 1965.

8. Русско-турецкий словарь / Сост. Д.А. Магазанник, М.С. Михайлов. – М.: ОГИЗ, 1946.

9. Казахи. Историко-этнографическое исследование. – Алматы: Казахстан, 1995.

10. Арабша-қазақша түсіндірме сөздік / Н.Д. Оңдасынов. – Алматы: Мектеп, том 1 – 1984; том 2 – 1989.

11. Әбдірахманов А. Қазақстан жер-су аттары. – Алматы: Қазақ ССР Ғылым академиясының баспасы, 1959.

*ӘОК 81-13*
***Қожахмет А. Қ.***
*Л.Н.Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*a_kozhakhmet@mail.ru*

## ҒЫЛЫМИ МӘТІНДІ ОҚЫТУДЫҢ ТАНЫМДЫҚ СИПАТЫ

**Аңдатпа.** Мақалада ғылыми мәтінді оқытудың танымдық сипаты жайында зерттеулер жүргізілді. Бұл – мемлекеттік деңгейде маңызы зор мәселе. Сол себепті бүгінде білім алушылар арасында ғылыми мәтінді оқыту басты назарға алынып жүр. Әсіресе, ғылыми мәтіннің танымдық ерекшелігін ескере отырып оқыту тиімді болып саналады. Кез келген ойдың түпкі тамыры танымнан бастау алады. Таным кез келген ғылыми мәтіннің ұлттық ерекшелігін танытады. Жұмыста ғылыми мәтінді оқытудың танымдық сипаты ғылыми терминдерді оқытумен байланыста қарастырылды. Себебі ғылыми мәтіннің ең негізгі белгісі терминдердің қолданысы. Тақырыпты ашу мақсатында ғылыми мәтінді оқытуға қатысты ғалымдардың оқулықтары негізге алынды.

**Түйін сөздер:** ғылыми мәтін, ғылыми мәтінді оқыту, терминология, таным, ғылым тілі.

*УДК 81-13*
***Кожахмет А. К.***
*Евразийский национальный университет имени Л.Н. Гумилева*
*Нур-Султан, Казахстан*
*a_kozhakhmet@mail.ru*

## ПОЗНАВАТЕЛЬНЫЙ ХАРАКТЕР ОБУЧЕНИЯ НАУЧНОМУ ТЕКСТУ

**Аннотация:** В статье исследуется познавательная природа обучения научному тексту. Это важный вопрос на государственном уровне. Именно поэтому сегодня основное внимание уделяется обучению студентов научным текстам. Преподавание научного текста особенно эффективно с учетом познавательных особенностей. В основе любой мысли лежит познание. Познание отражает национальное своеобразие любого научного текста. Познавательный характер обучения научному тексту рассматривается в связи с обучением научных терминов. Потому что главной особенностью научного текста

является использование терминов. Для раскрытия темы использовались учебники ученых по преподаванию научных текстов.

**Ключевые слова:** научный текст, обучение научному тексту, терминология, познание, язык науки.

*UDC 81-13*
***Kozhakhmet A.***
*L.N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*a_kozhakhmet@mail.ru*

## COGNITIVE NATURE OF TEACHING A SCIENTIFIC TEXT

**Abstract:** The article explores the cognitive nature of teaching a scientific text. This is an important issue at the state level. That is why today the main attention is paid to teaching students about scientific texts. Teaching a scientific text is especially effective in view of cognitive features. At the heart of any thought is knowledge. Cognition reflects the national identity of any scientific text. The cognitive nature of teaching a scientific text is considered in connection with the teaching of scientific terms. Because the main feature of a scientific text is the use of terms. To reveal the topic, textbooks of scientists on teaching scientific texts were used.

**Keywords:** scientific text, teaching scientific text, terminology, cognition, language of science.

Ғылыми мәтінді ұлттық таным негізінде оқыту – бүгінгі күннің маңызды мәселелерінің бірі. Қазақстан Республикасының Президенті Қ.Тоқаев «Халық бірлігі және жүйелі реформалар – ел өркендеуінің берік негізі» атты Жолдауында білім беру саласына, әсіресе ғылымды дамытуға ерекше екпін түсіреді. Ғылымды ұлтпен сабақтастықта дамыту ғылыми мәтінді сапалы жазу мен оқытуға келіп саяды.

Ғылыми мәтінді оқыту қазақ ғылым тілін дамыту үшін маңызды. Ғылым тілінің дамуы, әлбетте, терминдердің жүйеленуімен байланысты. Термин түзуде негізге алынатын бірден-бір ұғым – таным. Себебі кез келген ойдың түпкі тамыры танымнан бастау алады. Танымның негізінде уәж, ассоциация, болжам сынды ұғымдар жатыр. Термин түзуде де осы ұғымдар қолданылады. Кез келген тілдік бірлік танымдық, психологиялық, философиялық саралаудан өтіп барып ойға айналады. Бұл үдеріс термин түзуде де жасалады.

Қазақ терминдері қазақтың тілдік бірліктерімен түзілгені қашанда абзал. Ғалым Ш.Құрманбайұлының «Қазақ тілінің терминологиялық

қорының қалыптасуына негіз болған басты көз ұлт тілінің өз байлығы екендігі сөзсіз. Оның бастауларын әріден, түркі тілдерінен, төл тіліміздің тармақталып дербес даму тарихымен қараған жөн» [Құрманбайұлы, 2014, 303 б.] деген пікірі ойымызға дәлел болмақ. Қазақ терминологиясын ұлттық таным негізінде қалыптастыру қағидатын алға қойған А.Байтұрсынұлының қазақ тіл білімі мен әдебиеттану салаларына қатысты терминдері бүгінде қазақ тіл білімі мен әдебиетануының ғылыми аппаратына айналып отыр. *Біріншіден,* ғалымның әр термині қазақ халқының болмыс-бітіміне, мәдениетіне, ұлттық танымына сай құрылған болса, *екіншіден,* ұлттық тіліміздің ғылым аренасындағы орнын танытуда жылдар бойы қызмет еткен терминдер болып саналады. А.Байтұрсынұлының «Тіл-құрал» еңбегіндегі *сөйлем, салалас сөйлем, сабақтас сөйлем, құрмалас сөйлем, дыбыс, дауысты дыбыстар, зат есім, сын есім, сан есім, есімдік, етістік, үстеу, қосымша, жалғау, жұрнақ, бастауыш, баяндауыш, толықтауыш, анықтауыш, пысықтауыш,* т.с.с. терминдері ұлттық таным негізінде жасалған терминдердің озық үлгісі болып саналады. Когнитивті лингвистика саласын зерттеуші ғалым Э.Оразалиева А.Байтұрсынұлының зерттеулеріндегі тілдік бірліктерді танымдық аяда қарастыра отырып, мынадай пікір айтқан: «Сөз басын «Аңдату» деп бастап, «...анық көріп, сезіп, біліп тұрған айналамыздағы заттардың бәріне» зер салған ғалым тіл мен танымның тығыз байланысқан арақатынасын өз тұжырымдамасының өзегіне айналдырды. Бұл ретте тілші «адам санасының» үш негізін барлық тілтанымдық фактілерге арқау етті, сөйтіп, тілдің міндетін де осы үш арнадан өрбітті. Олар – ақыл, қиял, көңіл. «Ақыл ісі – аңдау, яғни нәрселердің жайын ұғу, тану, ақылға салып райлау, қиял ісі – меңзеу, яғни ойдағы нәрселерді белгілі нәрселердің тұрпатына, бернесіне ұқсату, бернелеу, суреттеп ойлау, көңіл ісі – түю, талғау» дей отырып, ғалым бүгінгі танымдық лингвистиканың мақсат-міндетін, негізгі ұғымдарын өзектеді» [Оразалиева, 2007, 29 б.]. Сөз саптауда ұстанған осындай қағидаттар негізінде жасалған А.Байтұрсынұлы терминдері қазақ тіл білімі, әдебиеттану салаларын ғылыми тұрғыда қалыптастырды әрі осы салаларды бүгінгі тұғырына жеткізді. Қай елдің термині болмасын, ол сол елдің тілдік ерекшеліктерінен, мәдениетінен, ділінен, танымынан хабар беруі тиіс. Себебі термин ғылым тілінің бір бөлшегі ғана емес, ол – ұлтты танытушы құралдардың бірі. Осы қағидат негізінде жазылған С.Қожанұлының «Есептану құралы», Ж.Күдериннің «Өсімдіктану», Е.Омарұлының «Пішіндеме», Ж.Аймауытұлының «Психология», М.Жұмабаевтың «Педагогика» оқулықтары бүгінде ғылыми мәтінді оқытудың негізі болып табылады. Бұл еңбектерде кездесетін *жақша,*

*өлшеуіш, айырым, бөлім, сабақша, тұқымша, аналық, аталық, тікше, ұқсастық, әдіс, әдістеме, кеңес, жанартау* сынды көптеген терминдер әлі күнге дейін ғылыми еңбектерде, күнделікті тұрмыс-тіршілікте қолданылып келеді. Олардың қай-қайсы да қарапайым халық үшін өте түсінікті.

Бүгінгі бекітіліп жүрген терминдер Алаш зиялыларының оқулықтарындағы терминдердей болса, ғылым тілі ұлттық сипатта қалыптасады. 2006-2009 жылдары аралығында ҚР Үкіметі жанындағы республикалық терминология комиссиясымен бекітілген көптеген терминдер халық арасында ұлттық тілде белсенді қолданылып жүр. Мысалы:

| **Қазақ тілінде** | **Орыс тілінде** |
|---|---|
| *дыбыс* | *аудио* |
| *айқұлақ* | *собачка (@)* |
| *түйіндеме* | *резюме* |
| *мәселе* | *проблема* |
| *сарапшы* | *эксперт* |
| *пайыз* | *процент* |
| *үдеріс* | *процесс* |
| *шатқал* | *каньон* |

Кез келген терминнің негізінде уәж бен ассоциация тұрады. Терминнің объективтілігі адресант пен адресат арасындағы байланысты орнатады. Тілімізге сіңіп, ғылым тілінде орнын тауып үлгерген бұл терминдер жалпыхалықтық тілде ұлттық тіліміздегі таңбалануымен қолданылып жүр. Ш.Құрманбайұлының 2014 жылы шыққан «Қазақ терминологиясы» атты еңбегінде терминге қойылатын талаптардың бірі ретінде терминнің уәжділігін көрсеткен. «Терминнің уәжділігі – термин мағынасының ұғынықты, өзі белгілейтін ұғымы жөнінде анық мәлімет беруі» [Құрманбайұлы, 2014, 502 б.]. Уәж бен таңба бір-бірімен үйлесім тапқан терминдерді І.Жарылғаповтың еңбектерінен көруге болады. І.Жарылғаповтың ғылымға енгізген *балмұздақ, оқырман, көрермен, суреткер, қаламгер, аялдама, дүниетаным* деген терминдері бүгінгі күні терминологиялық қорымыздағы сәтті терминдер қатарында. Бұл терминдердің сыртқы формасы мен ішкі мағынасы бір-біріне үйлесімді түзілген. Терминдерді ұлттық мәдениетімізді, ұлттық тәрбиемен қалыптасқан болмысымызды, жылдар бойы жинақталған танымымызды, тәжірибемен жүйеленген аялық білімімізді бір арнаға тоғыстыру арқылы жасаған жағдайда ұлттық терминологияның негізі қалыптасады. Ғалым Ж.Манкееваның «Тіл – тек коммуникативтік құрал

емес, сонымен бірге адам болмысының, оның мәдениетінің көрінісі. Өйткені мәдениет таңба, белгіден тысқары, яғни тілден тысқары өмір сүре алмайды. Адамды түгелдей дерлік таңбалық әлем қоршаған. Өйткені адам болмысының өзі таңбалық, тілдік болмыс. Адам бір мезгілде таңбаны тудырушы да, оны талдаушы да. Тіл – тек денотативті (белгілі сигналдық) коммуникация құралы ғана емес, сонымен бірге коннотативті (белгілі әлеуметтік-мәдени, идеологиялық мәні бар) құрал. Тілде әр халықтың тарихы, оның өмірі, тіршілігі, шаруашылығы мен мәдениеті жатыр» [Манкеева, 2014, 126 б.] деген пікірінен тіл мен ұлттың ажырамас ұғымдар екенін көруге болады. Қазақ терминологиясын ұлттық дәрежеге жеткізу үшін терминдегі ұлт, ұлттағы термин көрінісін зерттеп, соған сәйкес ұлттық сипаттағы терминдер жасау керек. Сол терминдер қазақ тілінде болса, ұлттық тілді ғылыми тұрғыда танытуға мүмкіндік туады. Ұлттық тілде түзілген терминдерді білім алушыларға меңгерту кірме терминдерді меңгертуден әлдеқайда жеңіл. Білім алушы қазақ болса, қазақы тәрбие алса, оған төл терминдерге негізделіп түзілген ғылыми мәтінін оқу тиімді.

