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

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

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

«Түркі тілдерін компьютерлік өңдеу»
атты I халықаралық конференция
ЕҢБЕКТЕРІ

ТРУДЫ
I Международной конференции
"Компьютерная обработка тюркских языков"

PROCEEDINGS
Of the I International Conference
on Computer processing of Turkic Languages (TurkLang-2013)

АСТАНА, 2013

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

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

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

«Түркі тілдерін компьютерлік өңдеу »
атты I халықаралық конференция
3-4 қазан, 2013 ж.

I Международная конференция
"Компьютерная обработка тюркских языков "
3-4 октября, 2013 г.

I International Conference
on Computer processing of Turkic Languages (TurkLang)
October 3-4, 2013

АСТАНА, 2013

УДК 81’322
ББК 81.1
Т 90

Т 90 ТҮРКІ ТІЛДЕРІН КОМПЬЮТЕРЛІК ӨҢДЕУ. Бірінші халықаралық конференция: Еңбектері/ Астана: Л.Н.Гумилев атындағы ЕҰУ баспасы, 2013- 340 бет
КОМПЬЮТЕРНАЯ ОБРАБОТКА ТЮРКСКИХ ЯЗЫКОВ. Первая международная конференция: Труды. – Астана: ЕНУ им. Л.Н. Гумилева, 2013. – 340 с.

ISBN 978-601-7454-85-2

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

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

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

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

УДК 81’322
ББК 81.1

Техникалық редакция: Бурибаева А.К.
Муканова А. С.
Ергеш Б.Ж.
Елибаева Г.З.

ISBN 978-601-7454-85-2

Программалық комитет

Шәріпбай Алтынбек Әміруұлы	Л.Н. Гумилев атындағы ЕҰУ, «Жасанды интеллект» ҒЗИ директоры, т.ғ.д., профессор	Төраға
Сулейманов Джавдет Шевкетович	Татарстан Республикасы ғылым академиясы вице-президенті, т.ғ.д., профессор	Тең төраға
Байбеков Сейтқасым Ниязбекұлы	Қазақ технология және бизнес университетінің президенті, т.ғ.д., профессор	Төраға орынбасары
Бөрібаева Әйгерім Кеулімжайқызы	«Жасанды интеллект» ҒЗИ аға ғылыми қызметкері	Ғалым хатшы

Адали Е.	Профессор	Түркия
Акалин Ш. Х.	Профессор	Түркия
Алтенбек Г.	Профессор	Қытай
Арипов М.М.	профессор	Өзбекстан
Байбеков С.Н.	профессор	Қазақстан
Желтов В.П.	профессор	Чувашия
Жұбанов А.К.	профессор	Қазақстан
Жүнісбек Ә.	Профессор	Қазақстан
Калимолдаев М.Н.	профессор	Қазақстан
Карабаева С.Ж.	профессор	Қырғызстан
Каримов Б. Р.	Профессор	Өзбекстан
Мусаев С.Ж.	профессор	Қырғызстан
Офлазер К.	Профессор	Катар
Силаму У.	Профессор	Қытай
Сиразитдинов З.А.	профессор	Башқұртстан
Тукеев У.А.	профессор	Қазақстан
Фатуллаев А.	Профессор	Әзірбайжан
Ыбраев Ш.Ы.	профессор	Қазақстан

Ұйымдастыру комитеті

Сыдықов Ерлан Батташұлы	Л.Н. Гумилев атындағы ЕҰУ ректоры	Төраға
Дихан Қамзабекұлы	Л.Н. Гумилев атындағы ЕҰУ проректоры	Тең төраға
Шәріпбай Алтынбек Әміржұлы	Л.Н. Гумилев атындағы ЕҰУ, «Жасанды интеллект» ҒЗИ директоры, т.ғ.д., профессор	Төраға орынбасары
Нұрбекова Жанат Құнапияновна	Ақпараттық технологиялар факультетінің деканы, п.ғ.д., профессор	Төраға орынбасары
Ергеш Бану Жантуғанқызы	«Жасанды интеллект» ҒЗИ ғылыми қызметкері	Техникалық хатшылар
Мұқанова Әсел Серікқызы		
Елибаева Ғазиза Қазбекқызы		

Бекманова Г.Т.	т.ғ.к., PhD доктор, Теориялық информатика кафедрасының меңгерушісі
Боранбаев С.Н.	ф.-м.ғ.д., Ақпараттық жүйелер кафедрасының профессоры
Адамов А.А.	т.ғ.д., Ақпараттық жүйелер кафедрасының профессоры
Омарбекова А.С.	т.ғ.к., Теориялық информатика кафедрасының доценті МА
Разахова Б.Ш.	т.ғ.к., Теориялық информатика кафедрасының доценті МА
Ниязова Р.С.	т.ғ.к., Теориялық информатика кафедрасының доценті МА
Аңдасова Б.З.	п.ғ.к., Теориялық информатика кафедрасының доценті МА
Қадеркеева З.К.	Теориялық информатика кафедрасының оқытушысы
Турмаганбетова Ш.	Теориялық информатика кафедрасының PhD докторанты
Кабдылова Д.	Теориялық информатика кафедрасының магистранты

МАЗМҰНЫ
СОДЕРЖАНИЕ
CONTENT

ПЛЕНАРЛЫҚ МӘЖІЛІС БАЯНДАМАЛАРЫ
ДОКЛАДЫ ПЛЕНАРНОГО ЗАСЕДАНИЯ
PLENARY MEETING

ТҮРІК ТІЛДЕРІН ЛАТЫНДАНДЫРУ: СТАНДАРТТАР ЖӘНЕ ТЕХНОЛОГИЯЛАР
ЛАТИНИЗАЦИЯ ТЮРКСКИХ ЯЗЫКОВ: СТАНДАРТЫ И ТЕХНОЛОГИИ
LATINIZATION OF TURKIC WRITING : STANDARDS AND TECHNOLOGY

КОМПЬЮТЕРЛІК ЖҮЙЕЛЕРДІ ҰЛТТЫҚ ЛОКАЛИЗАЦИЯЛАУ ЖӘНЕ ТЕРМИНОЛОГИЯ
НАЦИОНАЛЬНАЯ ЛОКАЛИЗАЦИЯ КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕРМИНОЛОГИЯ
THE NATIONAL LOCALIZATION OF COMPUTER SYSTEMS AND TERMINOLOGY

ТҮРІК ТІЛДЕРІНІҢ ЭЛЕКТРОНДЫ КОРПУСТАРЫ
ЭЛЕКТРОННЫЕ КОРПУСЫ ТЮРКСКИХ ЯЗЫКОВ
ELECTRONIC CORPORAS OF TURKIC LANGUAGES

МӘТІНДІ МОРФОЛОГИЯЛЫҚ ЖӘНЕ СИНТАКСИСТІК ӨҢДЕУ ЖҮЙЕЛЕРІ
СИСТЕМЫ МОРФОЛОГИЧЕСКОЙ И СИНТАКСИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ
SYSTEMS OF MORPHOLOGICAL AND SYNTACTIC PROCESSING OF TEXTS

СӨЙЛЕУЛЕРДІ СИНТЕЗДЕУ ЖӘНЕ ТАНУ ЖҮЙЕЛЕРІ
СИСТЕМЫ РАСПОЗНАВАНИЯ И СИНТЕЗА РЕЧИ
SPEECH RECOGNITION AND SYNTHESIS SYSTEMS

МӘТІНДЕРДІ СЕМАНТИКАЛЫҚ ӨҢДЕУ ЖҮЙЕЛЕРІ
СИСТЕМЫ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ
SYSTEMS OF SEMANTIC TEXT PROCESSING

ТҮРІК ТІЛДЕРІНЕ ОҚЫТУДЫҢ ТЕХНОЛОГИЯЛАРЫ МЕН ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕЛЕРІ
ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ ДЛЯ ОБУЧЕНИЯ ТЮРКСКИМ ЯЗЫКАМ
INTELLIGENT SYSTEMS AND TECHNOLOGIES FOR LEARNING TURKIC LANGUAGES

ПЛЕНАРЛЫҚ МӘЖІЛІС БАЯНДАМАЛАРЫ
ДОКЛАДЫ ПЛЕНАРНОГО ЗАСЕДАНИЯ
PLENARY MEETING

A.A. SHARIPBAY

Scientific-Research Institute "Artificial intelligence", Astana, Kazakhstan

PROBLEMS AND PROSPECTS OF COMPUTER PROCESSING OF THE KAZAKH LANGUAGE

1. The current state of the Kazakh language.

The Kazakh language is the national language of the Kazakh people, the indigenous population of the Republic of Kazakhstan and residing in many other countries around the world. So he, like any national language, which is the main means of communication of native speakers should develop.

The Kazakh language is the state language of the Republic of Kazakhstan and the demand for it, as in the internal relations of the country and in international relations has increased. So he, like any state language should be used in all areas of intellectual activity in our society and to get support from the state.

However, the status of the Kazakh language proves otherwise. First, the Kazakhs living in different countries can not be written to communicate among themselves, as they use different alphabets and spelling. Secondly, the Kazakhs living in their home country, can not fully use the Kazakh script due to the large diversity of terminology and spelling in printed matter. Third, the current Kazakh alphabet based on Cyrillic not always supported by modern computers and electronic means of communication and still produces linguistic differences.