Бұл зерттеу жұмысына арқау болып отырған ғылыми мәтінді оқыту мәселесі ғалым С.Әлісжанның еңбектерінен табылады. Ғалымның 2017 жылы студенттерге арналған «Ғылыми дискурстың танымдық негіздері (коммуникативтік-прагматикалық аспект)» атты құралында ғылыми дискурстың белгілері мен ерекшеліктері, негізгі ұғымдары, ғылыми мәтіннің мағыналық құрылымы, ғылыми мәтіндердегі субъект факторы қарастырылған. Еңбек «Ғылыми дискурс: құрылымы мен белгілері», «Ғылыми дискурстағы субъект категориясы» және «Ғылым дискурсты метафоралық модельдеу» деген үш тараудан тұрады. Автор ғылыми мәтін мен ғылыми білім ұғымдарын қатар қолданып, ғылыми білімді мамандардың ғылыми-танымдық іс-әрекетінің нәтижесі деп түсіндіреді. Сонымен қатар ғалым ғылыми мәтіннің танымдық сипатына ғаламның ғылыми бейнесі тұрғысынан келген. Ғылыми мәтінді оқыту әдістемесіне, ғылыми мәтінді жазуға қатысты Б.Динаева мен С.Сапинаның «Академиялық сауаттылықтың теориялық және практикалық негіздері» және Е.Оспановтың «Академиялық жазылым» атты оқу құралдары бар. Бұл екі оқу құралында ғылыми мәтінді оқыту мәселесіне академиялық жазба тұрғысынан келген.

Қорыта айтқанда, ғылыми мәтінді оқыту мәселесі А.Байтұрсынұлы, С.Қожанұлы, Ж.Күдерин, Е.Омаруұлы, Ж.Аймауытұлы, М.Жұмабаев сынды ғалымдардың зерттеулерінен басталып, бүгінде осы ғалымдардың ғылымның түрлі саласы бойынша жазылған оқулықтары ғылыми мәтін теориясының, ғылыми мәтінді оқыту әдістемесінің негізі болып отыр. Тілдің танымдық сипатын зерттеу тілдің жаңа қырларын,

яғни коммуникативтік, дискурстық, прагматикалық ерекшеліктерін ашады. Оқыту саласында бұл ерекшеліктерді зерттеу аса маңызды болып саналады. Себебі оқыту үдерісі теория мен практика ғана, онда оқытушы мен білім алушы арасындағы қарым-қатынас, білім алушының ақпаратты қабылдауы, оқытушының интерпретациясы бар. Ғылыми мәтінді ұлтпен сабақтастықта оқыту ең әуелі ұлттық мүддені қорғауға, ұлттық тілді сақтауға, ұлттық ғылымды дамытуға, ұлттық ғылыми тілді қалыптастыруға көмектеседі.

**Әдебиеттер тізімі**

1. Құрманбайұлы Ш. Қазақ терминологиясы. – Алматы: Сардар, 2014. – 936 б.
2. Манкеева Ж. Қазақ тіл білімінің мәселелері. – Алматы, 2014. – 640 б.
3. Оразалиева Э. Когнитивті лингвистика: қалыптасуы мен дамуы. – Алматы, 2007. – 312 б.

# ТҮРКІ ЖӘНЕ ШЕТ ТІЛДЕРІН ОҚЫТУДЫҢ ИНТЕЛЛЕКТУАЛДЫ ТЕХНОЛОГИЯЛАРЫ

# ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ОБУЧЕНИЯ ТЮРКСКИМ И ИНОСТРАННЫМ ЯЗЫКАМ

# INTELLIGENCE TECHNOLOGIES FOR LEARNING TURKIC AND FOREIGN LANGUAGES

*ӘОК 004.9*

***[1]Абишева Ж.М., [2]Разахова Б.Ш.***
*Л.Н.Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]zhanna0789@gmail.com, [2]razakhova_bsh@enu.kz*

## БІЛІМ БЕРУДЕГІ ГЕЙМИФИКАЦИЯ

**Аңдатпа**. Бұл жұмыс геймификация қосымшаларының білім берудегі ықпалына шолу мақсатында жазылды. Ол үшін геймификация әдісін білім беру саласында қолдану бойынша зерттеулер жүргізілді. Ойын элементтерін пайдалану арқылы қандай жетістіктерге жетуге болатыны, оның артықшылықтары мен кемішіліктері көрсетілді. Геймификация арқылы ойынға айналдырылған оқыту жүйесін ұстану қазіргі заманауи технологияларды пайдалануға және білім беру процесін цифрландыруға алып келеді. Ойын элементтерінің бірдей құралдар жиынтығын қолданатын, бірақ бірдей элементтерді әртүрлі тәсілдермен қолданатын қосымша тәсілдер ретінде геймификацияның ойын дизайнын таңдау өте маңызды.

**Түйін сөздер:** Геймификация, білім берудегі тенденция, заманауи технология, қашықтықтан оқыту, ойын элементтері, білім беруді цифрландыру, ойын механикасы, оқыту әдісі, ойын сценарийлері

*УДК 004.9*
***[1]Абишева Ж.М., [2]Разахова Б.Ш.***
*Евразийский национальный университет имени Л.Н. Гумилева*
*Нур-Султан, Казахстан*
*[1]zhanna0789@gmail.com, [2]razakhova_bsh@enu.kz*

# ГЕЙМИФИКАЦИЯ В ОБРАЗОВАНИИ

**Аннотация**. Эта работа была написана с целью обзора влияния приложений геймификации в образовании. Для этого были проведены исследования по применению метода геймификации в образовательной области. Было показано, каких успехов можно достичь, используя игровые элементы, его преимущества и недостатки. Следование системе обучения, преобразованной в игру посредством геймификации, приводит к использованию современных технологий и цифровизации образовательного процесса. В качестве дополнительных подходов, использующих один и тот же набор инструментов для элементов игры, но использующих одни и те же элементы по-разному, очень важно выбрать игровой дизайн геймификации. Во-первых, мы определили концепцию геймификации и объяснили, как она используется для разработки высокопроизводительных процессов в различных областях обслуживания. Во-вторых, мы определили основные принципы и аспекты геймификации, эффективные концепции. В-третьих, мы изучили связь между геймификацией и образованием с целью обучения через игровые элементы. Геймификация дополняет и поддерживает уже существующий учебный контент. Следует отметить, что Геймификация включает в себя не только баллы, значения и уровни, но и набор подходов. С помощью этого метода можно получить положительные результаты в образовании. Игровой режим и ценность геймификации, а также правила мотивации могут быть использованы в качестве мощных инструментов для образования и обучения. Приобретение собственных полезных элементов игры, правил, условий выигрыша, наград, наказаний, статуса, вклада и других индивидуальных действий может выявить сильные и слабые стороны ученика.

**Ключевые слова**: геймификация, тренды в образовании, современные технологии, дистанционное обучение, игровые элементы, цифровизация образования, игровая механика, методика обучения, игровые сценарии.

*UDC 004.9*
***[1]Abisheva Zh., [2]Razakhova B.***
*L.N.Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]zhanna0789@gmail.com, [2]razakhova_bsh@enu.kz*

# GAMIFICATION IN EDUCATION

**Abstract**. This paper was written to review the impact of gamification applications in education. For this purpose, studies were conducted on the application of the gamification method in the educational field. It was shown what successes can be achieved using game elements, its advantages and disadvantages. Following a learning system transformed into a game through gamification leads to the use of modern technologies and digitalization of the educational process. As additional approaches using the same set of tools for game elements, but using the same elements in different ways, it is very important to choose a gamification game design. First, we defined the concept of gamification and explained how it can be used to develop high-performance processes in various service sectors. Secondly, we have identified the main principles and aspects of effective concepts of gamification. Third, we have studied the relationship between gamification and education to learn through game elements. Both are ultimately designed to influence the same criteria: learning and getting positive results about it. Gamification complements and supports existing educational content. It should be noted that gamification includes not only points, values, and levels, but also a set of approaches. Gamification is a powerful way to attract attention and increase enthusiasm. The game mode and value of gamification and the rules of motivation can be used as powerful tools for education and training. Acquiring its own useful elements of the game, rules, winning conditions, rewards, punishments, status, contributions, and other individual actions can reveal the student's strengths and weaknesses.

**Keywords:** gamification, trends in education, modern technologies, distance learning, game elements, digitalization of education, game mechanics, teaching methods, game scenarios.

## Кіріспе

Негізінде, геймификация – назар аудартудың, зейін қоюдың және құлшынысты арттырудың күшті әдісі болып табылады. Геймификацияның ойын режимі мен құндылығын және ынталандыру ережелерін білім беру мен оқытудың мықты құралы ретінде пайдалануға болады. Ойынның өзіндік пайдалы элементтерін алу

арқылы ережелер, жеңіс шарттары, марапаттар, жазалар, мәртебе, үлестер және басқа да жеке әрекеттер білім алушының қарқынды және әлсіз тұстарын тауып бере алады. Испандық ғалым Сержи Вильяграсаның пікірінше, маркетингтің негізгі тұжырымдамасы тартымды, бірақ бір біріне ұқсамайтын идеяларды бір арнаға біріктіру болып табылатындықтан, денсаулық пен спорт, білім және мансап - сол сияқты, геймификацияда да көңіл көтеру мен білім алу идеясын бір арнада тоғыстыру [Вильяграса, 2014, б.40]. Білім беруді заманауи тұрғыда технологияландыру арқылы ойынның неше түрлі элементтерін пайдалана отырып, ойынсыз мәнмәтінді ойын түріне айналдыру болып табылады. Миннесота Университетінің доценті, зерттеуші Ричард Ландерстің пікірінше, геймификация ұғымы қазіргі заманғы қашықтықтан электронды білім беруді дамытудың негізгі тенденцияларының бірі болып табылады [Ландерс, 2014, б.760]. Негізінде, білім алушы ойын идеясына ілігіп, көбірек білуге немесе жеңуге деген қызығушылықпен алға ұмтылады, соңында ойын элементтері арқылы өзіне қажетті ой түйіп, идея немесе әрекеттің әсерінен өзіне қарқынды білім алу мүмкіндігін ашады.

**Геймификация ұғымы**

Ойындар еліктеуді тудыртады - ол кез келген оқу тәжірибесіндегі ең басты қажеттілік. Белгілі бір стратегияны өту, сайыстық есептерді шешу, кодтық алгоритмдерді ойын арқылы меңгеру немесе ойын элементтерімен қазақ тілін үйрену сияқты тәжірибені геймификация әдісі арқылы бейімдеуге болады.

Қазіргі таңда, алдыңғы қатарлы мемлекеттердің көпшілігі сандық ойындарды ұстанады. Американдық ғалым Брайан Арнольдтың жүргізген зерттеуі бойынша ойыншылардың жас диапазоны жыл сайын жасарып, ал ересек ойыншылар болса бұрыннан да әлдеқайда жақсы ойнайды [Брайан Арнольд, 2014, б.32]. Планшеттер мен смартфондардың көбейе түсуіне байланысты, бұдан былай үстелдік және портативті компьютерлер, теледидарлар және ойын консольдері онлайн ойын жүйесіне қосылудың жалғыз жолы болып табылмайды. Ойын өз кезегінде туризм аймағын қамтып, сауда, өндірістік және білім әлеміне еніп, пайдалы білім алу әрі адамды оқуға жігерлендіретін құрал ретінде өзін көрсете білді. Көптеген оқу орындары мен дамыту орталықтары ойын арқылы білім беру тәжірибесін дамытып жатқандықтан, ойын сценариі мен ойын элементтерін, механикаларын және ойыннан тыс кездегі фреймворктарды интеграциялау сияқты геймификация әдісіне көбірек көңіл бөліне бастады. Білім алу құқығынан айырылған жастар арасында білім беруді сақтау мен

дағдыларды меңгеруді жақсарту үшін оқуға қатысуды ынталандырушы бірден бір құрал ретінде геймификацияны қалай және қандай деңгейде пайдалануға болады деген сұрақ туындайды.

Көптеген кәсіпорындар қызметкерлерді сыйақы мөлшерімен марапаттау, көшбасшылар тақтасы және төсбелгілер арқылы жұмысқа деген қызығушылығын арттыру мақсатында ынталандыру бағдарламалары мен мобильді қолданбаларды әзірлеу үшін геймификация әдісін кеңінен қолдануда. Әлі де қалыптасу сатысында болса да, білім беру саласында геймификациялауды зерттеушілер мен оны пайдаланатын педагогтар арасында геймификция әдісі үлкен қолданысқа ие, өйткені олар білім алушылардың нәтижелілігі мен шығармашылық ізденісін ынталандырудың бірден бір жолы ойын элементтері екенін біледі. Геймификацияның негізгі құндылығы – оқыту процесіне ерікті түрде қатысудың жоғарғы көрсеткіші болып табылады. Бұл өзара әрекеттесу жаңа қатысушыны тартып қана қоймай, сонымен қатар білім алушының күнделікті қолдануын ынталандыратын жаңа әдетке және мәдениетке әкелуі мүмкін. Ойынның қызығушылығына бір кірген соң, онымен өткізілетін уақыт мөлшері де ұлғаяды, жеткен жетістіктермен қоса сол ойынға деген сенімділік те арта түспек.

Мобильді технология жетістіктерінің арқасында ойын мүмкіндіктерінің өрісі одан әрі кеңейіп, білім алушыларға кез келген уақытта, кез келген жерден жүйеге қосылуға мүмкіндік бар. Смартфоны немесе планшеті бар кез келген адам ойыншы бола алады. Қазіргі таңда көптеген тегін мобильді ойындар бар және олардың ішіндегі ең танымалдылары отбасымен, достарымен әлеуметтік өзара әрекеттесу және қарым қатынас құралы болып табылады. Мобильді ойындардағы әлеуметтік желі функциялары барлық уақытта желіде болуды қамтамасыз ететін ойынның кеңінен таралуын қолдайды; онлайн ойынның тартымдылығы тек кім ойнайтынында ғана емес, сонымен қатар жеке желіде басқа ойыншылармен сайысқа түсіп, бақ сынасу болып табылады. Бұл қолжетімді ойын технологияларының кең таралуы, барлық жерде болуы өзара әрекеттесуге ыңғайлы жаңа уақыт шеңберлерін ашады.