The main reason for this state of the Kazakh language is the lack of basic standards on phonetics, spelling, grammar rules, terminology, etc. In the textbooks on the Kazakh language, there are many contradictory statements regarding the linguistic foundations of language.

To illustrate the effect of such a provision of the Kazakh language, you can give a short history of his writing.

It can be assumed that the source of writing Kazakh language are ancient Turkic writing, which are preserved in the runic monuments were first deciphered in 1893 by the Swedish scientist Thomsen (about it you can learn more in the museum of writing in the L.N.Gulilyov ENU).

A new Arabic-based Turkic writing method called usul jadid was invented by prominent Tatar enlightener Ismail Gaspirali(1851-1914). Using this method Akhmet Baytursynuly converted the Kazakh writing system to the Arabic alphabet in 1912. He refined the Kazakh phonetic system and developed the new Arabic-based Kazakh alphabet which consisted of 28 letters. This alphabet was used in our country until 1929, while the Kazakhs living abroad (particularly in China) still use it.

Soviet rule changed the Kazakh alphabet twice: first it was changed to the 29-lettered Latin-based alphabet in 1929, then 42-lettered Cyrillic-based alphabet in 1940. The latter reform has preserved all 33 letters of the Russian alphabet and added 9 letters for specific Kazakh sounds.

The latest reform is absurd from the point of view of the progressive linguistics, as it was done by force without features of the Kazakh language and destroyed its internal unity and grammatical patterns. This reform is not only not allowed to develop, but still damages the Kazakh language. Not necessary to be an expert in linguistics to verify this, because this reform is required to save the Kazakh text grammatical rules of the Russian language, i.e. writing and even reading the Russian words in the Kazakh text should be in accordance with the rules of Russian language. As a result of this reform in the Kazakh orthography has accumulated a significant amount of contradictions, which affect the culture of writing, spelling interfere with learning and as a result the Kazakh language is deformed by the day, changing their unique sound and grammatical patterns. For clarity, imagine that for the enrichment of the Russian language is proposed to add a sound of English, while keeping unchanged the rules of writing and reading English. It is clear that such a proposal for the enrichment of the Russian language is a complete nonsense. But despite the

absurdity of such enrichment of the Kazakh language, some believe that the correct pronunciation and spelling of Russian words in the Kazakh text is a sign of literacy.

Thus, forced cyrillization of writing of the Kazakh language has led to the fact that today a phonetic system and spelling norm of the Kazakh language were wrong and evil. Therefore, today it is necessary to correct the situation to the Kazakh language in the future, not to lose their identity and harmony.

Not understanding of these problems by experts and figures in linguistics and language policy is bewildering. Until now, publishes textbooks and scientific books whose authors can't distinguish between the concept of "sound" and "letter", does not know what a "phoneme", etc. Especially surprising the fact that many textbooks and training manuals on the Kazakh language contradict each other.

2. Problems of reforming of the Kazakh language and the change of the alphabet.

The current state of the Kazakh language not only allows it to full functioning as a state language, but also directly prevents it. The Kazakh language is deformed day by day, changing its sounding and disrupting the natural grammatical rules. Therefore it is necessary to **reform the Kazakh language,** to eliminate these and other contradictions.

For organizing and unification of the sound system is expected to develop and approve standard phonetics of the Kazakh language, which would require changing its alphabet and development spelling norms, as well as the processing of grammatical rules, etc.

It is important to note that changing of the alphabet of the Kazakh language is not necessary to engage in a modification of the current based on the Cyrillic alphabet, and it is necessary to pass at once to the Latin alphabet, to leave all the linguistic problems in Cyrillic and not to be confused where the new rules, and where the old.

At present we have developed the project of national standard of phonetics of the Kazakh language, which, based on the theory of A.Baitursynov determined the number of sounds of the Kazakh language, carried out their classification, built signs for the sounds of the phonetic transcription of the Kazakh language. The proposed project is a phonetic alphabet sounds of the Kazakh language examination will be held in the International Phonetic Association. These works are performed within the project for 2012-2014. "Creation of acoustic corpus of the Kazakh language and revision of its phonetic structure, representation of Kazakh phonemes in the International Phonetic Alphabet", financed by the MES on a priority "Fundamental and applied research in the field of economic, social and human sciences".

In this paper, we have only confirmed that the Kazakh language has only 28 sounds, 9 of them vowels and 19 consonants. Among the 5 vowel sounds а,о,ұ,ы,е are the phonemes, and 4 sounds ә,ө,ү,і – their allophones. All 19 consonants б,ғ,г,д,ж,з,й,қ,к,л,м,н,ң,п,р,с,т,у,ш are phonemes. In 1929, before developing a Latinized alphabet the number of consonants of the Kazakh language increased to 20, added sound хы, denoting the Latin letter h. This sound like the sound of в, ф, borrowed in 1940 from the Russian language does not violate the phonetic and phonological patterns of the Kazakh language, as they will help to correctly pronounce the international terms such as: валюта, вакуум, вакцина, вариант, вектор, вексель, вето, викторина, вирус, виртуал, вице, вокал, вольт, хадис, хаки, халат, хаос, химия, хлор, хор, хром, хроника, хрусталь, хунта, факт, факультет, фаза, файл, фауна, федерация, фельетон, физика, филармония, фильм, фонетика, формула, форфор, фосфор, фотон, фракция, функция.

Thus, we can assume that in the sound system of the Kazakh language will be 31 sounds, they are: ***а, ә, б, в, ғ, г, д, ж, з, е, й, к, қ, л, м, н, ң, о, ө, п, р, с, т, у, ұ, ү, ф, х, ш, ы, і.***. This proves that the need to change the alphabet of the Kazakh language.

3. The problems of computer processing of the Kazakh language

At present the development of any natural language is unthinkable without information technologies. There are already developed multimedia dictionaries, automation programs for ***teaching, machine translation, speech recognition and synthesis, text generation, text content***

interpretation, etc. for English, French, Russian, Japanese, Chinese and other languages. This level of development is possible for the Kazakh language too.

Computer processing of the Kazakh language involves the creation and use of the appropriate information resources in the Kazakh language. The databases of spelling and terminological dictionaries, the Kazakh language training systems, programs translators from the Kazakh language to another and vice versa, code conversion systems of the text in the Kazakh language from one graphics to another, speech recognition and synthesis system, etc. In this regard we carry out the project "Automation of recognition and generation of the Kazakh written and spoken languages" which is also financed by MES since 2011 on a priority "Information and telecommunication technologies".

It is known that all computers and telecommunications equipment manufactured in the world support and will support the English alphabet. To input and process characters of other alphabets one needs special code tables and drivers, their development and installation requires a considerable amount of money and intellectual efforts.

It is known that development of information resources is a science intensive and costly process. It is carried out by the use of modern information technologies that support the English alphabet by default. Today in many countries, Germany and Russia, are being discussed issues of development of information resources in the English-based alphabets, which not requires the development of fonts and program drivers, various sorting, search and recognition programs (e.g. for scanners). And all that requires large extra costs and highly qualified human resources. By this reason, in order to considerably facilitate development of information resources in Kazakh and reduce their costs it is necessary to ***convert the Kazakh writing system to the Latin graphics.*** We must also retain the same sequence of letters, this will enable us to use *built-in sorting, search and recognition programs* included in any system packages that rapidly change depending on the computer hardware specifications. Otherwise operating on the basis of Cyrillics alphabet of the Kazakh language generates a collateral and unjustified problem of support of the Kazakh alphabet without which decision possibility of use of modern information technologies is absolutely excluded. Some drivers are incompatible with each other that leads to a redoing of previously created information resources and require an additional labor and material costs.

Now because of the lack or inconsistency of such drivers, even in attempt of transfer of the text in the Kazakh language from one computer to another it is necessary to perform anew work on its set and formatting. At mass penetration into an everyday life of computers with various configurations and program systems the volume of such unnecessary works will increase. And if now not to take drastic and reasonable measures, the problem will pass from category difficult solvable in unsoluble, and we should in empty be spent as, first, quite often there will be vain works on creation of earlier created information resources, secondly, development and adjustment of necessary drivers demand considerable financial means and intellectual efforts.

For those who worries that because of alphabet change we will lose and we won't be able to read earlier published literary, cultural, scientific, it is possible to tell that there are no serious problems. For this purpose it is necessary to enter only once into memory of the computer the necessary text presented on the basis of any graphics. ***Translation of the Kazakh text from one graphics on other graphics can be automated, and then this text can be printed out for reading in any graphics*** [www.alphabet.kz]. So in 2013, we have prepared a project "Scientific, methodological, technological and methodological base for the conversion of the Kazakh writing to the Latin alphabet" and won grant funding of the Ministry of Education and Science of the Republic of Kazakhstan for 2013-2015.

Currently, under the guidance of my research made a lot of work on the computer processing of the Kazakh language. There are strong theoretical and practical groundwork in the following directions:

Phonetics:

– number and classification of sounds of the Kazakh language, their linguistic characteristics and methods of notation;

– the mathematical model of the phonetic system of the Kazakh language, which presupposes the separation of sounds into natural classes, suitable for automatic processing of speech;

– the applied model of the automatic transcriptor is constructed, allowing to pass from alphabetic structure of a word to a phonetic transcription.

Morphology:

– the mathematical model of morphological rules of the Kazakh language is constructed;

– methods and analysis algorithms and synthesis of words and word forms of the Kazakh language are developed.

Syntax:

– the mathematical model of syntactic rules of the Kazakh language is constructed;

– methods and analysis algorithms and synthesis of word combinations and offers of the Kazakh language are developed.

Speech Recognition:

– methods and algorithms of delimitation of speech, segmentation of a speech signal within the limits of wide phonetic classification are developed;

– methods and algorithms of recognition of pairs sounds with application two-threshold scalar recognizers and some specific sounds of the Kazakh language are developed;

– the algorithm of recognition of separate Kazakh words of the limited dictionary with use of a dynamic time number is developed;

– for fast search of the data the tree of transcriptions is constructed of the big dictionary;

– the concept recognition syntactically related sentences are developed.

Speech Synthesis:

– algorithms of transcription of the Kazakh words and sentences are developed;

– synthesis algorithms of the Kazakh speech from the free text are developed;

– the algorithm of training of model of synthesis of the Kazakh speech to the certain announcer is developed.

There are realizations of the methods set forth above and algorithms in the form of program applications which show fruitfulness of used approaches and mathematical models. Three participants are defended candidate dissertation on the specialty 05.13.11 – Mathematical and software of computers, systems and computer networks.

Three participants work over theses for a doctor's degree PhD on the topic of the project.

Five participants are working on master's thesis on the topic of the project.

The progress in automation of processing of the Kazakh writing and recognition and synthesis of the Kazakh speech will bring direct and immediate results in all spheres of intellectual activity of our country that confirms **prospects of a computerization of the Kazakh language**

4. Prospects for computer processing of the Kazakh language.

In order that it is correct to make a computerization and development of the Kazakh language it is offered to solve 77 problems in the following 7 directions:

I. Standardization of Kazakh language:

I.1. A standard phonetics;

I.2. A standard the alphabet and encoding of letters;

I.3. A standard orthography;

I.4. A standard grammar (morphology and syntax);

I.5. A standards industrial terms;

I.6. A standards onomastics and toponymy

I.7. A standard of measurement of Kazakh language knowledge.

II. Electronic dictionaries of Kazakh language:

II.1. Electronic orthographic dictionary;

II.2. Electronic orthoepic dictionary;

II.3. E-semantic dictionary of word forms;

II.4. Electronic Phraseological dictionary;
II.5. Electronic dictionaries;
II.6. Multilingual electronic dictionaries;
II.7. Electronic dictionary diphones.

III. Formalization of Kazakh language

III.1. Formalization of the phonetic and phonologic rules of sounds;
III.2. The formalization of the rules of word-formations with endings;
III.3. The formalization of the rules of word-formations with the suffix;
III.4. The formalization of the rules of formation of phrases;
III.5. The formalization of the rules of formation of simple sentences;
III.6. The formalization of the rules of formation of complex sentences;
III.7. The formalization of rules for the preparation of transcription and diphones.

IV. Automation of processing writing

IV.1. Drivers and converters with any graphics (encoding) to another schedule (encoding) and vice versa;
IV.2. Morphological analyzer;
IV.3. Generator word forms;
IV.4. The parser;
IV.5. Generator sentences;
IV.6. Converter of text in a semantic network;
IV.7. Semantic search engine.

V. Speech technologies of Kazakh language

V.1. Creation of an acoustic database;
V.2. Creation of an acoustic database;
V.3. Recognition of continuous speech;
V.4. Synthesis of individual words;
V.5. Synthesis of speech;
V.6. Hardware implementation of speech recognition;
V.7. Hardware implementation of the speech synthesis.

VI. E-learning Kazakh language

VI.1. E-learning and certification of the Kazakh language for the simple;
VI.2. E- learning and certification of the Kazakh language for basic level;
VI.3. E- learning and certification of the Kazakh language for intermediate level;
VI.4 E- learning and certification of the Kazakh language for good level;
VI.5. E- learning and certification of the Kazakh language for the higher level;
VI.6. E- learning and certification of the Kazakh language proficiency level;
VI.7. Educational Web-portal of the Kazakh language.

VII. Automatic translators

VII.1. Creation of Kazakh-Russian-English-Chinese translation and audio dictionary
VII.2. Creation of Kazakh-Russian and Russian-Kazakh translator;
VII.3 Creation of Kazakh-Russian and Russian-Kazakh audio translator;
VII.4. Creation of Kazakh-English and English-Kazakh translator;
VII.5. Creation of Kazakh-English and English-Kazakh audio translator;
VII.6. Creation of Kazakh-Chinese and Chinese-Kazakh translator;
VII.7. Creation of Kazakh-Chinese and Chinese-Kazakh audio translator;

Д.Ш. СУЛЕЙМАНОВ

НИИ «Прикладная семиотика» АН РТ, Казань, Татарстан

К ВОПРОСУ ВНЕДРЕНИЯ ТАТАРСКОГО ЯЗЫКА В КИБЕРПРОСТРАНСТВО

Введение

Востребованность языка как когнитивного и коммуникативного средства, а также его развитие, соответственно, и дальнейшее сохранение его как культурного явления, очевидно, в значительной степени зависят от активности функционирования языка в компьютерных информационных технологиях.

Данный тезис раскрывается нами на примере внедрения татарского языка в так называемое киберпространство (то есть пространство взаимодействия человека и компьютерных систем и технологий). При этом, как будет показано ниже, обеспечение функционирования татарского языка в компьютерных системах и технологиях является актуальным не только в плане повышения его активности и конкурентоспособности среди других языков в качестве средства накопления информации и общения с компьютером, но также и в плане создания новых технологий хранения и обработки информации на основе татарского языка в силу целого ряда когнитивных особенностей его структуры и лексического корпуса.

Очевидно также, что для обеспечения равного функционирования татарского и русского языков как государственных в Республике Татарстан, необходимо, чтобы татарский язык так же, как и русский, стал рабочим языком компьютеров. Соответственно, наряду с задачей использования татарского языка в инфокоммуникационных технологиях и создания специальных программ обработки татарского языка, ставится также задача татарской локализации их интерфейсной оболочки, т.е. средств общения компьютера с человеком.

Исследования и разработки по внедрению татарского языка в компьютерные технологии в Республике Татарстан начались практически с конца 1980-х годов, с разработки первых драйверов периферийных устройств, текстового редактора и татарского корректора, необходимых для компьютерного издания татарских книг, газет и журналов и ведения делопроизводства. В 1993 г. для решения задач в рамках научно-прикладной программы Академии наук РТ «Компьютерное обеспечение функционирования татарского языка как государственного» и для разработки средств компьютерного обеспечения татарского языка как государственного в рамках Государственной программы РТ по сохранению, изучению и развитию языков народов Республики Татарстан была создана Совместная научно-исследовательская лаборатория Академии наук РТ и Казанского государственного университета «Проблемы искусственного интеллекта».

Фундаментальные исследования и прикладные разработки по поддержке татарского языка в информационных технологиях изначально осуществляются в трех основных направлениях:

1) внедрение татарского языка в информационные технологии («Татарский язык в ИТ»),

2) разработка и адаптация информационных технологий для татарского языка («ИТ для татарского языка»),

3) использование когнитивных возможностей татарского языка для создания новых информационных технологий («Татарский язык для ИТ»).

1. Внедрение татарского языка в информационные технологии

Первое направление исследований и разработок «Татарский язык в ИТ» непосредственно связано с проблемой сохранения языка, повышения его активности в мировом инфокоммуникационном пространстве, использования татарского языка в киберпространстве как когнитивного и как коммуникативного средства, т.е. средства

представления, накопления и передачи информации, обеспечения паритетного функционирования татарского и русского языков как государственных в Республике Татарстан, а также предоставления возможности носителям языка прямого общения с компьютерными системами без языка посредника. Данное направление работ включает базовую и полную локализации компьютерных систем, то есть адаптацию их под татарский язык.

В настоящее время эта задача решена в полном объеме для татарского языка на основе кириллической графики. Учеными Академии наук РТ и КФУ разработаны экранные и клавиатурные драйверы, драйверы печати и шрифтовое обеспечение для татарского языка на кириллической основе и предложены в качестве стандарта для применения в информационных технологиях в Республике Татарстан. На их основе принято Постановление КМ РТ «О стандартах кодировки символов татарского алфавита для компьютерных применений» (N 1026 от 9 декабря 1996 года).

Данное Постановление помогло унифицировать драйверы устройств, которые в первое время создавались различными группами и отдельными специалистами по своему усмотрению, и практически как вирус распространились по различным компьютерам, закрепляя разную раскладку одних и тех же татарских букв на кодовых страницах, создавая «разночтение». Унификация кодовой страницы, соответственно, и драйверов устройств, помогла ликвидировать начавшийся хаос в делопроизводстве, когда татарские тексты, набранные на одной машине, не читались на другой или отображались некорректно.