**Геймификация әдісі және оның негізгі аспектілері**

Геймификация әдісі білім беру саласында қолданылған кезде, тәжірибелік, өзін-өзі басқару және өмір бойы білім алу мүмкіндіктері шегін экспоненциалды түрде кеңейте түседі.

Геймификация мидың атқарылған жұмыс үшін сыйақы алуға ұмтылу қасиетін пайдаланады. Әрбір әрекеттен кейін әртүрлі

жетістіктер түрінде берілетін марапаттарды кез келген онлайн ойындардан байқауға болады.

Геймификациядағы негізгі ұстаным – мінез-құлықты интерактивті реттеуді қамтамасыз ету және ойынның барлық функционалдығын тез меңгеру үшін білім алушыдан тұрақты түрде өлшенетін кері байланыс алып отыру болып табылады.

Геймификацияның тағы бір ұстанымы – ойын немесе әрекеттің басталу тарихы және сол туралы аңыз құру. Білім алушы ойын технологияларының көмегімен өзінің білім алу барысында жеткен жетістіктерінің хронологиясын құра алады, оларға шолу жасай алады. Бұл білім алушылар арасында іске араласушылық, бәсекелестік, ортақтастық сезімін, қандай да бір ойластырылған мақсаттарға жетуге деген қызығушылықты тудырады. Сонымен қатар, білім алушының ойынға қатысу дәрежесін сақтай отырып, ойын дағдыларының дамуын қамтамасыз ететін жаңа дағдылар мен қабілеттіліктерді меңгеруіне байланысты тапсырмалардың кезең-кезеңімен өзгеруі және күрделенуі қолданылады.

Ресейлік зерттеуші Вадим Матониннің мақаласында аталып өткендей, геймификацияның ескеру қажет негізгі аспектілері мынандай [Матонин, 2017, б.37]:

- динамикалық компонент – нақты уақытта білім алушының назарын және реакциясын қажет ететін әртүрлі сценарийлерді пайдалану;
- ойын механикасы – ойын сыйақылары, атақ-мансап, бонустар, элементтер сияқты геймплейге тән сценарийлерді пайдалану;
- эстетикалық компонент – эмоционалды түрде еліктіретін ойын графикасынан көрнекі әсер жасау;
- әлеуметтік компонент – білім алушылар арасындағы өзара әрекеттесуді, коммуникацияны және тәжірибе мен ақпарат алмасуды қамтамасыз ету.

Геймификацияның тиімді тұжырымдамасы – білім алушылардың назарын аудару, қызықтыру, сергіту, сынақтан өткізу және тәрбиелеу болып табылады. Румыниялық зерттеуші Юлиан Фурдудың мәлімдеуінше, оқыту процесінде геймификацияны қолданудың кейбір артықшылықтары мыналар [Фурду, 2017, б.59]:

Ойын кезінде «ойын-сауық» терминін оқытумен біріктіру арқылы үздік оқу тәжірибесіне қол жеткізуге болады.

Жақсы ұйымдастырылған ойын стратегиясы білім алушыларды белсендірек етеді, ал қатысудың жоғары деңгейі кері байланысты, есте сақтауды арттыруға көмектеседі.

*Кері байланыс.* Геймификация метриканы қолданатындықтан, білім алушының қаншалықты жетістікке жетіп жатқанын бақылап отыруға болады. Білім алушының көзқарасы бойынша, тесттер мен тапсырмалар және басқа да іс әрекеттер кері байланыстың әртүрлі деңгейлерін қамтамасыз етеді, осылайша білім алушылар не үйренгенін немесе нені үйрену керек екеніне өздігінен сараптама жасай алатын болады.

*Ең жақсы оқу ортасы.* Оқу тәжірибесі дербестендірілген; білім алушылар оңтайлы әдіс арқылы өздігінен дами алады. Марапаттау жүйесі білім алушыларға шынайы өмірлік жағдайлар мен тапсырмаларды орындауға көмектесетін тиімді бейресми оқу ортасын қамтамасыз етеді.

Геймификация ұпайлар, белгілер және атақ-мансап деңгейлерімен қамтамасыз етілген беткі деңгейдің артықшылықтарынан әлдеқайда жоғары, өйткені циклдік оқытудың ғылыми принциптері мен есте сақтау қабілеті бір жүйеде тоғысқанда мінез-құлықтың өзгеруіне септігін тигізбей қоймайды. Геймификация жан-жақты болып табылады, өйткені ол арқылы көптеген оқыту қажеттіліктерін қанағаттандырып қана қоймай, өнімді сату, тұтынушыларды қолдау, әлеуметтік дағдылар, хабардарлықты арттыруға болады, нәтижесінде ұйымдардың өнімділігі артады.

Сонымен қатар, шетел ғалымдарының пікірінше, геймификация білім алушылардың оқу-тәрбие процесіне жақсырақ араласуына ықпал етеді. Мысалы, Диан Шаффхаузер оқу іс-әрекетінде ойын әдістерін қолданудың келесі артықшылықтарын атап көрсетті [Шаффхаузер, 2009, б.28]:

- білім алушының жеке тәжірибесінің оқумен байланысы (ойын ортасын зерттелетін құбылыстарды, объектілерді, жағдайларды визуализациялау арқылы пайдалану);
- оқу барысында бәсекеге қабілеттілік пен салауатты бәсекелестік рухын қалыптастыру;
- топтық жұмыс;
- жүйелі ойлауды дамыту;
- білім алушының үздік нәтижеге қол жеткізуі үшін итерацияны қолдану.

Дегенмен, ағылшындық зерттеуші Себастьян Детердингтің мақаласында көрсетілген мәліметтерге сүйенсек, геймификацияны шамадан тыс көп немесе теріс пайдалануда ескеру қажет кейбір кемшіліктері бар [Детердинг, 2011, б.13]. Ойын элементтерін пайдалануды міндетті ету арқылы белгілі бір ережеге негізделген тәжірибе жинақталуы мүмкін. Шеберлікті емес, күш-жігерді марапаттау керек және білім алушылар әрбір сәтсіздікті жеңіске жетудің тағы бір

мүмкіндігі ретінде қабылдауды үйренуі тиіс. Ойын арқылы білім беру құрылымы, сәтсіз талпыныс орын алған жағдайда, білім алушылар тапсырманы қайта қайталай алатындай етіп жасалуы керек. Кері байланыс білім алушылардың іс-әрекетін түзету және олардың келесі іс-әрекетіне ынталандыру мақсатында болуы қажет. Сонымен қатар, ойын барысында қолданылатын метрикалар мен сабаққа нақты қатысу деңгейі теңестірілуі керек. Тапсырмалар дизайны мен контент қойылымы мүмкіндігінше бейтарап болуы үшін мұқият ойластырылған болуы керек, бірақ сонымен бірге зеріктіргіш және қызықсыз болып көрінбеуі керек.

**Геймификацияның білім беру саласындағы тигізер пайдасы**

Кәсіпкерлер, студенттер, пациенттер, қызметкерлер немесе тұтынушылар болсын, барлық ұйымдар мүдделі тараптарды ынталандыруы және қызықтыра білуі керек. Білім беру саласында да білім алушының қызығушылығын оятып, оқуға деген ынтасын арттыру ең алғы мақсаттардың бірі болмақ. Ойын элементтерін пайдалану - мақсатқа жетудің бірден бір тәсілі. Канадалық ізденуші Карен Робсонның жүргізген зерттеулеріне сүйене отырып, геймификация жүйесін пайдалану арқылы білім беру саласында пайда әкелетін мынандай нұсқаулықтарды ұсынамыз [Робсон, 2015, б.411]:

1. Ойынның мақсаты қандай? Оқыту процесін геймифика-циялау үшін біз қол жеткізгіміз келетін мақсатты анықтау. Ол қаржылық, әлеуметтік немесе экологиялық мақсаттарға негізделуі керек. Жоғары оқу орындары осы мақсаттарға жету үшін қажетті мінез-құлық пен нәтижелерді жасау және түзету үшін геймификацияны пайдалану әлеуетін бағалауы керек. Екі немесе үш мақсатқа емес, бір мақсатқа назар аудару күрделілікті және ұйымдастыру кезеңдері, динамикалық, эмоциялық әсерлердің бір-біріне қайшы келу мүмкіндіктерін азайтады. Білім беру орындары сонымен қатар геймификацияның басқа да әдіс тәсілдерін пайдалану арқылы әртүрлі шаралары мен мақсаттарын анықтауы керек. . Ойынға айналдыру шаралары мен оқу орнының білім берудегі мақсаттары арасындағы себепті байланысты анықтау өте маңызды.

2. Ойын ойнау. Адамдар ойын процесін алдап көргісі келеді. Ойыншылар да, бақылаушылар да, көрермендер де сөз байласып, ережелерді бұзу арқылы ойын нәтижесін өз пайдасына шешуге тырысады. Осындай адами әрекеттердің оң және теріс жақтарын түсіну. Бір жағынан, ол басқа ойыншыларды, бақылаушыларды және көрермендерді ығыстыруы мүмкін жағымсыз эмоцияларға, әділетсіздікке әкелетін динамика жасай алады. Дегенмен, ережелерді

бұзу процесінен алынған оң білім мен өзгерістер болуы мүмкін. Мысалы, бұл мінез-құлық геймификацияланған процестің механикасын өзгертуге негіз бола алады, осылайша әділдік пен шыншылдық принциптерін тереңдетуге және нәтижелерді жақсартуға болады.

3. Тәжірибені жаңарту және бекіту. Күнделікті қолданысқа енгізіліп отырған заманауи жаңа ойын технологияларын пайдалану, трендтік ойындармен ілесіп жүру білім алушының білімге деген көзқарасын өзгертпек. Әр түрлі ойын құралдары қолданылған кезде ойын тәжірибесі де өзгереді. Кез келген стратегиялық инвестиция сияқты, басқарушылық міндеттер мен стратегиялық мақсаттарға назар аудару маңызды. Бұл ойынға айналдырылған тәжірибенің ішкі және сыртқы құрылымы, мазмұны үнемі бақыланатынын білдіреді.

4. Барлық әртүрлі рөлдерді тану. Көптеген геймификация мысалдары тек дизайнер мен ойыншылар арасындағы байланысқа бағытталған. Бұл маңызды, себебі ойыншыға қатысты нәтижелер арасындағы байланыстарды терең түсінуге ықпал етеді. Дегенмен, бір желіде бірнеше ойыншының болуы және олардың ойынға қатысуы процесте әртүрлі мінез-құлық пен нәтижелерді қалай қуаттандырып, бағыттай алатынын түсіну маңызды.

5. Ойын немен аяқталды? Соңында ойын тәжірибесі өз мәресіне жетіп, нәтиже алынуы тиіс. Тәжірибені үнемі жаңарту және оны қолданысқа енгізу барысында алынған нәтиженің ұйымға тигізер пайдасы зор. Ойынның соңы – геймификацияланған процестің өміріндегі соңғы кезең деп аталады. Дизайнерлер бұл фазаның бар екенін білуі керек және олар қатысып отырған көрермендер мен бақылаушылардың қайта оралуға, жаңа ойын процестеріне қатысуға дайын болатындай етіп баптай алуы тиіс.

**Қорытынды**

Біз бұл жұмыста геймификация әдісін білім беру саласында қолдану арқылы қандай жетістіктерге жетуге болатынын көрсете отырып, сонымен қатар кемшілікерін де атап өттік. Біріншіден, біз геймификация ұғымын анықтап, оның әртүрлі қызмет көрсету салаларында жоғары өнімділік процестерін дамыту үшін қалай қолданылатынын түсіндірдік. Екіншіден, геймификацияның негізгі ұстанымдары мен аспектілерін, тиімді тұжырымдамаларын анықтадық. Үшіншіден, ойын элементтері арқылы оқыту мақсатында геймификация мен білім беру арасындағы байланысты зерттедік. Геймификация бұрыннан бар оқу мазмұнын толықтырады және қолдайды. Геймификация ұпайларды, мәндерді және деңгейлерді ғана емес, сонымен қатар тәсілдер жиынтығын да қамтитынын атап өткен жөн. С

помощью этого метода можно получить положительные результаты в образовании.

**Әдебиеттер тізімі**

1. Arnold, B. J. (2014). Gamification in education. Proceedings of the American Society of Business and Behavioral Sciences, 21(1), 32-39.

2. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining" gamification". In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments (pp. 9-15).

3. Furdu, I., Tomozei, C., & Kose, U. (2017). Pros and cons gamification and gaming in classroom. arXiv preprint arXiv:1708.09337,56-62.

4. Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. Simulation & gaming, 45(6), 752-768.

5. Robson, K., Plangger, K., Kietzmann, J. H., McCarthy, I., & Pitt, L. (2015). Is it all a game? Understanding the principles of gamification. Business horizons, 58(4), 411-420.

6. Schaffhauser, D. (2009). Which came first--The technology or the pedagogy?. The Journal, 36(8), 27-32.

7. Villagrasa, S., Fonseca, D., Redondo, E., & Duran, J. (2014). Teaching case of gamification and visual technologies for education. Journal of Cases on Information Technology (JCIT), 16(4), 38-57.

8. Матонин, В. В. (2017). Тренды современного образования: геймификация. Вестник Бурятского государственного университета. Образование. Личность. Общество, (2), 36-40.

*ӘОК 372.851*
***[1]Бекбауова А.У., [2]Ғилманова Н.Т.***
*Қ.Жұбанов атындағы Ақтөбе өңірлік университеті,*
*Ақтөбе, Қазақстан*
*[1]Mirra478@mail.ru, [2]gilmanova2000@list.ru*

# CLIL ТЕХНОЛОГИЯСЫН АЛГЕБРА ПӘНІН ОҚЫТУДА ҚОЛДАНУ

**Аңдатпа.** Мақала алгебраға интеграцияның тұжырымдамалары мен принциптеріне пәндік-тілдік кешенді тәсіл (CLIL) әдісін қолдануға арналған. CLIL әдісінің негізгі мақсаты – «Үштілділік» ұлттық жобасы аясында ағылшын тілін жаратылыстану ғылымдарымен ықпалдастыру арқылы оқушылардың оқуға деген қызығушылығын арттыру, ағылшын тілін меңгеру дағдыларын дамыту. Интеграция ұғымдары мен принциптеріне (CLIL) пәндік-тілдік кешенді көзқарас әдісінің қағидалары мен үлгісі келтірілген. Математикадан «Жай бөлшектерге амалдар қолдану» тақырыбы бойынша пәнді тіл арқылы меңгертуге және пән мазмұны арқылы тілді меңгеруге бағытталған сабақ жоспары ұсынылған. Әр сабақта диалогқа қажетті пәндік терминология мен сөз тіркестерін енгізу, сонымен қатар диалог пен монолог арқылы оқушылардың бір-бірімен қарым-қатынасын дамытуға арналған жаттығулар сияқты алгебраны оқытуда CLIL-ді қолданудың қажетті шаралары көрсетілген.

**Түйін сөздер**: Интеграция, жоғары ретті ойлау дағдылары, оқыту мен оқудағы жаңа тәсілдер, инновация.

*УДК 372.851*
***[1]Бекбауова А.У., [2]Ғилманова Н.Т.***
*Актюбинский региональный университет имени К.Жубанова*
*Актобе, Казахстан*
*[1]Mirra478@mail.ru, [2]gilmanova2000@list.ru*

# ПРИМЕНЕНИЕ ТЕХНОЛОГИИ CLIL В ПРЕПОДАВАНИИ АЛГЕБРЫ

**Аннотация.** Статья посвящена примеру применения метода предметно-языковому комплексному подходу к понятиям и принципам интеграции (CLIL) к алгебре. Основной целью метода CLIL является повышение интереса учащихся к обучению за счет интеграции

английского языка с естественными науками, развитие у них навыков владения английским языком в рамках национального проекта «Үштілділік». Приведены принципы и модель метода предметно-языковому комплексному подходу к понятиям и принципам интеграции (CLIL). Представлен план урока по математике на тему «Apply operations on simple parts», нацеленные на изучение предмета через язык и изучение языка через содержание предмета. Указаны необходимые меры для применения CLIL при изучения алгебры, такие как введение на каждом уроке предметной терминологии и фраз, необходимых для диалога, а также упражнений на развитие общения учащихся друг с другом в диалоге и монологе.

**Ключевые слова:** Интеграция, мыслительные навыки высокого порядка, новые подходы в преподавании и обучении, инновация.

*UDC 372.851*

***[1]Bekbauova A., [2] Gilmanova N.***
*Aktobe regional university named after K.Zhubanov*
*Aktobe, Kazakhstan*
*[1]Mirra478@mail.ru, [2]gilmanova2000@list.ru*

## APPLICATION OF CLIL TECHNOLOGY IN TEACHING ALGEBRA

**Abstract.** The article is devoted to an example of applying the method of the subject-language complex approach to concepts and principles of integration (CLIL) to algebra. The main goal of the CLIL method is to increase students' interest in learning by integrating English with the natural sciences, developing their English language skills within the framework of the national project "Үштілділік". The principles and model of the method of the subject-language complex approach to concepts and principles of integration (CLIL) are given. A lesson plan in mathematics on the topic "Apply operations on simple parts" is presented, aimed at studying the subject through the language and learning the language through the content of the subject. The necessary measures for the application of CLIL in the study of algebra are indicated, such as the introduction of subject terminology and phrases necessary for dialogue at each lesson, as well as exercises to develop students' communication with each other in dialogue and monologue.

**Key words:** Integration, high thinking skills, new approaches to teaching and learning, innovation.

Мемлекет басшысы Қасым – Жомарт Тоқаев Тамыз конференциясында үштілділік бағдарламасында: «Үш тілді білім беру жүйесін енгізу – күн тәртібіндегі өзекті тақырып. Ең алдымен, бұл мәселені жан – жақты талдап, салмақты шешімге келу қажет. Бұл өте маңызды. Қазақ тілі үш тілдің біреуі болып қалмайды. Үш тілдің біріншісі, негізгісі, бастысы, маңыздысы бола береді» деп Елбасы өнегелі ойын айтты.

CLIL кіріктірілген оқыту бағдарламасы – бірнеше тілді меңгерген, әлеуметтік және кәсіптік бағдарға қабілетті, мәдениетті тұлғаны қалыптастыруды мақсат етеді[1].

CLIL әдісін 1994 жылы көптілді білім беруді зерттеуші Дэвид Марш Еуропадағы тілдік білім берудің жағдайын үйлестіру бойынша жұмыс істеу барысында енгізді.

«CLIL» әдісінің негізгі мақсаты: ағылшын тілін жаратылыстану бағытындағы пәндермен кіріктіріп оқыту арқылы оқушылардың білім алуға қызығушылықтарын арттырып, олардың ағылшын тілінде сауатты түрде «Үштілділік» ұлттық жоба аясында білім алуын, дамуға дағдыландыру қажет. Пән мен тілді кіріктіріп оқыту әдісі негізінде сабақта берілетін тапсырмалар мен жұмыстар тілді дамыту бағытында жүреді. Яғни тілдік емес пән оқытушылары тек пәннің мазмұнына ғана емес, сонымен қатар сол пән арқылы оқушылардың тілдік дағдыларын байыта түседі. Бұл әдіс арқылы бірден екі мақсатқа жетуге болады: ағылшын тілі арқылы сабақ мазмұнын түсіндіру және пәндерді меңгеру арқылы ғылыми тіл дағдыларын дамыту болып табылады [2,3,4].

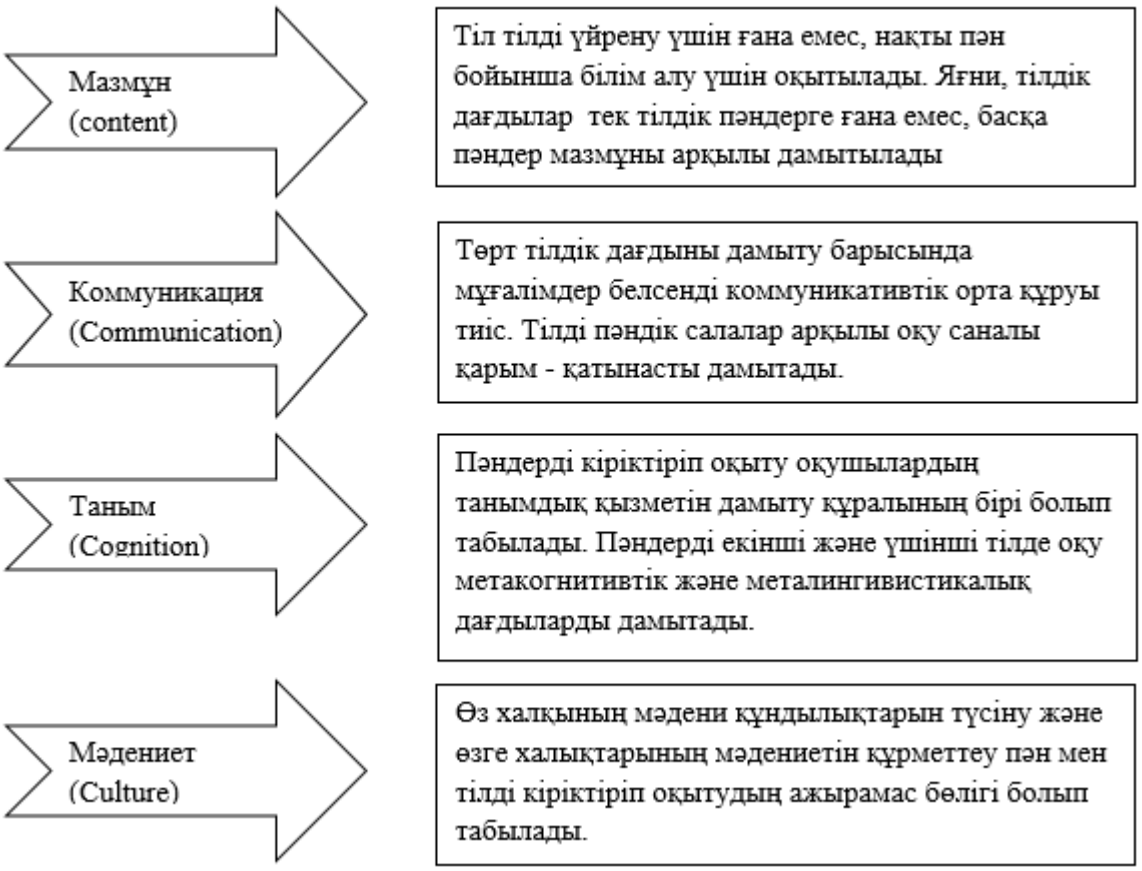
Халықаралық сарапшылар CLIL әдісін төмендегідей анықтайды:

•«тілді мәдениетаралық ойлауға дайындық және жалпы білім беруді жақсартудың құралы ретінде қарайтын әдіс»;

•«пән арқылы тілді зерттеуге және пәнді зерттеуге бағытталған әдісі»;

•«пәндерді екінші немесе үшінші тілді зерттеу арқылы және пәндік облыстарды зерттеу арқылы тілдік дағдыларын жетілдіру, яғни бір мезгілде пәндік және тілдік мақсаттарға қол жеткізу».

Пән мен тілді кіріктіріп оқытудың өзара тығыз байланысты төрт негізгі қағидатын қарастырайық:




Пән мен тілді кіріктіріп оқыту қағидаттары төмендегідей:

- ерекше дарынды оқушыларға мәндіді де саналы оқуға;
- сабақ құрылымында проблемалық жағдайлар үлесін ұлғайтуға;
- жеке тұлғаның зерттеушілік типін қалыптастыратын оқушының ойлау әрекетін белсендіруге;
- оқушыға бір мезгілде әрекетті орындауда барлық мақсат қоюдан нәтижеге дейінгі барлық процесіне бақылау жасауға;
- білім берушілік, дамытушылық және танымдық үш бірлікті ұстаздың шексіз шығармашылығы тұрғысынан жүзеге асыруға;
- сабақтың ақпараттық сиымдылығын арттыруға;
- оқушылардың уәждемесін арттыруға, олардың оқу танымдық әрекетін белсендіруге, шаршағандары мен қажуларын азайтуға;
- сабақ үрдісінде өтілетін тақырыптың адам және қоғам өмірімен тікелей байланысына мән беруге;
- оқушылардың шығармашылық ойлауын дамытуға;
- оқушылардың алған білімдерін өмірде қолдана білулеріне көмектесетін әдіс.

Яғни,оқыту, тәрбиелеу, дамыту, ынталандыру төрттігін басшылыққа ала отырып, атап көрсеткен қағидаттар негізінде

оқушылардың білімін бағалаймын. Бұл оқушының сабақта өзін - өзі және өзгені бақылауына, бағалауына мүмкіндік туғыза алады.

CLIL-технологияларының **үш түрлі моделі анықталған[5]**:

− soft (жұмсақ ) CLIL, language-led деген, он жерде арнайыланған мәнмәтіндердің лингвистикалы ерекшелігіне мән беріледі;

− hard (қатты) CLIL не subject-led, мұнда пәннің 50 пайызы ағылшын тілінде оқытылады;

− үшінші модель аталып отырған модельдің аралығында орташа дәрежелерге ие болады және кей пәндеріміз шетел тілдерінде оқытылу барысында пайдаланады (partial immersion).

1-кесте – CLIL пән-тілдік ықпалдандырылған оқытудың тұжырымдамалық негізі

| | |
|---|---|
| Дефинициялар | Шетел тілдері немесе 2-ші тіл негізінде мазмұн бойынша оқытудың әдістер жиынтықтары, олардың негізінде бір уақытта тілдердің өздері оқытылатыны анық |
| Мақсаттары | Шетел тілдерінің негізінде пәндерді оқыту және шетел тілін оқытылатын пән арқылы |
| Методология | Оқытудың қосбірлікті мақсатына қол жеткізу |
| Тұжырымдамалық компоненттер | 4 C<br>– «Content» - пәнді-мазмұндық компоненттер<br>– «Communication» - әлеуметтік тілдік компоненттер<br>– «Cognition» - танымды-оқыту компоненттері<br>– «Culture» - мәденилік компоненттер |
| Аспектілер | – Мәдениет аспектілері<br>– Әлеуметтік аспектілері<br>– Тілдік аспектілері<br>– Пәндік аспектілері<br>– Оқыту аспектілері |
| Қолданудың моделдері | – soft (жұмсақ) CLIL<br>– hard (қатты) CLIL<br>– middle CLIL |

Көптеген зерттеушілер пәндік-тілдік интеграцияда екі негізгі тәсілді бөліп көрсетеді:

- Content-driven education, оқу процесі бірінші кезекте оқу пәнінің мазмұнын зерттеуге бағытталған;

• Language-driven education, пәндік мазмұнға негізделген шет тілін үйренуге бағытталған.