На базе принятых стандартов по соглашению с фирмой Майкрософт были разработаны соответствующие драйверы устройств и внедрены в операционную среду Windows NT и Office-2000. В настоящее время пакет драйверов TATWIN, включенный в программный комплекс поддержки татарского языка TatSoft 2, позволяет вести делопроизводство на татарском языке на кириллической основе во всех приложениях операционной системы MS Wındows’95, ’98, ’2000, ’XP, Vista, Wındows’7, Wındows’8, а также работать в Интернете. Соответствующая информация имеется на web-сайте фирмы Майкрософт. Таким образом, татарский язык стал вторым тюркским языком (после турецкого языка), подготовленным для реализации специалистами самой республики (а не разработчиками фирмы), и доступным в среде Windows при ее инсталляции на любом рабочем месте.

Сотрудничество Академии наук РТ с Московским бюро фирмы Майкрософт, начавшееся уже в 95-е годы с татарской локализации ОС Windows’95, нашло перспективное продолжение. В 2005-2010 годах осуществлена полная татарская локализация основных продуктов фирмы Майкрософт. Научно-исследовательским институтом «Прикладная семиотика» Академии наук РТ и лабораторией «Проблемы ИИ» КФУ разработан татарский интерфейс операционной системы и, таким образом, татарский язык, наряду с такими мировыми языками как английский и русский, стал родным языком для операционной системы Windows и таких активно используемых пользовательских программ как Word, Exel, Power Point.

Татарская локализация операционной среды MS Windows и ее приложений ведет к активному внедрению татарского языка в инфокоммуникационные технологии, развитию татарского языка и распространению его в мировом информационном пространстве. Очевидно, что только становясь языком компьютерных технологий, языком накопления, обработки, передачи информации, языком общения с компьютерными системами, татарский язык, впрочем, также и языки других народов, имеет возможность стать полнокровным государственным языком в республике, языком культуры, языком науки, языком общения в киберпространстве.

2. Разработка и адаптация информационных технологий для татарского языка

В рамках *второго направления* «ИТ для татарского языка» разработаны пакеты прикладных программ для работы с татарским языком, программные средства, позволяющие компьютеризировать делопроизводство, издание газет и журналов, проверять корректность

татарских текстов, автоматизировать рабочие места специалистов. Осуществляются 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 прагматически-ориентированных лингвистических моделей и создания на их базе систем автоматизированной обработки татарского языка. Важными и активно разрабатываемыми и, очевидно, судьбоносными являются вопросы татарской терминологии в киберпространстве.

В настоящее время создана полнофункциональная компьютерная модель морфологии татарского языка, причем, учитывая структурную специфику татарского языка и исходя из прикладных задач, разработаны три различные модели морфологии. Генеративная модель морфологии, основанная на правилах словоизменения, хотя и уступает другим моделям по быстродействию, обеспечивает полноту анализа словоформы, позволяя в полной мере учитывать агглютинативный характер языка, распознавая словоформы потенциально неограниченной длины. Парадигматическая модель татарской морфологии обеспечивает быстрое распознавание словоформ и анализ корректности татарских словоформ с точностью до 95 %, используется в поисковой системе УИС «Россия» (ЦИТ МГУ, г. Москва) и в среде MS Windows и ее офисных приложениях. Причем, скорость распознавания составляет до 100 слов в 0.014 секунд, что перекрывает требования заказчика на целый порядок. Кроме того, в рамках совместного проекта с Белкентским университетом (Турция) разработана двухуровневая модель морфологии татарского языка, реализованная в среде известной программной оболочки PC KIMMO и используемая в составе татарско-турецкого машинного переводчика. Создана также структурно-функциональная модель татарских аффиксальных морфем, являющаяся «инвентарной базой» для построения различных прагматически-ориентированных морфологических моделей и на ее базе построен интегрированный программно-информационный комплекс «Татарская морфема». Данный комплекс практически является автоматизированным рабочим местом (АРМом) для разработчиков различных лингвопроцессоров, а также для осуществления учебно-исследовательской деятельности в татарском языкознании, может быть успешно использован как исследовательский инструмент и для других языков.

Еще одна полезная программа - татарско-русский машинный переводчик татарских фамильно-именных групп, созданная на основе словаря компонент и правил, учитывающих специфику образования татарских собственных имен, является незаменимым инструментом в автоматизированных системах ЗАГС и Паспортно-визовой службы, а также для автоматической генерации татарских имен и фамилий на основе модели компонент татарского имени. Специалистами института осуществлена татарская локализация оптического распознавателя текстов FineReader московской фирмы ABBYY. Данная программа, благодаря встроенной морфологии татарского языка, распознает татарские тексты с такой же точностью и быстротой, как и русские и английские.

Важной задачей, которая выполняется институтом, является создание и поддержка электронного корпуса татарского языка, практически представляющего собой машинный фонд 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) программные модули: лингвопроцессоры (машинные переводчики, синтезатор речи, распознаватель текста и речи и др.), АРМы специалиста (учителя, редактора, лингвиста и др.), интеллектуальная многоязычная машина поиска. Задача создания электронного корпуса татарского языка является фундаментальной научно-практической проблемой, решение которой даст возможность быстрого и удобного доступа к различным лингвистическом ресурсам большого объема посредством использования вычислительных машин. Очевидно, наличие богатой лингвистической базы, отображающей татарский язык практически во всех его проявлениях в речи и тексте, включая диалекты, позволит проводить достоверные научные исследования на основе данного фактографического материала, а не только на основе лингвистической интуиции самого

исследователя и ряда примеров из доступных источников, как это делается, как правило, в настоящее время. Реализация данного проекта приведет к формированию соответствующей инфраструктуры (татарский контент и средства работы с татарским контентом) для полноценного представления татарского языка в сети Интернет.

Одним из интересных и полезных продуктов, разработанных институтом совместно с фирмой ABBYY и ИЯЛИ АН РТ, являются Многоязычные электронные словари Lingvo'x3 с татарским языком, представляющий собой практически настольную библиотеку из 154 различных словарей на 12 языках мира, в числе которых имеется и татарский язык. Ценность данного электронного словаря для татаро-язычного пользователя, кроме многих других возможностей, заключается в том, что через татарско-русскую языковую пару доступны переводы во всех 154 словарях на 11 языках мира (то есть, включив татарско-русский словарь объемом порядка 60000 словарных статей, потенциально мы получили татарско-английский, татарско-французский, татарско-испанский, татарско-немецкий, татарско-китайский, татарско-турецкий и др. двуязычные словари). Линейка словарей, включенных в Лингво-оболочку, с выходом новых версий, постоянно расширяется. Уже появилась новая версия Lingvo'x5, в которой через татарско-русский словарь доступны переводы слов на 20 языках народов мира.

Незаменимым инструментом в делопроизводстве и издательском деле является программа WordCorr – морфологический корректор татарских текстов для Microsoft Word, который позволяет находить и исправлять ошибки в татарских текстах, при этом предлагая возможные корректные варианты. Функционирует во всех операционных системах Windows'95 '98 '2000 'XP, Vista, Win 7, Win 8 и приложениях.

Практически с 1990-х годов осуществляется активная работа по разработке электронных обучающих программ татарскому языку, а также программ обучения предметов на татарском языке. Ряд последних разработок доступны в Интернете, среди них: Татар Телле Заман (ТТЗ) – мультимедийный электронный учебник по татарскому языку, Татар-онлайн – мультимедийный Интернет-учебник по татарскому языку, мультимедийный учебник 5 класса для дистанционного Интернет-обучения татарскому языку.

Программа «Татар Телле Заман» содержит более 2000 татарских слов, более 2500 рисунков и фотографий, озвученные диалоги на различные темы и 11 увлекательных лингвистических игр, три типа различных упражнений, позволяющих тестировать знания обучаемого, возможности для совершенствования татарского произношения вслед за диктором. Многоязычный интерфейс (русский, татарский (кириллица, латиница), английский) системы позволяет изучать татарский язык как в русскоязычной, так и англоязычной среде. Татарские версии электронных мультимедийных учебных пособий Химия-8 и Физика-7, разработанные совместно с московской фирмой «Просвещение-Медиа» при содействии Министерства образования и науки РТ и Издательства «Магариф», благодаря комплексу разнообразных мультимедийных возможностей (видеосюжеты, анимация, звук, качественные иллюстрации, сотни интерактивных заданий и т.д.) обеспечивают увлекательный и эффективный процесс обучения. Разработано и передано в школы республики электронное мультимедийное учебно-методическое пособие «Татар теле-5». Электронное пособие содержит учебный материал по 6 темам, 123 упражнения, разделенных на 27 типов; включает гипертекстовый справочный материал по татарскому языкознанию, руководство пользователя и анимационную контекст-подсказку по запросу пользователя в он-лайн режиме. Программное обеспечение и технологии разработки и реализации мультимедийных учебных пособий, разработанные с ориентацией на татаро-язычную среду, в основе своей являются универсальными, независимыми от языка и успешно могут быть использованы также при создании электронных учебных пособий для других проблемных областей и для других языков.

В институте активно разрабатываются прагматически-ориентированные речевые технологии. В настоящее время синтезатор татарской речи и распознаватель корректности произнесенного татарского предложения внедряются в состав лингвистического пакета EF

(Education First) для использования в уникальной дистанционной системе обучения “Ана теле”, инициированного Президентом Республики Татарстан и обеспечивающего обучение татарскому языку 24 часа в сутки в течение 7 дней в неделю.