Сабақ барысында оқушының білімін, қызығушылығын, логикалық ойлау дағдыларын арттыру мақсатында "Миға шабуыл", "... қатысты 10 сөзді атаңыз", "Лабиринт" әдістерімен қатар learning.apps платформасы арқылы графикалық диктант пен тест тапсырмалары берілді. Сабақ соңында «Репортаж» әдісі арқылы рефлексия жасалды. Сабақта бағалау парақтары таратылып, өзіндік бағалау жүргізілді.

Ағылшын тілінде алгебра пәнін оқытудың қалыпты жағдайдан айырмашылығы бар. Негізгі айырмашылық оқушылардың ағылшын тілін еркін меңгермегендігінде, соған сәйкес CLIL әдісі қолданылған сабақтарда оқушылар ақыл – ой әрекетінен гөрі танымдық әрекетін іске асыруға болады. Алгебра пәнінде кіріктірілген бағдарламада арнайы тілдік мақсаттар көрсетілген және осыны ескере отырып әр сабаққа пәндік терминология және диалогқа қажетті сөз тіркестерін енгізу, диалогтік және монологтік түрде оқушыларды бір – бірімен тілдесуін дамытуы үшін түрі жаттығулар жасау көзделіп отыр.

Математикалық пәндерде Clil технологиясын қолданудың маңыздылығы, ең алдымен білім алушылардың математикалық терминдерге байланысты сөздік қорын кеңейте отыра, өз бетімен ізденуге, шығармашылық қабілеттерін шыңдауға және математикалық есептерді ағылшын тілінде жеткізе білуге дағдылануға үлкен мүмкіндік туғызады.

**Әдебиеттер тізімі**

1. Методическое пособие «Использование подхода предметно-языкового интегрированного обучения (CLIL) в Назарбаев Интеллектуальных школах»

2. Kambatyrova, A. (2012). Көптілді білім контекстінде Content Language Integrated Learning (CLIL) тәсілінің мысалында тіл мен пәндік оқытуды интеграциялаудың әдістемелік ерекшеліктері [Methodological features of Content Language Integrated Learning (CLIL) in the context of multilingual education]. *Bilim Educational Research Journal,* 4, 73-78.

3. Алметов Н.Ш., Методы и приемы CLIL. ЮКГУ им М.Ауэзова, 2016

4. Усманова З.Ф., Заяц Т.Ф., Мукажанова Г.Ж. Реализация технологии CLIL в условиях полилингвального обучения // Филология и лингвистика в современном мире: материалы I Междунар. Науч. Конф. (г. Москва, июнь 2017 г.) – М.: Буки – Веди, 2017.

5. Математика:Жалпы білім беретін мектептің 5 – сыныбына арналған оқулық. 1 – бөлім. Т.Алдамұратова, Е.Байшоланов, 4 – бас., -Алматы: Атамұра, 2015, 96 – бет.

6. Қазақстан Республикасында 2017 -2021 жылдарға дейінгі білім беруді дамытудың мемлекеттік бағдарламасы, Астана.2017ж.

7. URL:*http://infourok.ru(МЕТОДЫ* И ПРИЕМЫ CLIL. ( Алметов Н.Ш., ЮКГУ им. М.Ауэзова, 2016)

8. URL:*http://zkoipk.kz/math-textbook-project/5-class/1797-matem-5.html*

9. URL:*https://moluch.ru/archive/152/43250/* CLIL әдісіне негізделген әдістемелік нұсқаулық. Астана 2013ж.

10. URL:*http://adilet.zan.kz/kaz/docs/V1600013182*Жаратылыстану – математика, (математика, физика, информатика, химия, биология) пәндері бойынша Кіріктірілген оқыту бағдарламасы (ИОП) – шілде, 2013ж.

*ӘОК 004*
***[1]Алханов А.А., [2]Туребаева Р.Д., [3]Маткаримов Б.Т.***
*Л.Н. Гумилев атындағы Еуразия ұлттық университеті*
*Нұр-Сұлтан, Қазақстан*
*[1]alkhanov_aa_1@enu.kz, [2]58stud@mail.ru, [3]matkarimov_bt@enu.kz*

# АВТОМАТТАНДЫРЫЛҒАН БІР ДҰРЫС ЖАУАПТЫ ТАҢДАУ СҰРАҚТАРЫН ҚҰРУ ЖҮЙЕСІ ҮШІН БІЛІМ ҚОРЫНЫҢ МОДЕЛІН ЖАСАУ

**Аннотация.** Қазіргі уақытта білім берудегі бағалау оқу процесінің өте маңызды бір бөлігі болып табылады. Жақсы ұйымдастырылған бағалау мұғалімдерге оқушылардың үлгерімін бақылауға және оқушыларға сәйкес оқу жоспарын құруға көмектеседі. Дегенмен, жақсы бағалау көп уақыт пен процедураларды талап етеді. Тест тапсырмаларын құрастыру өте көп уақытты қажет ететін және жалықтыратын процесс. Оқушыларды дәл тексеру және оқудағы олқылықтарды табу үшін тест тапсырмалары жарамды және сенімді болуы керек, әйтпесе тестілеуден алынған нәтижелер сенімсіз болуы мүмкін және қате жоспарлауға және олқылықтарды одан әрі жоюға әкелуі мүмкін. Бұл жұмыс осыған байланысты жұмысқа шолу жасайды және тест сұрақтарын өңдеу және құру үшін қазақ, орыс және ағылшын тілдерінің онтологиялық моделін ұсынады. Іс жүзінде дұрыс жобаланған онтологиялық модель сарапшылардың мінез-құлқын имитациялайтын жүйелерді құруда үлкен рөл атқарады. Сондықтан, бұл жұмыстың мақсатының бірі онтологиялық модельдің мысалын көрсету болып табылады.

**Түйін сөздер:** бір дұрыс жауапты таңдау сұрақтары, онтология, бағалау, сұрақты автоматты түрде генерациялау, тіл онтологиясы, білім қоры, дистракторлар генерациясы.

*УДК 004*
***[1]Алханов А.А., [2]Туребаева Р.Д., [3]Маткаримов Б.Т.***
*Евразийский Национальный Университет имени Л.Н. Гумилева,*
*Нур-Султан, Казахстан*
*[1]alkhanov_aa_1@enu.kz, [2]58stud@mail.ru, [3]matkarimov_bt@enu.kz*

# СОЗДАНИЕ МОДЕЛИ БАЗЫ ЗНАНИЙ ДЛЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ГЕНЕРАЦИИ ВОПРОСОВ С НЕСКОЛЬКИМИ ВАРИАНТАМИ ОТВЕТОВ

**Аннотация**. В настоящее время оценивание в образовании является

очень важным аспектом процесса обучения. Хорошо организованное оценивание помогает учителям отслеживать прогресс своих учеников и составлять для них соответствующий учебный план. Тем не менее, качественное оценивание требует много времени, усилий, проведения множества процедур и соблюдения правил. Составление тестовых заданий — очень трудоемкий и утомительный процесс. Чтобы качественно протестировать учащихся и найти пробелы в обучении, тестовые задания должны быть валидными и надежными, иначе результаты, полученные в результате тестирования, могут быть недостоверными и привести к неправильному планированию работы. В этой статье представлен обзор соответствующей работы и предлагается онтологическая модель казахского, русского и английского языков для обработки и генерации тестовых вопросов. На практике правильно спроектированная онтологическая модель играет большую роль в построении систем, имитирующих поведение экспертов. Поэтому одной из целей этой работы является иллюстрация примера онтологической модели.

**Ключевые слова**: вопросы с несколькими вариантами ответов, онтология, оценивание, автоматическая генерация вопросов, языковая онтология, база знаний, генерация дистракторов.

*UDC 004*

***[1]Alkhanov A.A., [2]Turebaeva R.D., [3]Matkarimov B.T.***
*L.N. Gumilyov Eurasian National University*
*Nur-Sultan, Kazakhstan*
*[1]alkhanov_aa_1@enu.kz, [2]58stud@mail.ru, [3]matkarimov_bt@enu.kz*

## CREATING A KNOWLEDGE BASE MODEL FOR AN AUTOMATED MULTIPLE CHOICE QUESTIONS GENERATING SYSTEM

**Abstract.** Nowadays assessment in education is a very important aspect of the learning process. Well organised assessment helps teachers to track progress of their students and to build appropriate study plan for the learners. Nevertheless, quality assessment requires a lot of time and procedures to be undertaken. Compiling test items is a very time consuming and tedious process. In order to test students accurately and find gaps in learning, the test items have to be valid and reliable, otherwise the results gained from the testing might be unreliable and lead to wrong planning and further addressing the gaps. This paper provides an overview over related work and proposes an ontology model of Kazakh, Russian and English languages for processing

and generating test questions. Practically, a properly designed ontology model plays a great role in building systems that mimic experts` behavior. Therefore, one of the intensions of this work is to illustrate an example of ontology model.

**Keywords:** multiple choice questions, ontology, assessment, automatic question generation, language ontology, knowledge base, distractor generation.

**Кіріспе**

Жасанды интеллектті (ЖИ) қуаттандыру үшін екі негізгі әдіс бар. Әдістердің бірі ЖИ жүйелерін мысалдардан үйренуді қамтитын машиналық оқытуды пайдалану болып табылады. IBM компаниясының Уотсон ЖИ-і көрсеткендей, машиналық оқыту бүгінгі күннің өзекті тақырыбы болып табылады. Машиналық оқыту жүйелері үлгіге сәйкес келетінін көру үшін оқу жиындарынан үйренгендерін жаңа деректермен салыстыру үшін статистикалық үлгі санаттарын пайдаланады . Олар , мысалы, деректер үлгілерін зерттеу және оларды алаяқтықпен байланысты үлгілермен салыстыру арқылы ұсталуы мүмкін алаяқтықты болжау үшін жиі пайдаланылады . Машиналық оқыту (МО) қазіргі уақытта ақпараттық технологиялар әлеміндегі ең перспективалы қолданбалы салалардың бірі болып табылады, қолданбаларды қолдану аясы шексіз дерлік. Машиналық оқытуды білім беру саласында қолдану қазіргі уақытта зерттеушілер мен ғалымдар арасында өзекті тақырып болып табылады [7]. Көптеген зерттеулер машиналық бағалаудың адам мамандарының бағалауы сияқты дәл екендігін көрсетіп , осы жаңа технологияның орасан зор әлеуетін көрсетеді [12].

ЖИ қуатының көзі ретінде онтологияларды пайдалану тағы бір жол болып табылады. Онтология - бұл олардың ерекшеліктері мен қарым-қатынастарын көрсететін пәндік аймақтағы немесе домендегі ұғымдар мен категориялардың жиынтығы. Бұл әдіс ақпаратты кодификациялайтын білім қорын құрудан басталады, содан кейін әртүрлі деректер бөліктері арасындағы қарым-қатынастарды көрсететін онтологияны әзірлеу. Онтологияға негізделген ЖИ жүйеге мазмұн мен қарым-қатынастарға негізделген қорытынды жасауға мүмкіндік беру арқылы адам өнімділігін еліктейді. Функционалды болу үшін онтологияға негізделген ЖИ оқу жинақтарын пайдалануды қажет етпейді және өте бейімделген нәтижелерді бере алады. Онтологиялық модель ақпаратты жүйелеу және құрылымдау арқылы ақпараттық іздеуді ұйымдастыруға мүмкіндік береді [6].

**Тақырыпқа қатысты жасалған еңбектерге шолу**

Белгілі бір салада онтологияны құрастырып оның негізінде тест сұрақтарын, жауаптарын және дистракторларын құрастыруды бірқатар авторлар қарастырған. Олардың барлығы осы саланың әрі қарай дамуына үлесін қосқан.

Тест сұрақтарын автоматты түрде жасау әдісі [11] жұмысында сипатталды. Олардың ұсынылған әдістемесі доменді де, мультимедиялық білімді де білдіретін онтологияларға негізделген . Мақалада сонымен қатар OWL (Ontology Web Language) Semantic Web стандартты технологиясының көмегімен тест сұрақтар мен сәйкес сурет аннотацияларын жасайтын осы стратегияның прототипін іске асыру сипатталған. Сұрақ элементтері жалпы онтологияға негізделген әдістемелер арқылы жасалғандықтан, ұсынылған әдістеме бірегей домен сипаттамаларына тәуелсіз. Қарапайым табиғи тілді генерациялау әдістері ұсынылған прототипті іске асыруда тест сұрақтарындағы элементтерді жобалау үшін қолданылады.

Онтологияға негізделген бір дұрыс жауапты таңдау сұрақтарын (БДЖТС) генерациялау жүйелері БДЖТС шығарса да, бұл БДЖТС элементтерінің білім беру параметрлерінде қаншалықты қолайлы екенін көру үшін аз зерттеу жүргізілді. Стандарттарды белгілеу және онтологияға негізделген БДЖТС генерациялау жүйелерін жобалау мен әзірлеуге қойылатын талаптарды анықтау үшін мұндай бағалау қажет. Құрастырылған БДЖТС-ң тиімділігі мен олардың білім беру құндылығы тұрғысынан бағалау арқылы осы мәселені шешу әрекеті [1] жұмысында жасалды. Олар көп таңдаулы сұрақтарды құруға арналған онтологияға негізделген құрал OntoQue эксперименттік бағасын берді. OntoQue – домен онтологиясы негізінде автоматты түрде бағалау элементтерін шығару арқылы нұсқаушыларға көмектесетін сұрақ жасау жүйесі. Көрсетілген онтологиялық доменнен OntoQue қозғалтқышы бағалау элементтерінің жинағын жасайды. БДЖТС түрі , true/false (T/F) және толтыру (T) элементтері опциялар арасында. OntoQue қозғалтқышы онтология нысандарын (өмірге жарамды талаптардың комбинациясы) итерациялау арқылы сұрақтарды жасайды. Олар негізінен осы сұрақтар түрлерін талқылады, өйткені олардың зерттеу мақсаты жасалған БДЖТС элементтерінің сапасын бағалау болды.

Келесі [2] жобасында онтологияға негізделген БДЖТС құруды бағалау үшін сәйкес білім беру теориясын және оларды бағалаудың тиісті әдістемелерін қолдану үшін пайдаланады. Олар онтологиядағы идеялардың ұқсастығын сенімді түрде бағалай алатынын анықтауға ерекше қызығушылық танытады, осылайша бұл шара жасалған БДЖТС қиындығын бақылау үшін пайдаланылуы мүмкін . Олар баяндамада

зерттеудің қысқаша мазмұнын, сондай-ақ олар жасаған негізгі қадамдарды және олар жасаған қорытындыларды береді. Олар өздерінің алгоритмін бағалау үшін автоматты «шешу» тәсілін қолданды, ол «дистракторды таңдау алғышартына» негізделген . Дегенмен, кіріс онтологиялары өте күрделі немесе өте қарапайым болғандықтан, шынайы бағалау мүмкін болмады.

OWL-DL онтологияларының ауқымын шынайы бағалауда қолдануға болатын бірнеше/көп таңдау сұрақтарын (БДЖТС) генерациялау тұрғысынан зерттеді, сонымен қатар Элементтерге жауап беру теориясы (ЭЖБТ) принциптерін пайдалана отырып жасалған бағалау элементтерінің тиімділігіне [5] жұмысында егжей-тегжейлі зерттеу жүргізілді. Олар егжей-тегжейлі автоматты тест сұрақ жасау (ATG) жүйесін және оның кеңейтілген нұсқасын, Extended-ATG жүйесін сипаттады. ATG жүйесі (бірінші жүйе) формальды онтологиядан белгілі бір өлшемдегі бірнеше таңдаулы сұрақтар жинағын шығару үшін пайдалы болды. Ол доменге қатысты бағалау жүргізуге қажетті сұрақтарды ғана таңдау үшін эвристика жинағын пайдалану арқылы жұмыс істейді. Extended-ATG жүйесін қалыптастыру үшін олар осы жүйеге жасалған БДЖТС қиындықтарының мәндерін анықтау және сұрақтар жиынының жалпы қиындық деңгейін реттеу (жаңа жүйе) сияқты жаңа функцияларды қосты. Осы жаңа қасиеттерді шешу үшін қолданылатын жаңа стратегиялар олардың зерттеулерінде талқыланады. Яғни, сұрақтың қиындық деңгейін анықтау әдісі және сұрақтар жиынтығының қиындығын бақылау алгоритмі. Авторлар сонымен қатар олардың ATG және E-ATG жүйелері арасындағы айырмашылықтарды түсіндірді. Олар ATG жүйесі БДЖТС сабақтарын генерациялау үшін тек екі предикатты пайдаланғанымен, E-ATG жүйесінде мұндай шектеулер жоқ және предикатқа негізделген сабақтарды өндірудің әртүрлі қызықты үлгілерін пайдаланады деп мәлімдейді. Бұл предикат үлгілері үлкен нақты әлемдегі сұрау жиындарын мұқият эмпирикалық зерттеу арқылы табылды. Бұған қоса, олар жаңа жүйеде артықшылыққа қатысты сұрауларды (мысалы, ең биік тау, ең үлкен өзен және т.б.)

Олар өз жұмыстарында әртүрлі онлайн онтологиялардан БДЖТС құру арқылы ұсынылған әдістердің орындылығы мен пайдалылығын тексерді. Алынған сұрақтарды домен сарапшылары ұсынған сұрақтармен салыстыру арқылы олар ұсынылған сұрақтарды таңдау эвристикасының тиімділігін зерттеді. Ұсынылған жүйе бойынша есептелген сұрақтардың қиындық ұпайлары сыныптағы эксперименттер деректері бойынша ЭЖБТ көмегімен анықталған олардың нақты қиындық деңгейімен айтарлықтай корреляцияланғаны анықталды.

Олардың нәтижелері E-ATG жүйесі адамдар жасаған сұрақтармен салыстыруға болатын (семантикалық ұқсастығы бойынша) доменге қатысты сұрақтар жиынтығын жасай алатынын көрсетті . Бұл әдіс нақты педагогикалық мақсаттарға қол жеткізу үшін автоматты түрде жасалған сұрақтар жиынының жалпы қиындық деңгейін басқару үшін де пайдаланылуы мүмкін . Олардың болашақтағы мақсаты – автономды түрде құрастырылған сұрақтардың психометриялық қасиеттерін (сенімділік пен негізділік сияқты) зерттеу үшін нақты жағдайдағы ауқымды эксперимент жүргізу.

Білім қорынан емтихан сұрақтары мен жауаптарын автоматты түрде құру процесін әзірлеу [9] еңбегінде жазылған. Олардың жұмысының өнімі смарт оқытуды ұйымдастыру мен жоспарлауға арналған компьютерлік программа болды. Олар RDF-құжаттарындағы онтологияға негізделген ұсыну үлгілері мен SPARQL сұрауды бағалаудың егжей-тегжейлі мысалдарын берді. Сонымен қатар, олар RDF-файлдарымен өзара әрекеттесу үшін сәйкес кітапханалары бар Python көмегімен іске асырылған сұрақ құрастыратын компьютерлік программаны ұсынды.

Білім қорынан тест сұрақтарын автоматты түрде жасау жүйесін әзірлеу [10] мақаласында сипатталған. Олардың жұмысы пайдалы болып табылады, өйткені онда RDF мәтіндерінде онтологияны құрудың және SPARQL сұрауын іске асырудың егжей-тегжейлі мысалдары қамтылған. Сондай-ақ ол Python бағдарламалау тілінің сұрақтарды құрастыратын программаны іске асыруын, сондай-ақ RDF файлдарымен жұмыс істеуге арналған маңызды кітапханаларды егжей-тегжейлі сипаттайды.

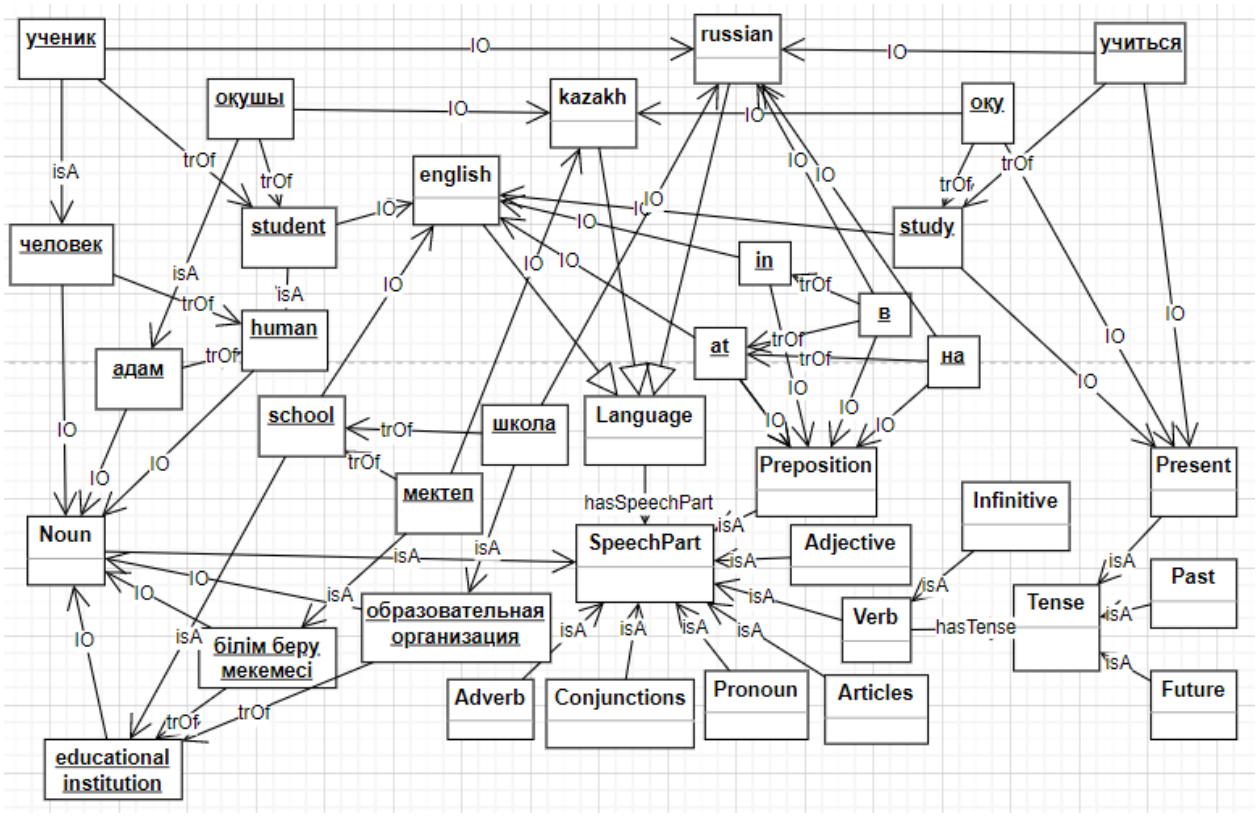
Блум таксономиясында инкапсуляцияланған дұрыс педагогикалық принциптерді пайдалана отырып, кез келген домен онтологиясынан тест сұрақтарын жасайтын авторлар әзірлеген құралдар сипатталып, [3] мақаласында бағаланды. Зерттеу жұмыс құжатта дизайн ғылымын негіз ретінде пайдалана отырып ұсынылған. Әр түрлі 90 ерекше онтологиядан барлығы 5230 сұрақ дайындалды және 8 арнайы сарапшы кездейсоқ таңдалған 81 сұрақты бағалады. Параметрлік емес дисперсияны талдау үшін сипаттамалық статистика және Крускал - Уоллис тесті қолданылды. Жалпы, жасалған сұрақтардың 69 пайызы тестілеуге жарамды деп бағаланды, ал 33 пайызы орташа және жоғары күрделілікте болды. Онтологиялар, дистракторларды шығару тактикасы және Блум сұрақтарының деңгейлері бойынша құрылған сұрақтардың сапасындағы елеулі айырмашылықтар анықталды: ең жоғары сапалы сұрақтар білімнің қолданылуын тексеретін және семантикалық стратегияларды пайдаланатын сұрақтар болды.

Көп таңдаулы тест тапсырмаларын жасаудың автоматтандырылған тәсілі [4] мақалада сипатталған. Әдіс тақырып-сөздерге немесе түйінді сөздерге (бір немесе бірнеше сөз) негізделген нұсқаулық сөйлемдерді таңдаудан басталады. Жауап кілті мәтіндегі ең жақсы кілт сөзден анықталады. Осыдан кейін жүйе осы мәлімдемеден жауап кілтін алып тастап, оны сұраққа (баған) түрлендіреді. Кез келген сыртқы ақпаратты немесе білім қорын қажет етпей, қате опциялар немесе дистракторлар мүмкіндіктерге негізделген кластерлік тәсіл арқылы автоматты түрде жасалады. Ұсынылған жұмыс негізінен Табиғи Тілді Өңдеу (ТТӨ) және Машиналық оқыту (МО) негізіндегі әдістерге негізделген.

Ұқсастықты есептеу үшін онтологияларды қолданатын және оларды машиналық оқыту әдістеріне енгізетін әдістерге [8] жұмысында шолу жасалды; олар семантикалық ұқсастық өлшемдері мен онтологиялық ендірулер онтологиялардың фондық білімдерін қалай пайдалана алатынын және онтологиялардың машиналық оқыту үлгілерін жақсартуға көмектесетін шектеулерді қалай қамтамасыз ете алатынын көрсетті. Жұмыс тест сұрақтарын құру туралы болмаса да, онтологияны қолдану туралы пайдалы ақпарат бар.

**Білім қорының моделін жасау**

1-суретте көрсетілген диаграмма үш тілдің, қазақ, орыс және ағылшын тілдерінің онтологиялық моделінің шағын бір бөлігі болып табылады. Онтология кластардан, ішкі класстардан, объектілерден және қатынастардан тұрады. Класстар мен ішкі класстар, ортасында көлденең сызығы бар блоктар болып табылады, ал объектілер асты сызылған мәтіні бар блоктар. Түсіндіруді қажет ететін бірнеше қысқартылған қатынастар түрлері бар, атап айтқанда олар: *IO – экземпляры/объектісі*, *trOf – аудармасы*, *isA – ол/болып табылады*. Ұсынылған онтология моделі тест сұрақтарының кез келген түрін құрастыруға және интеллектуалды қорытындылар жасауға ыңғайлы болып табылады. Бұл білім қорын граф түрінде көрсетуге болады. Граф екі өлшемді динамикалық массив немесе вектор көмегімен де жүзеге асырылуы мүмкін.