Среди перспективных работ института можно выделить следующие проекты.

1. Разработка Интеллектуальной многоязычной поисковой машины (ИМПМ). Актуальность работ по созданию ИМПМ связана с необходимостью создания машинного фонда (ресурса электронных коллекций) татарского языка, сложившейся языковой ситуацией в республике Татарстан, появлением новых лингвистических и интеллектуальных технологий многоязыкового поиска, основанных на глубоком разрешении лексической многозначности. Кроме того, потребность в многоязыковых поисковых технологиях обусловлена тем фактом, что ряд развитых государств имеют несколько официальных языков, что дает проекту перспективу дальнейшего коммерческого использования.

2. Разработка программы распознавания татарской речи. Как прогнозируется специалистами, одним из основных направлений развития в сфере высоких технологий в ближайшие годы будут речевые технологии, особенно, автоматическое распознавание речи (АРР). Ожидается широкое внедрение технологий АРР в ведущие сектора экономики. По оценкам аналитиков, объём рынка продукции, использующей АРР, будет сравним с рынками таких высокотехнологичных товаров как микропроцессоры, персональные компьютеры, программное обеспечение.

3. Разработка татарско-русского машинного переводчика, а также машинных тюркоязычных переводчиков в паре с татарским языком. Если особая актуальность машинных переводчиков первой группы объясняется необходимостью доступа к англоязычным базам знаний в Интернете через русский язык (априори предполагается, что русско-английский переводчик имеется) и необходимостью поддерживать равное функционирование татарского и русского языков как государственных в Республике Татарстан, то вторая группа - среди родственных языков, эта работа привлекательна в силу относительной простоты и малой затратности решения этой задачи (в некоторых случаях практически это простая конвертация текстов, например, для татарско-башкирской пары языков), а также в силу культурологической функции такого переводчика, помогающего сближению родственных народов.

3. Использование когнитивных возможностей татарского языка для создания новых информационных технологий

Третье направление исследований «Татарский язык для ИТ» связано с актуальной задачей создания интеллектуальных операционных систем и интеллектуального программного инструментария на основе использования потенциала естественных языков, их семантических и синтаксических конструкций, а также лексического корпуса. Очевидно, что естественный язык является основой для любой символической системы, определенным образом организованной, имеющей свой синтаксис и свою семантику (сюда же включается любая логика, математика и др.). Соответственно, вместе с языком в этих системах реализуется и ментальность языка (точнее, ментальность этноса, передаваемая через язык).

Что является важным для компьютерных технологий? Известно, что критичными, соответственно, важными для компьютерных технологий являются такие показатели как *время обработки* информации, *объем памяти* для хранения информации (сжатие информации), *активность знаний* и возможность *задания нечетких команд* (однозначно воспринимаемых в определенном контексте). Последние два свойства являются необходимыми характеристиками для интеллектуальных систем и технологий. Соответственно, весьма актуальными и перспективными являются когнитивные исследования в языке с целью определения таких структур, схем, формул, которые в естественном языке реализуют указанные свойства и могут быть эффективно использованы при создании искусственных языков и систем программирования, а также любых других средств описания, хранения и обработки информации.

Как известно, операционные системы, языки программирования, средства обработки информации, практически все программное обеспечение, используемое сегодня, разработаны на основе английского языка и, соответственно, на основе менталитета английского языка (менталитета, отражаемого через английский язык - западного менталитета). Английский язык является языком флективно-аналитического типа (флексия – когда допускается и префиксное, и инфиксное, и постфиксное изменение формы слова; аналитический тип – когда новое значение образуется сочетанием слов), практически с нулевой морфологией (по сравнению с агглютинативными языками). Отсюда следует, что сложный смысл образуется словосочетаниями и это приводит к большой комбинаторике при анализе. А это, в свою очередь, ведет к увеличению самых критичных показателей в вычислительных системах - объема требуемой памяти и времени при обработке информации. Выход из такой ситуации – исключение большого контекста, глубины конструкций, в итоге - упрощение смысла, семантики, соответственно и «интеллектуальных показателей». Таким образом, в основе самого английского языка заложен «интеллектуальный» тупик для вычислительных машин, заставляющий их не "умнеть", а искать выход через повышение быстродействия системы и увеличение памяти, т.е. через развитие «физики» (hardware), а не «мозгов» (software).

Еще один недостаток технологий, основанных на английском языке, заключается в том, что сам строй языка, его синтаксис, «сопротивляется», даже противоречит одному из главных свойств интеллектуальности системы – *активности знаний.* Как известно, английский язык относится к языкам типа SVO (Subject-Verb-Object). То есть, «Субъект: Действие-Информация» (*I'll go to the cinema tomorrow afternoon with my friend ...*). Таким образом, сначала требуется выполнить, потом рассуждать, анализировать. Решение принимается не на основе информации, а информация подается в рамках выбранного действия. То есть, не информация диктует, какое именно действие необходимо совершить, какие методы, алгоритмы применять для ее обработки, а наоборот, действие, средство, схема, алгоритмы заставляют форматировать, структурировать, модифицировать информацию.

В отличие от индо-европейских языков, тюркские языки относятся к языкам типа SOV (Subject -Object-Verb). Соответственно реализуется схема: «Субъект: Информация-Действие». Например, смысл английского предложения, приведенного выше, будет передаваться следующим татарским предложением: *Min (я) ... irtege (завтра) toshten (обед) song (после) dustym (друг) belen (с) kinoga (в кино) baram (иду) (букв.: Я ... завтра после обеда с другом в кино иду).* То есть, в татарском предложении сначала раскрывается информация, анализ ситуации, а затем уже в конце предложения приводится действие, отображаемое, как правило, глаголом.

Как показывают исследования, проводимые в НИИ «Прикладная семиотика» АН РТ, а также зарубежными исследователями [1], тюркские языки, как агглютинативные языки, обладающие регулярной морфологией и, вместе с тем, естественной сложностью, разрешаемой по контексту, являются эффективным инструментом для создания интеллектуальных систем обработки информации [2-4]. В силу минимальных показателей временных и емкостных оценочных функций для генерации и анализа цепочек словоформ (за счет регулярности) достигается эффективность при накоплении и обработке информации. Компактность передачи смысла текста на поверхностном, лексическом, уровне объясняется также возможностями языка синтетически, т.е. словоформой, кодировать смысл, который для других языков (английский, русский) формируется аналитически, чаще всего, несколькими предложениями.

Агглютинативность морфологии, минимальность исключений, наличие мощного мета-аппарата, синтаксическая мотивированность активности информации в татарском тексте, позволяют ставить задачу о возможности создания языка промежуточной трансляции, т.е. языка-посредника на базе татарского языка, и даже разработки новых операционных систем на основе новой идеологии.

Заключение

В докладе изложены результаты деятельности НИИ «Прикладная семиотика» Академии наук РТ и НИЛ «Проблемы искусственного интеллекта» АНТ И КФУ за последние 10-15 лет в области создания стандартов и программных средств обеспечения паритетного функционирования татарского языка в инфо-коммуникационных технологиях в качестве одного из государственных языков в Республике Татарстан. Показан ряд потенциальных когнитивных возможностей татарского языка, позволяющий ему стать формальной базой для построения новых средств описания, хранения и обработки информации. Также подчеркивается, что только становясь языком компьютерных технологий, языком накопления, обработки, передачи информации, языком общения с компьютерными системами в киберпространстве, татарский язык имеет возможность стать полнокровным государственным языком, языком общения, языком науки.

Литература

1. Heintz J. and Schonig C. Turcic Morphology as Regular Language // Central Asianic Jornal (CFJ), 1989. -P.1-24.
2. Suleymanov D.S. Natural possibilities of the Tatar morphology as a formal base of the NLP // In Proceedings of the First International Workshop “Computerisation of Natural Languages” (Varna, Sept. 3-7, 1999). –Sofia (Bulgaria): Information Services Plc, 1999. -P.113.
3. Сулейманов Д.Ш. Естественные когнитивные механизмы в татарском языке // В Тр. Межд.семинара Диалог-2002 “Компьютерная лингвистика и интеллектуальные технологии” (г.Протвино, 6-11 июня 2002 г.): в 2 т. / Под ред. А.С.Нариньяни. – М.: Наука, 2002. –С. 500-507.
4. Suleymanov D.S. Natural cognitive mechanisms in the Tatar language // In the Collection of the Vienna Proceedings of the Twentieth European Meeting in Cybernetics and Systems Research. Edited by Robert Trappel. Vienna, Austria, 6-9 April, 2010. – P. 210-213.

E. ADALI

Istanbul Technical University, Computer Engineering and Informatics Faculty, Istanbul, Turkey

TURKS' EXPERIENCES WITH THE LATIN ALPHABET

A Brief History of the Alphabet

An alphabet is a set of symbols or characters that represents the sounds of a language in writing. The purpose of an alphabet is to establish an exact, one-to-one correspondence between each sound and its symbol. However, some languages use a diphthong (two vowels), or assign multiple consonants for one sound.

The Phoenicians developed the first alphabet in the 18th century BC. This alphabet consisted of just consonants. Over time, the Aramaic, the Hebrew and the Arabic alphabets were derived from the Phoenician alphabet. In the ancient Hellenistic periods, Anatolian nations added vowels to the Phoenician alphabet and adopted it as the Greek alphabet. The classical Latin alphabet evolved from the Greek alphabet called the Cumaean alphabet, which was adopted and modified by the Etruscans who ruled in the early years of Rome.

In honor of two Köktürk princes Kul Tigin and his brother Bilge Kağan, two monuments were erected in the Orkhon Valley (now in Mongolia) in 732 and 735. The oldest Turkic alphabet, which is called the Göktürk alphabet, was used on the Orkhon-Yenisev inscriptions in the 8th century; however, the exact origins of the Göktürk alphabet is uncertain. The website of the Language Committee of the Ministry of Culture and Information of the Republic of Kazakhstan lists 54

inscriptions from the Orkhon, 106 from the Yenisev, 15 from the Talas, and 78 from the Altai area. Another old alphabet used by the Turks is the Uyghur alphabet.

In the 8th century, some Turkic groups started migrating away from Middle Asia in different directions. The Oğuz Turks, who moved to Anatolia, first founded the Selçuk States, then the Ottoman Empire in 1299 and finally the Turkish Republic in 1923. During this move, they accepted Islam and adopted the Arabic alphabet. In 1928, Turks started using the Turkish Alphabet, which was derived from the Latin alphabet.

The Yakut, Azeri, Uzbek and Turkmens developed their own Latin based alphabets during the early periods of the 20th century, but they were not able to use them due to USSR oppression. Nowadays Turkmenistan, and Azerbaijan, use the Latin alphabet but Uzbekistan still uses both the Latin and Cyrillic alphabets.

Why Need a New Alphabet?

Today, even though most countries use the Latin alphabet, there are some who still use different alphabets, such as the Arabic, Cyrillic, Hebrew, Greek, and Georgian alphabets. Can all nations use the same alphabet? How can they choose which alphabet to use? Answers to these questions are as follows:

Historical and Cultural Reasons

Nations in this group adopted or developed their alphabets a long time ago and have been using them ever since. Over time, they modified these alphabets according to their own language needs. These nations believe that their alphabets are the best, at least for their own languages. They also believe that their alphabets are part of their heritage. Since they see their alphabets as trademarks of their languages and nations, they do not want to change them even though they experience some technical difficulties. We can put Arabic, Hebrew, Greek and Georgian alphabets in this category.

Political Constrains

In the past or even today, some nations exist as mandates of or as minorities in another nation. Imperial states dictate an alphabet to their mandates or minority nations. In history, Arab States, USSR, and today the Chinese government, dictate their alphabets. Therefore, some countries located in the Middle East and North Africa still use the Arabic alphabet, while those in the former USSR territory use the Cyrillic alphabet.

Religious Reasons

Holy books of religions are written in the alphabets of the languages used in the countries the religions originated from. Koran is a good example of this case. Nations other than the Arabs, such as Selçuks, Turks and Iranians who want to read the Koran decided to adopt the Arabic alphabet. However, when they realized that the Arabic alphabet is not fully suitable for their languages, they had to modify it and added some new letters and symbols.

Several North African nations also changed their languages and now they are speaking Arabic even though they are not ethnically of Arab origin.

Technical reasons

In the past, not many people were literate. Books were handwritten by penmen and thus, they were very expensive. Invention of the printing machine made it affordable for everyone to get a book and this encouraged them to learn how to read and write.

Classical printing machines use discrete letters. The Arabic script is a longhand style and the form of a letter varies at the beginning, in the middle and at the end of a word. Therefore, a typographic house needs more than 700 different characters. Today, computer systems can solve this problem easily.

In some cases, the current alphabet may not be suitable for the language of the nation. Turkish language and the Arabic alphabet is a good example for this case. Although Turkish has 13 vowels, the Arabic alphabet has only three vowels. On the contrary, the Arabic alphabet has many consonants, but Turkish does not have as many.

In the second half of the 19th century, the telegraph system was used for military and commercial purposes. The telegraph system uses the Morse alphabet, which is designed for the

Latin alphabet. During the First World War and the War of Liberation of Turkey, the Turkish army and diplomats used the Latin alphabet for telegraph messages.

Nowadays, we have a similar problem in the field of information systems. Most of the international standards, such as the character set, the Internet, the Electronic Data Interchange (EDI), etc., are developed for the Latin alphabet.

If a Nation Changes its Alphabet, What Will it Lose?

The decision of changing the current alphabet is not an easy one. Especially in this century, it is even more complicated. The following cases must be considered:

1. All books, commercial documents, official papers, official records, signboards, and nameplates have been written in the old alphabet.
2. People know the old alphabet. When the new alphabet becomes the statutory alphabet, in a short time, everybody becomes illiterate.
3. In order to educate the population, many teachers will be required. Therefore, before changing the alphabet, a sufficient number of teachers must be educated.
4. The sound analysis of the language must be performed. Required vowels and consonants of the language must be identified. If a known alphabet will be adopted, the letters of the alphabet must be analyzed to check if they are sufficient to represent the sounds of the language or not.

Experiences of the Turks

Turks first used the Göktürk alphabet, then the Uyghur alphabet. Later, they started using the Arabic alphabet in the 9th century and finally accepted the Turkish alphabet, which is based on the Latin alphabet, in 1928.

When the Turks realized that the Arabic alphabet was not adequate for Turkish, they added some new characters and some new signs to this alphabet. They modified the shape of the characters and gave them artistic features.

The first printing house in the Ottoman Empire was opened in 1726. Around the 1850's books and newspapers were printed in the state. During this period, Turks realized the following difficulties of the Arabic alphabet.

- Arabic letters are written in running hand form. In other words, the letters are not discrete.
- Each letter has three forms; 1^{o} At the beginning of the word; 2^{o} In the middle of the word and 3^{o} At the end of the word.
- There are no capital forms of the letters. Therefore, a proper noun cannot be typed.
- There are only three vowels (a, i, u) in the Arabic alphabet. Turkish basically has 8 vowels, which are a, e, ı, i, o, ö, u, ü. In addition, Turkish has two types of "e" sounds (lips and round) and the "a, ı, i, ü" letters also have a long form.
- In the old times, a student could learn to read and write in Arabic in 4 or 5 years.

To overcome these difficulties, the following actions were taken:

– Turks tried to modify the Arabic alphabet. They added the Turkish vowels (a, e, ı, o, ö and ü) and some Turkish consonants (ç, j, g, p). The following is an example for this: □□□□ (dede). Letter "he" is used as the vowel "e". Also, for military uses Enver Paşa developed a special alphabet, which has discrete letters.

– In the 19th century Turks used the Latin alphabet for commercial documents. Example: Hadji Bekir (Hacı Bekir), Hussein Djahid (Hüseyin Cahit), Istamboul (İstanbul). As you can see, they used the French spelling rules for the Turkish words.

– When the Turkish Republic was five years old, in June of 1928, the president of the state, Mustafa Kemal decided to develop the Turkish alphabet based on the Latin alphabet. After this decision was made;

1. Firstly, a linguistic committee was formed. This committee worked on the sound analysis of the language and on developing the letters of the Turkish alphabet.
2. Many teachers were educated.

3. Many courses were opened throughout the country for the elderly. Roughly 2.500.000 people, of which 1.124.916 were elderly, were educated in five years. The population of Turkey in 1927 was 14.832.725. In Table-1, the number of courses, teachers and students are shown.

Table-1: Some figures about teaching new alphabet

Students of Public School for Elderly People					
Years	Number of courses	Number of teachers	Number of student	Number of graduates	Rate %
1928/29	20.489	16.922	1.045.500	526.881	50,39
1929/30	12.937	11.307	544.534	245.663	45,11
1930/31	9.602	8.940	352.902	172.322	48,82
1931/32	5.915	5.437	205.349	99.491	48,44
1932/33	5.107	4.084	157.639	80.559	51,10
Total	54.050	46.690	2.305.924	1.124.916	48,78

The population of Turkey in 1927 was 14.832.725

- On the first of November of 1928, the code of the Turkish alphabet was declared by the Turkish government.
- The letters of the Turkish alphabet are: a, b, c, ç, d, e, f, g, ğ, h, ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, z. This alphabet consists of 8 vowels (a, e, ı, i, o, ö, u and ü) and 21 consonants. Some characters (ı ; I, ğ; Ğ, ş; Ş) are new for the Latin alphabet. In Turkish some vowels have two or three forms. To solve this problem, some special signs such as " ^ " and " ' " were added. The " ^ " sign makes a sound softer and/or longer; the " ' " sign means stop reading. The vowels and the consonants of Turkish are given in Table-2 and Table-3 respectively. In Table-4, the vowels and the consonants of Turkish are shown from a different perspective.