1-сурет

Суретте көрсетілген онтологияны қолданып, дұрыс жауапқа жақын келетін, яғни өте сапалы дистракторларды құрастыруға болады. Жоғарыда көрсетілген онтология тілдерді үйренуге арналған тест сұрақтарын құрастыруға ыңғайлы болып табылады, одан бөлек оны қазақ тілінен басқа тілдерге компьютерлік аудармашы программасын жасау үшін қолдануға болады. Онтология моделін қолданып төмендегідей түрлі тест сұрақтарын құрастыруға болады:

1) «мектеп» сөзінің ағылшын тіліндегі аудармасы қандай ?
A. student
B. human
C. school
D. study
2) «оқу» сөзі қазақ тілінің қай сөз табына жатады ?
A. сан есім
B. сын есім
C. зат есім
D. етістік
3) Who is a student?
A. educational institution
B. human

C. school
D. study

Бірінші тест сұрағы қазақша сөздерінің ағылшынша аудармасын тексеру үшін құрастырылған, екінші тест сұрағы қазақ тіліндегі сөйлемдерінің құрылымын тексеруге арналған, ал үшінші сұрақ ағылшын тілінде оқушының кім екенін анықтау үшін құрастырылған, демек объектінің қай топқа немесе классқа жататынын тексеру үшін құрастырылған тест сұрағы. Теория жүзінде осындай онтология моделі әртүрлі білім салаларына қатысты сұрақтарды құрастыруға және сонымен қатар автоматтандырылған түрде интеллектуалды қорытынды жасауға мүмкіндік бере алады. Жоғары аталған артықшылықтарға қоса осы онтология моделі тест сұрақтарын, олардың дұрыс жауаптарын және дистракторларын қазақ, орыс және ағылшын тілдерінде генерациалауға мүмкіндік береді.

**Қорытынды**

Бұл мақала тақырып бойынша кейбір байланысты жұмыстарға шолу жасайды және онтология редакторының көмегімен жүзеге асырылуы мүмкін және автоматтандырылған тест сұрақтарын құру үшін пайдаланылуы мүмкін үш тілге арналған онтология үлгісін көрсетеді. Ұсынылған онтологиялық модель БДЖТС үшін сенімді дистракторларды жасауға мүмкіндік береді және нақты тілдің сөздік қорын бағалау үшін тест сұрақтарын құру үшін пайдаланылуы мүмкін. Дегенмен, мәліметтер қорын автоматтандырылған жинақтау әдістері мүмкін болса да, әрбір тіл үшін сөздердің көптігін ескере отырып, мәліметтер қорын толтыруда қиындықтар туындауы мүмкін. Сонымен қатар, әзірше сөз жалғаулары мен мезгіл шақтарын тест сұрақтарын құрастырған кезде ескеру шешілмеген мәселе болып тұр.

**Әдебиеттер тізімі**

1. Al-Yahya, M. (2014). Ontology-based multiple choice question generation. *The Scientific World Journal, 2014*.

2. Alsubait, T., Parsia, B., & Sattler, U. (2016). Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz, 30*(2), 183-188.

3. Cubric, M., & Tosic, M. (2020). Design and evaluation of an ontology-based tool for generating multiple-choice questions. *Interactive Technology and Smart Education*.

4. Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment. *Multimedia Tools and Applications*, *80*(21), 31907-31925.

5. EV, V., & Kumar, P. S. (2017). Automated generation of assessment tests from domain ontologies. *Semantic Web*, *8*(6), 1023-1047.

6. Globa, L., Novogrudska, R., Koval, A., & Senchenko, V. (2018). Examples of ontology model usage in engineering fields. *Ontology in Information Science*.

7. Kučak, D., Juričić, V., & Đambić, G. (2018). MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS. *Annals of DAAAM & Proceedings*, *29*.

8. Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. *Briefings in bioinformatics*, *22*(4), bbaa199.

9. Omarbekova, A. S., Nurgazinova, G. S., Sharipbay, A. A., Barlybayev, A., & Bekmanova, G. T. (2017, August). Automatic formation of questions and answers on the basis of the knowledge base. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1-3). IEEE.

10. Omarbekova, A., Sharipbay, A., & Barlybaev, A. (2017, February). Generation of test questions from RDF files using PYTHON and SPARQL. In *Journal of Physics: Conference Series* (Vol. 806, No. 1, p. 012009). IOP Publishing.

11. Papasalouros, A., Kotis, K., & Kanaris, K. (2011). Automatic generation of tests from domain and multimedia ontologies. *Interactive Learning Environments*, *19*(1), 5-23.

12. Zhai, X. (2019). Applying machine learning in science assessment: opportunity and challenge. *Journal of Science Education and Technology*.

# COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH» СЕМИНАРЫНЫҢ МАТЕРИАЛДАРЫ

# МАТЕРИАЛЫ СЕМИНАРА «COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH»

# MATERIALS OF THE SEMINAR «COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH

В данный раздел включены аннотации статей, которые докладывались на семинаре «COMPUTATIONAL MODELS IN TURKIC LANGUAGE AND SPEECH». Полный текст данных статей будут опубликованы в отдельном сборнике конференции.

*УДК 81.32*

***[1]Гатиатуллин А.Р., [2]Прокопьев Н.А., [3]Сулейманов Д.Ш.***
*[1, 2, 3] Институт прикладной семиотики*
*Академии Наук Республики Татарстан,*
*[1, 2]Казанский Федеральный Университет*
*Казань, Татарстан, Россия*
*[1]ayrat.gatiatulin@gmail.com, [2]nikolai.prokopyev@gmail.com,*
*[3]dvdt.slt@gmail.com*

## К ВОПРОСУ СОЗДАНИЯ ЕДИНОГО КОНЦЕПТУАЛЬНОГО ПРОСТРАНСТВА ТЮРКСКИХ ЭЛЕКТРОННЫХ КОРПУСОВ С ПОМОЩЬЮ ИНТЕРНЕТ-ПОРТАЛА «ТЮРКСКАЯ МОРФЕМА»

**Аннотация.** Сфера создания электронных корпусов естественных языков как инструмента их сохранения и развития, а также как ресурсной базы для разработчиков NLP-технологий и исследователей языков испытывает быстрый рост количества и объема электронных корпусов для множества языков, в том числе и для тюркских языков. Вместе с тем для множества тюркских языков корпусы отсутствуют в силу наличия у разработчиков проблем с технической реализацией и поддержкой функционирования данных больших и требовательных к техническому обеспечению ресурсов. В то же время, существующие корпусы тюркских языков характеризуются проблемами унификации и взаимной интеграции, ведущими к сложностям в проведении компаративистских исследований, а также в разработке параллельных корпусов, многоязычных систем обработки текстов и в решении теоретических и прикладных задач.

В данной статье представлен инструментарий для интеграции тюркских корпусов с веб-порталом «Тюркская морфема», многоязычным ресурсом, содержащим как языконезависимые грамматические, синтаксические и семантические модели, так и специфичные для языков тюркской семьи элементы. Рассмотрены два варианта интеграции корпусов: создание нового корпуса в рамках портала и связывание существующего корпуса с порталом. Применение данного инструментария позволит решить проблемы унификации корпусной разметки и разработки программного обеспечения для обработки естественного языка, обогатить ресурсы портала за счет создания единого концептуального пространства тюркских электронных корпусов.

**Ключевые слова:** электронный корпус, концептуальное пространство, лингвистический ресурс, тюркология.

*УДК 81'33, 004.04*

***[1]Рахимова Д., [2]Турганбаева А., [3]Абдразах Ы.***
*[1,2,3]Казахский национальный университет им. Аль-Фараби,*
*[1,2]Институт информационно-вычислительных технологий*
*Алматы, Казахстан*
*[1]di.diva@mail.ru*

# О ПРОБЛЕМЕ ВЫЯВЛЕНИЯ МОРФОЛОГИЧЕСКИХ ОШИБОК В КАЗАХСКОМ ЯЗЫКЕ

**Аннотация.** В настоящее время объем электронной информации увеличивается. Различные интеллектуальные прикладные системы обработки естественного языка очень быстро и активно вошли в потребность рядовых жителей, например, такие как: Информационно-поисковые системы, диалоговые системы, средства машинного перевода и автоматического возобновления, средства проверки орфографии для анализа и обработки текстов на естественных языках. Для реализации и работы той или иной системы необходимо выполнять предобработку данных. Один из них — проверка слов на правильность, морфологический анализ. В данной статье представлен анализ ошибок, часто используемых пользователями, составлена классификация по морфологическим ошибкам. Для решения проблемы представлен подход к выявлению некорректных слов в казахском языке. Разработан подход для выявления некорректных слов из полу структурированных данных, основанный на алгоритме стемминга с основами лексики по морфологической модели CSE (Complete Set of Endings). Анализируя

методы и системы выявления ошибок в тексте, были разработаны модели выявления неверных слов казахского языка, составлен словарь правильных слов казахского языка. Корпус собран на казахском языке с использованием программного кода, созданного с учетом особенностей и особенностей казахского языка . Эксперимент проводился на данных из социальных сетей и новостных веб-порталов. Выявление некорректных слов и последующее их исправление повысит статус информации на вашем сайте, поможет вашему бизнесу расти, собрать целевую аудиторию, а также обработает необходимую информацию, анализируя комментарии пользователей, и поисковые системы найдут ключевые слова на вашем сайте. В целом по результатам эксперимента точность определения некорректных слов составила более 90%. Основной вклад в эту работу вносит автоматизированный подбор корпуса для казахского языка. Разработка подхода к выявлению некорректных слов казахского языка с учетом лингвистических свойств языка и создание лингвистических ресурсов и программ для сбора данных в полу структурированных данных.

**Ключевые слова:** Казахский язык, морфологические ошибки, неправильные слова, алгоритм стемминга, полный набор окончаний.

*УДК 811`322:*

***[1]Абдурахмонова Н., [2]Исмаилов А.Ш.***
*[1]Национальный университет Узбекистана, Ташкент, Узбекистан*
*[2]Андижанский машиностроительный институт,*
*Андижан, Узбекистан*
*[1]abdurahmonova.1987@mail.ru, [2]alisherismailov1991@gmail.com*

# РАЗРАБОТКА УЗБЕКСКОГО СТЕММЕРА ДЛЯ УЗБЕКСКОГО ЯЗЫКА

**Аннотация.** В этой статье проводится сравнительное исследование различных видов алгоритмов стемминга. Чтобы получить хороший результат для стеммера, были исследованы лингвистические модели в соответствии с особенностями языка, а именно его агглютинативностью.

За последние несколько десятилетий объем база данных, производимых по всему миру, значительно вырос. Первый рост данных начался около пятидесяти лет назад, когда исследователи столкнулись с первыми научными публикациями из самых разных областей, позже это стало социализацией Интернета. Исследователям нужны были некоторые механизмы для поиска информации из этих объемных

данных. В настоящее время эти механизмы составляют большую часть процесса поиска информации. Эти механизмы представляют собой полную, обширную и специализированную область исследований в области информационных технологий. Основная цель информационного поиска состоит в систематическом анализе данных и извлечении некоторых связанных данных или документов, которые необходимы пользователю или требуемой информации. Потребность в информации означает, что механизм отвечает не только на прямые вопросы (например, что такое основа?), но также ищет документы, относящиеся к введенному термину или группе терминов, которые могут быть конкретными (например, основа) или не быть конкретными. В данной ситуации пользователь не уверен, что ищет алгоритм поиска корней слов. Эти конкретные вопросы, термины и слова также известны как запросы, и они также будет представлять собой ввод пользователя в систему. Результатом может быть заголовок документа или группа заголовков документов, и их можно ранжировать в соответствии с процентом соответствия между документами и запросом в процессе расчета.

**Ключевые слова:** стемминг, поиск информации, обработка естественного языка

*УДК 378.147 + 519.767.6*

***[1]Прокопьев Н.А., [2]Сулейманов Д.Ш.***
*[1, 2] Институт прикладной семиотики*
*Академии Наук Республики Татарстан,*
*[1]Казанский Федеральный Университет*
*Казань, Татарстан, Россия*
*[1]nikolai.prokopyev@gmail.com, [2]dvdt.slt@gmail.com*

# СИСТЕМА КОНТРОЛЯ ЗНАНИЙ НА ОСНОВЕ ПРАГМАТИЧЕСКИ ОРИЕНТИРОВАННОГО АЛГОРИТМА ОБРАБОТКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ВОПРОСНО-ОТВЕТНЫХ ТЕКСТОВ

**Аннотация.** Автоматизированный контроль знаний является важным компонентом систем электронного образования. Более того, как показывает анализ последних публикаций, в современных системах электронного образования практически не реализована возможность автоматизированного контроля ответа, свободно формулируемого обучаемым на естественном языке. Реализация такого алгоритма позволит осуществлять более эффективное управление процессом

образования в режиме онлайн за счет дифференциации учебного контента и индивидуализации обучения на основе более тонкого автоматизированного анализа и оценки знаний обучаемого.

В статье представлена реализация двухуровневого алгоритма обработки естественно-языковых вопросно-ответных текстов на основе прагматически-ориентированного подхода. Сформулирован метод итеративной экспериментальной верификации данного алгоритма. Реализован прототип системы контроля знаний, использующей прагматически-ориентированные модели и методы для обработки ответов. Реализован модуль для экспериментальной верификации, позволяющий проводить анализ результатов применения алгоритма к реальным ответам обучаемых и оценку его эффективности.