Table-2: The Vowels of Turkish

Vowels	**Unrounded**		**Rounded**	
	Wide	Narrow	Wide	Narrow
Back vowel	a	ı	O	U
Front vowel	e	i	Ö	Ü

Table-3: The Consonants of Turkish

Consonants		**Labial**	**labio-dental**	**Dental**	**Plato-alveolar**	**Palatal**	**Velar**	**Glottal**
Voiceless stop	Hard	P		t	Ç	k (front)	k (back)	
Voiced stop	Soft	B		d	C	g (front)	g (back)	
Voiceless fricative	Hard		f	s	Ş			
Voiced fricative	Soft		v	z	J			
Nasal		m		n				
Liquid				l, r				

Approximant						Y		h

Table-4: The Vowels and Consonants of Turkish

Vowels and Consonants						
a	front	f		l	front	T
a	Back	g	front	l	back	U
a	Long	g	back	m		Ü
b		ğ		n		V
c		h		o		Y
ç		i		ö		Z
d		i		p		
e	Open	j		r		
e	Close	k	front	s		
e	Long	k	back	ş		

4. The most important feature of the Turkish alphabet is that, it is a phonetic alphabet, which means that one letter represents one sound of the language. In the Turkish alphabet, diphthongs (two vowels) and multiple consonants are not used for one sound.

5. In the second half of 1928, newspapers used both alphabets (Arabic and Turkish). Starting from the first day of January 1929, newspapers and books were printed with the new alphabet. All official documents, official records, signboards and nameplates were written with the new alphabet in one year. This process was completed by the end of 1929.

6. Today, children can learn to read and write with the Turkish alphabet in 4 to 6 months. In 1927, the percentage of literacy in the population was less than 20%, but today it is 95%.

7. After changing the Arabic alphabet, Turks started working on improving the Turkish language. In 1927, the Turkish language had many Arabic and Persian words. Statistics show that 38% of the words were Turkish and 58% were Arabic and Persian. The grammar of the language was similar to the Arabic or the Farsi grammar. In the 1990's the percentage of Turkish words were increased up to 85% to 90%.

The Results of the New Alphabet

An alphabet may be changed in a short time or over a long period. Although Turks made the switch in one year, the Uzbek are doing it in 20 years. Both methods (fast and slow) have some advantages and disadvantages.

The Fast Method

- The new generation can learn the new alphabet in regular schools.
- Public schools must be opened for the elderly. A sufficient number of teachers must first be educated before these schools can be opened.
- Typewriters should be changed or modified, if they are still being used.
- IT systems must be set for the new alphabet.
- All signboards and nameplates have to be changed in a short time.

The Slow Method

- The new generation can learn the new alphabet in regular schools.
- Elderly people will not be willing to learn the new alphabet.
- It will not be necessary to open public schools for the elderly.
- All signboards and nameplates have to be written with both alphabets.

Recommendation for Nation Wishing to Change Their Alphabets

- Today, the Latin alphabet has come to be regarded as the World standard. So, if a nation decides to change its current alphabet, a Latin based alphabet would be the best choice.
- A phonetic alphabet will be the best choice.
- Changing the current alphabet with a new one causes some loss of heritage. To preserve the old documents, books and records, they should be converted to the new alphabet. Today, OCR techniques will aid in this process.
- The sound analysis of the language must be done. Based on this analysis, the appropriate letters of the alphabet should be determined.
- Some additional characters may be required. If so, these should be chosen from the ISO/IEC 8859-1 table. The alphabet of relative nations will be useful when selecting the new alphabet.
- The alphabet transformation should be completed as soon as possible.

References

[1] İ. Ergenç, "Konuşma Dili ve Türkçenin Söyleyiş Sözlüğü", Multilungual Yabancı Dil Yayınları, ISBN 975-6542-06-x, 2002

[2] M. Ş. Ülkütaşır, "Atatürk ve Harf Devrimi", TDK yayınları : 384,ISBN 975-16-0361-7, 2000

[3] B. N. Şimşir, "Türk Yazı Devrimi", TDK yayınları : ISBN:9751604206, 1992

С.Ж. МУСАЕВ, С. Ж.КАРАБАЕВА, А.И.ИМАНАЛИЕВА

Кыргызский государственный университет строительства, транспорта и архитектуры им. Н.Исанова, Бишкек, Кыргызстан

ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ РАЗВИТИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В КЫРГЫЗСТАНЕ

Сегодня важную роль в жизни современного общества играют автоматизированные информационные технологии. Все люди должны иметь возможность пользоваться преимуществами глобальных информационных ресурсов. В настоящее время ИНТЕРНЕТ состоит из миллиарда страниц информации и продолжает разрастаться. Однако при этом «всемирная паутина» представляет собой чрезвычайно демократичную среду, состоящую из неисчислимого количества Web-сайтов, создаваемых отдельными лицами и неформальными группами. Виртуальные сообщества людей, разбросанные по всему земному шару, но объединенные общими интересами, обсуждают буквально все, начиная с языков, находящихся под угрозой исчезновения, и кончая особенностями национальной культуры.

В XXI веке в мире формируется разделительная граница между странами, создающими информационное общество, и странами, отстающими в области информатизации. Новое разделение мира на развитые и отстающие страны - это «цифровое или информационное разделение», показывающее уровень информатизации государств. Поэтому государственные органы любой страны вынуждены принимать определенные меры, чтобы не попасть в группу «информационно отсталых» стран.

До 2000 года кыргызское языкознание развивалось практически без привлечения математических аппаратов и возможностей вычислительной техники для создания и изучения языковых моделей.

Начиная с 1990 года, Институт теоретической и прикладной математики Национальной академии наук Кыргызской Республики ведет работу по алгоритмизации кыргызского языка. Разработан и реализован на компьютере единый алгоритм словоизменения в кыргызском

языке, составлена контрольно-обучающая программа со случайным формированием заданий, которая используется в учебных заведениях.

В системах обработки знаний на естественном языке одной из центральных является задача разработки лингвистических ресурсов. Лингвистические ресурсы представляют собой базы данных, включающие концептуальную информацию в виде различных словарей, парадигматических и формальных моделей естественного языка, а также специализированые лингвистические процессоры обработки самой модели.

На сегодняшний день необходимый уровень решения проблем моделирования, процесса понимания смысла текстов и проблемы синтеза речи у нас пока еще не достигнут, хотя работы в области компьютерной лингвистики ведутся во всех развитых странах мира. Тем не менее, можно отметить серьезные научные и практические достижения в области компьютерной лингвистики.

В последние 10 лет существования суверенной Кыргызской Республики процесс информатизации общества вышел на новый уровень. Каждый год число пользователей глобальной сети ИНТЕРНЕТ увеличивается в геометрической прогрессии. Стремительно растет сам рынок информационных технологий, что в свою очередь вызывает рост потребности в соответствующих специалистах.

Лингвистика! Она должна повернуться лицом к новым задачам, выдвигаемым компьютеризацией. Компьютерная лингвистика в настоящий момент находится на подъеме в Кыргызстане. Для отечественных лингвистов недавно открылись те области приложения знаний о языке, которые традиционны для зарубежного сообщества профессиональных лингвистов. В Кыргызстане появляется спрос на лингвистов, работающих в области рекламы и в сфере публичной политики и в связи с этим появилось необходимость открытия в ВУЗах таких специальностей. С 2007 года в Кыргызском государственном университете строительства, транспорта и архитектуры им. Н. Исанова начала свою работу новая кафедра «Компьютерная лингвистика».

Кафедра компьютерной лингвистики Кыргызского государственного университета строительства, транспорта и архитектуры им. Н.Исанова основана в 2007 году и является выпускающей кафедрой, которая готовит специалистов в области компьютерной лингвистики применительно к кыргызскому языку. Нужных и востребованных специалистов на сегоднящний день. Так как на рынке труда специалисты по компьютерной лингвистике на сегодняшний день очень востребованы, а программы адресной подготовки компьютерных лингвистов необходимы.

Компьютерную лингвистику могут изучат люди с разным базовым образованием. Был бы у них интерес к растущим сейчас в цене и популярности лингвистическим технологиям, которые они смогут применить в уже знакомой или совсем новой для них области. Это могут быть и лингвисты, и математики, и социологи, и даже маркетологи — никаких специальных ограничений здесь нет.

На кафедре ведутся работы по созданию компьютерных лингвистических моделей для кыргызского языка и готовятся лингвистические ресурсы. В частности, подготовлен электронный словарь терминов по информационным технологиям на кыргызском языке.

При внедрении кыргызского языка в компьютерные технологии предполагается разработка пакетов прикладных программ для автоматизации профессиональной деятельности и для автоматизации обработки текстов на кыргызском языке, создание кыргызско-язычного интерфейса для пользовательских систем.

Под руководством директора института новых информационных технологий Бейшенбека Такырбашевича Укуева и корпорации Microsoft был переведен Windows-7, версия office -14 интерфейс на кыргызский язык.

Выпускники специальности «Компьютерная лингвистика» на проектировании дипломных проектов разрабатывают новые обучающие мультимедийные программы английского и кыргызского языков. А также студенты кафедры работают над разработкой перевода на

кыргызский язык сайтов Facebook, Twitter, Google и Wikipedia и улучшения качества их работы.

В свете очерченных перспектив, диктуемых жизнью назрела необходимость и в специальных отечественных учебных изданий, статьей и журналов по компьютерной лингвистике.

Поток зарубежных публикаций по вопросам компьютерной лингвистики огромен. Выходит много монографий, сборников, издаются журналы «Компьютерная лингвистика». Ежегодно проводится не менее десятка международных научных симпозиумов по компьютерной лингвистике, по машинному переводу, по применению компьютеров в управлении, в гуманитарных науках, в словарном деле, в обучении и т. д.