**Ключевые слова:** контроль знаний, обработка естественного языка, прагматически-ориентированный метод, электронное образование, электронная оценка

*УДК 81'33, 004.04*

***[1]Рахимова Д., [2]Сулейменов Е.,[3]Ахмет Г.***
*[1,3]Казахский национальный университет им. Аль-Фараби,*
*[1,2]Институт информационно-вычислительных технологий,*
*Алматы, Казахстан*
*[1]di.diva@mail.ru*

## ЗАДАЧА СИНТЕЗА КАЗАХСКОГО ЯЗЫКА НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

**Аннотация**. В настоящее время многие тексты синтезируются с помощью компьютера. Синтез предложений становится все более распространенным в различных областях. Например, он используется в интеллектуальных системах, способных объяснить пользователю ход решения конкретной задачи, системах поддержки принятия решений, помогающих пользователю принимать решения на основе разработанной альтернативы, умных домах, голосовых помощниках, чат-ботах и так далее. Машинное обучение — один из самых эффективных подходов к решению задачи синтеза текста. Алгоритмы машинного обучения могут определять, как выполнять важные задачи, анализируя примеры. В задаче синтеза предложений с помощью машинного обучения можно заменить ряд компонентов всей системы нейронными сетями, что позволяет не только приблизиться по качеству к существующим алгоритмам, но даже значительно их превзойти. В статье исследуются аналоговые системы синтеза предложений,

рассматривается проблема синтеза предложений в казахском языке. Ряд параллельных блоков вопросов и ответов на английском и русском языках были переведены на казахский язык и проверены. Таким образом, был собран параллельный корпус на казахском языке, состоящий из 60000 вопросов и ответов. Задача синтеза предложений решалась на базе системы чат-ботов методом seq2seq. На основе выбранного метода и собранного корпуса был проведен ряд экспериментов и получены результаты.

Исследование выполнено и профинансировано грантом Проекта ИРН AP 09259556 Министерства образования и науки Республики Казахстан.

**Ключевые слова:** казахский язык, синтез предложений, машинное обучение, метод seq2seq, NMT, RNN.

*УДК 004.934*

***[1]Кожирбаев Ж., [2]Есенбаев Ж., [3]Исламгожаев Т.***
*National Laboratory Astana, Назарбаев Университет,*
*Нур-Султан, Казахстан*
*[1]zhanibek.kozhirbayev@nu.edu.kz, [2]zhyessenbayev@nu.edu.kz,*
*[3]talgat.islamgozhayev@nu.edu.kz*

## ПЕРЕВОД РЕЧИ В РЕЧЬ ДЛЯ КАЗАХСКОГО ЯЗЫКА: ОБЗОР

**Аннотация**. Мы представляем здесь нашу работу по обзору перевода речи в речь для казахского языка. Эту задачу перевода можно решить, применяя каскадный метод. Система каскадного метода состоит из таких модулей, как автоматическое распознавание речи, машинный перевод и синтез речи. Второй подход использует сквозной метод, который переводит речь с одного языка на речь на другом языке, не полагаясь на генерацию промежуточного текста.

Мы исследовали работу, связанную с модулями каскадного решения для казахского языка. Последние достижения, основанные на прямом речевом переводе для других языков, также кратко освещены в этой статье. Недавний прогресс в самоконтролируемом обучении речи и модели многоязычного кодера-декодера демонстрирует впечатляющую производительность в задачах перевода речи с небольшим набором данных. Данное исследование финансируется Комитетом науки Министерства образования и науки Республики Казахстан (грант № AP13068635, № AP08053085 и №BR11765535).

**Ключевые слова:** перевод речи, сквозной перевод речи, каскадный перевод речи, предварительно обученные модели.

*УДК 004.912; 004.62*
***[1]Кадырбек Н., [2]Мансурова М., [3]Момбеков М., [4]Кыргызбаева М.***
*Казахский национальный университет им. Аль-Фараби,*
*Алматы, Казахстан*
*[1]nurgaliqadyrbek@gmail.com, [2]mansurova.madina@gmail.com*
*[4]marzhan.kyrgyzbaeva@gmail.com*

# АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ПРЕДИКАТНО-АРГУМЕНТНОЙ СТРУКТУРЫ (PAS) НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИ

**Аннотация.** Данная работа посвящена разработке инструмента для автоматического извлечения предикативно-аргументной структуры из текстов на казахском языке. В работе освещаются вопросы, которые касаются структуры казахского языка, сбора текстовых данных и их моделирования. Мы представляем новую схему анализа структуры аргументов предикатов, которая может быть обучена на основе частично аннотированных корпусов. Эта семантическая маркировка требует применения хороших моделей векторизации, таких как GloVe и Word2Vec. Преимущество частичного аннотирования заключается в том, что оно улучшает адаптивность к различным предметным областям. В ходе работы устанавливается базовый список семантических аргументов предиката с незначительным дополнением с целью расширения практической значимости инструмента. Аргументы выражают причинно-следственную перспективу предложения. В рамках работы исследована семантическая модель казахского языка, разработаны модель глубокого обучения для автоматической маркировки неструктурированных текстов и инструмент разметки данных.

**Ключевые слова:** семантический анализ, предикатно-аргументная структура, Обработка естественного языка, машинное обучение, BERT.

*УДК 81'33, 004.04*
***[1]Ahmet Toprak, [2]Metin Turan***
*İstanbul Commerce University, Стамбул, Турция*
*[1]ahmetoprak190363@gmail.com, [2] turanmetin@gmail.com*

# СОЗДАНИЕ КОРПУСА АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТУРЕЦКОГО ЯЗЫКА С ПОМОЩЬЮ NLP

**Аннотация**. Сегодня информация производится и потребляется очень быстро. Особенно за последние 15 лет Интернет стал частью повседневной жизни, количество интернет-информации и ресурсов превзошло все ожидания. В среде, где данные настолько плотны и разбросаны, становится чрезвычайно важным эффективно извлекать значимую информацию. Получение необходимой информации из таких больших данных — сложный и сложный процесс для человека. Существует множество методов борьбы с этим процессом. Они оцениваются в наборах данных, доступных в литературе. Автоматические системы реерирования текстов являются одним из таких методов, которые упрощают утомительную работу и экономят время, а также облегчают получение необходимых документов.

Однако для турецких текстовых систем реферирования недостаточно данных. В этом исследовании корпус, который включает в себя как полный текст, так и резюме, написанные на турецком языке для использования в системах реферирования текста, автоматически генерируется с помощью веб-поиска на основе методов обработки естественного языка. Резюме полных текстов в корпусе были созданы с использованием заголовков новостных сайтов, взятых из Интернета. Чтобы измерить способность резюме в наборе данных представлять полнотекстовые документы, было рассчитано сходство между резюме и полным текстом, и средний уровень сходства текста Wordnet для созданного набора данных составил 39,6%. Это исследование принесет большую пользу литературе с точки зрения исследования турецкого языка, где ощущается нехватка наборов данных.

**Ключевые слова:** Сводный корпус турецкого языка, Обобщение текста, Веб-поиск, Wordnet, NLP.

*УДК 004.91*

***[1]Макипова Э., [2]Ахметов И.***
*[1]Казахстанский Институт Менеджемента Экономики и Прогнозирования*
*[1,2] Институт Информационных и Вычислительных Технологий*
*Алматы, Казахстан*
*[2]iskander.akhmetov@gmail.com*

## ВОСКРЕШЕНИЕ: РЕКОНСТРУКЦИЯ ХАЗАРСКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

**Аннотация.** Расшифровка или реконструкция мертвых языков является сложной задачей, особенно когда цель состоит в том, чтобы

реконструировать язык, в котором не осталось совсем или осталось очень мало текстов, например, хазарский язык или раннеславянские и угорские языки. В этой статье мы предлагаем историческую перспективу хазарского народа, его языка и современных потомков, а именно чувашей и татар. Затем мы поговорим о том, как информатика может помочь исследователям в реконструкции и расшифровке языка. Наконец, мы тестируем подход к поиску хазарских/булгарских слов-кандидатов в чувашском и татарском языках путем (1) нормализации слов двух языков и (2) их сравнения с учетом семантических понятий для решения проблемы омонимии и (3) исключения общетюркских слов и заимствований из русского языка.

*УДК 004.932.72*

***[1]Амангельды Н., [2]Разахова Б.Ш., [3]Турсынова Н.А.,***
***[4]Кадыр Е., [5]Сарбупеев Н.А.***
*[1,2,3,5] Евразийский национальный университет им. Л. Н. Гумилева,*
*[4]Astana IT University, Нур-Султан, Казахстан*
*[1]nurzadaamangeldy@gmail.com,[2]utalina@mail.ru,*
*[3]ntursynova000@gmail.com,[4]yenglik.kadyr02@gmail.com*

## СЕГМЕНТАЦИЯ И ОПРЕДЕЛЕНИЕ КЛАССА РАКА КОЖИ (ЗЛОКАЧЕСТВЕННЫЙ ИЛИ НЕЗЛОКАЧЕСТВЕННЫЙ) НА ОСНОВЕ МОДЕЛИ MASK-RCNN

**Аннотация.** В этой статье представлено решение для определения класса рака кожи, доброкачественного или злокачественного, на основе RCNN. Это позволит врачам-онкологам быстро определить вид опухоли и предотвратить заболевание, тем самым сэкономив время. Изучена архитектура и уточнена эффективная модель исскуственной нейронной сети. В этом проекте нейронная сеть была самообучена через библиотеки команд Detectron2. Нейронная сеть была построена путем чтения конфигураций и разделения их на каждую управляющую часть, обеспечивающую работу программы. В модель успешно внедрились конфигурации Mask-RCNN и достигла хорошего результата 95-97%.

**Ключевые слова:** Нейронная сеть, искусственный интеллект, CNN, модель, рекуррентный нейронный сеть, злокачественный, доброкачественный

*УДК 004.94*

***[1]Кудубаева С. А., [2]Нургазина Д. М.***
*Евразийский Национальный университет им. Л.Н.Гумилева,*

*Нур-Султан, Казахстан*
*[1]saule.kudubayeva@gmail.com, [2]dana_nurgazina@inbox.ru*

## СЕМАНТИЧЕСКИЙ РАЗБОР ЖЕСТОВОГО ЯЗЫКА ПРИ ПЕРЕВОДЕ С ТЕКСТА НА КЖЯ

**Аннотация.** В статье затрагивается тема семантики казахского жестового языка и ее особенности. Ведь семантика – это раздел лингвистики, который изучает смысл выражения и сочетания знаков. Автор затрагивает вопрос насколько похожи словесный язык с жестовым языком и как они отличаются друг от друга. Потому что как всем известно, слова появились, когда древние люди показывая какие-то предметы имитировали звуки. Основное содержание исследования составляет анализ свойств частей речи казахского языка в жестовом языке, правила перевода глаголов и наречия времени. По причине того, что в жестовом языке все слова используются в начальной форме, то есть в инфинитиве. Возникает необходимость знания порядка расположения слов в предложении, так как язык жестов является отдельной наукой для исследования. К тому же приводятся данные, проливающие свет на части речевой системы казахского жестового языка (КЖЯ), в частности на классификации имени существительного и глагола. Жестовый язык – это многомодальная средство коммуникации, в котором играют важную роль корпус, голова говорящего, жестикуляция и выражение лица. В этой работе подробно приведены формулы перевода с текста на жестовый язык, а на основу берется субъект-объект-предикат. Но стоит правильно понимать, что все таки необходимо в первую очередь знать смысл предложении, здесь как и в словесном языке существуют очень много исключении, которые еще требуют дальнейшего исследования. А также примеры показывающие взаимосвязь субъектно-объектные отношения - это обозначение кем является говорящий, «дополнением» или «подлежащим» высказывания.

**Ключевые слова:** казахский жестовый язык, семантика, сурдоперевод, части речи.

*УДК 519.95*

***[1]Игнатьев Н., [2]Тулиев У.***
*[1,2] Национальный университет Узбекистана, Ташкент, Узбекистан*
*[1]n_ignatev@rambler.ru, [2]mirzobek.tuliyev@gmail.com*

## АНАЛИЗ СЕМАНТИЧЕСКОЙ СВЯЗАННОСТИ СЛОВ НА ОСНОВЕ ОБЩЕГО СЛОВАРЯ

**Аннотация.** Рассматривается методика создания наборов (групп) из слов (понятий) естественного языка по данным из текстовых документов. Считается, что множество документов разделено на два непересекающихся класса, для описания которых используется общий словарь. Наборы слов применяются для снижения размерности исходного пространства через формирование по ним латентных признаков. Выбор латентных признаков обосновывается по значениям контентной аутентичности документов. На основе ранжирования признаков по результатам разбиения документов на кластеры производится отбор семантически связанных слов по темам.

**Ключевые слова:** Тематическое моделирование, латентные признаки, онтология, общий словарь

ISBN 978-601-326-645-9

«TURKLANG 2022»
«Түркі тілдерін компьютерлік өңдеу»
атты Х халықаралық конференция
ЕҢБЕКТЕРІ
16-18 маусым 2022 ж.

ТРУДЫ
X Международной конференции
«Компьютерная обработка тюркских языков»
«TURKLANG 2022»
16-18 июня 2022 г.

PROCEEDINGS
of the X International Conference
on Computer processing of Turkic Languages
«TURKLANG 2022»
16-18 June 2022

Басуға 13.06.2022 жылы қол қойылды
Шартты баспа табағы 21,5 Пішімі $60/84_{1/16}$
Таралымы 20 дана
Тапсырыс 0105
«Булатов А.Ж.» жеке кәсіпкер
010000, Нұр-Сұлтан қ.,
Керей, Жәнібек хандар к-сі, 22-38
Тел.: 8/7172/223-418
e-mail: masterpo08@mail.ru