Проблема в том, что на сегодняшний очень нужны организация ежегодных конференций, симпозиумов и форумов по компьютерной лингвистике для обсуждения проблемы компьютеризации тюркских народов на ряду с лингвистикой, а также издание монографий, сборников и журналов в тюркоязычных странах. Евразийский национальный университет имени Л.Н. Гумилёва, Министерство образования и науки Республики Казахстан совместно с Академией наук Республики Татарстан организовали 1-ую Международную конференцию на тему "Компьютерная обработка тюркских языков. Латинизация письменности". Для того, чтобы компьтерная лингвистика имела активное применение и развитие было бы хорошо организация ежегодных конференций в тюркоязычных странах и это стало бы мощным инструментом, ведущим наши народы к научному прогрессу, который будет сближать тюркие народы. Сегодня в тюркоязычных странах делается немало по работе компьютерной лингвистике и современные ученые тюркского мира должны работать сообща над созданием и использованием терминов информационных технологий. Уже второй раз был проведен форум по теме «Стандартизация и унификация терминов информационных технологий тюркоязычных стран», который был организован Ассоциацией информатиков Турции господином Кораем Озером и профессором университета Хаджеттепе в Турции Шукру Халык Акалином.

В настоящее время перед кыргызскими учеными стоят проблемы преобразовании и компьютеризации кыргызского языка. Проблемы преобразовании кыргызского языка заключается в стандартизации орфографии кыргызского языка. Мы сталкиваемся с трудностями и сложностью при проведении семантического анализа текста или в машинном переводе на компьютере. При переводе парных слов «*ата мекен, бака жалбырак, көз мончок, үч бурчтук* и т.д.» с помощью машинного перевода компьютер переводит каждое слово отдельно например, *ата мекен – ата (отец)* и *мекен (родина)* при морфологическом разборе *үч бурчтук – үч (числительное)* и *бурчтук (существительное)*. Для устранение этих проблем нам приходится писать эти слова слитно при проведении семантического анализа слова, необходимо разработать такой механизм устранения орфографических норм кыргызского языка.

Для проведения научно-прикладных работ по созданию машинного фонда кыргызского языка, кыргызского речевого интерфейса, функционально-структурной модели кыргызских морфем как формальной и информационно-справочной базы при построении лингвопроцессоров, морфологических и синтаксических анализов, генерации текста, распознавания и синтеза речи, интерпретации смысла тестов, семантический поиск в интернете, разработок обучающих игровых программ на кыргызском языке необходимы немалые деньги. Эти проблемы в республике можно успешно осуществлять в рамках Государственных программ.

В республике на сегодня практически полностью выполнен комплекс организационных и директивных мероприятий и создана необходимая база для обеспечения функционирования кыргызского языка в компьютерных технологиях. Под руководством д. ф-м. н., профессора Павела Сергеевича Панкова, который работает над этой тематикой, выполняется немало работ по компьютеризации кыргызского языка.

Составлен полный список аффиксов в кыргызском языке, составлен словарь терминов по информационным технологиям на кыргызском языке.

Разработано новое понятие субъекта, как перманентно неустойчивого объекта такого, что малые по энергии внешние воздействия (команды) вызывают у него большие по энергии существенно различные реакции и изменения внутреннего состояния, и введено новое понятие языка, как системы таких команд.

Введено и реализовано на компьютере определение математической модели понятий, давшее возможность неязыкового независимого представления естественных языков.

На ряду с этими работами в 2002 году Э.Д. Асанов создал программу «Тамга-КИТ». Внедрение, дополнение и развитие этой программы стало главным путем в компьютеризации кыргызского языка.

Данная программа состоит из 20 компонентов, для того чтобы можно было использовать кыргызский язык. «Тамга-КИТ» облегчает работу на кыргызском языке и эта программа дала широкий путь в развитии политики государственного языка в информационной сфере.

Главная особенность программы «Тамга-КИТ» - это возможность проверять грамматику (орфографию, стилистику) текстов на кыргызском языке. Стратегия программы основано на социально-коммуникационном развитии государственного языка в сопровождение с современными информационными технологиями.

В 2011 году был создан языковый пакет «KyrSpell 2.2 - Проверка орфографии кыргызского языка для MS Office 97-2013» специально для кыргызского языка.

Данный комплект программ позволяет проверять на орфографические ошибки, находит синонимы, антонимы а также родственные слова (функция тезаурус), и кроме этого осуществлять расстановку переносов текстов на кыргызском языке в приложениях MS Office. Проверка орфографии осуществляется стандартными средствами, что обеспечивает проверку в любых приложениях, где существует соответствующая функция (например, в MS Word, MS Excel, MS PowerPoint, MS Outlook, Outlook Express и др.)

Для прогресса в автоматизации обработки кыргызского языка необходимо разработать следующее:

- стандартизация орфографии кыргызского языка;
- разработать и активно применять на практике корректор кыргызских текстов на основе генеративной морфологии, помогающий пользователю обнаруживать и исправлять ошибки в тексте, электронный русско-кыргызский словарь, словарь бытовых терминов, англо-кыргызско-русский и русско-кыргызский словники компьютерных терминов, электронные многоязычные словари, толковые словари и орфографические словари;
- разработать и реализовать на компьютере единый алгоритм словоизменения в кыргызском языке, создать акустическую базу данных кыргызского языка и аппаратная реализация распознавание и синтеза речи.

Таким образом когда мы сегодня говорим о компьютерной лингвистике, нужно понимать, что области применения лингвистических технологий стремительно расширяются. При этом следует придерживаться правила, что язык должен идти на ряду с техникой. В современную эпоху глобализации и формирования информационной цивилизации важно решение проблем развития единства тюркских народов и повышения статуса национальных языков и формирования единого информационного пространства тюркских народов.

Кыргызский и казахский языки, как и другие тюркские языки, относятся к типу агглютативных, и имеют стройные системы правил последовательного присоединения и написания окончаний, с малым числом исключений.

С переходом бывших тюркоязычных стран в составе СССР на латинский алфавит необходим переход национального языка к латинской системе алфавита. И необходимо создать **общетюркский единый алгоритм словоизменения в тюркских языках.**

Осуществление вышесказанных предложений способствовала бы сохранению и развитию каждого из тюркских народов, развитию тюркской цивилизации в системе глобальной цивилизации.

Ш.А.НАЗИРОВ, Х.Х.ХОМИДОВ, А.И.АЛНИЯЗОВ, К.С.РАХМАНОВ, А.З.МАХМУДОВ

Ташкентский университет информационных технологий, Ташкент, Узбекистан

ФОРМАЛИЗАЦИЯ КОНСТРУКЦИЙ ПРЕДЛОЖЕНИЙ УЗБЕКСКОГО, ТУРЕЦКОГО И КАРАКАЛПАКСКОГО ЯЗЫКОВ

При создании электронного переводчика естественных языков (в данном примере тюркские языкы) требуется формализации конструкций предложений.

В настоящее время нами были сравнительно анализированы формализация конструкций узбекских, турецких и каракалпакских языков. Подобным же образом, можно формализировать конструкций предложений и для других тюркских языков.

В данной статьи при формализации конструкций предложений рассмотрены 62 самые распространенные предложения узбекского, турецкого и каракалпакского языков.

В таблице 1 приведена сравнительная конструкция предложений узбекского, турецкого и каракалпакского языков [1-3].

Таблица 1

№	Тип предложений	Узбекское название	Турецкое название	Каракапакское название
1	Не распространенное предложение	**+ тўл:** Ø + **ҳол:** Ø + **кес:** Талабалар кетишди.	**özne** + tüml: Ø **+ zarf tüml**: Ø **+ yüklem**: Öğrenciler gittiler.	баслаўыш + толықлаўыш: Ø + пысықлаўыш: Ø + баянлаўыш: Талабалар кетти.
2	Распространенное предложение с определением	аниқ + эга + аниқ + тўл + аниқ + ҳол + кес: Университет талабалари берилган топшириқларни жиддий ёндашиб бажардилар.	**İ**sim tamlaması (özne) + sıfat tamlaması (nesne) + sıfat tamlaması + yüklem: Üniversite öğrencileri verilen ödevleri ciddi şekilde yaptılar.	анық. + басл. + анықл. + тол. + анық. + пысықл. + баянл: Университет талабалары берилген тапсырмаларды жуўапкерли жандасып орынлады.
3	Не распространенное предложение с определением	аниқ + эга + тўл: Ø **+** ҳол: Ø **+** аниқ; + кес: Университет битирувчилари талабгор кадрлардир.	**sıfatfiil** + özne + tümleç: Ø **+** zarf tümleci: Ø **+** sıfat; + yüklem: Üniversiteden mezun olan öğrenciler istenilen kadrodur.	анықл. + басл. + толықл..: Ø **+** пысықл.: Ø **+** анықл. + баянл.: Университет питкериўшилери талапкер кадрлар болып есапланады.
4	Распространенное обще-вопросительное предложение	эга + тўл + ҳол + кес: ми?: Талаба вазифаларини вақтида бажардими?	cümle özne + nesne + zarf tümleci + yüklem: mı?: Öğrenci ödevlerini zamanında yaptı mı**?**	басл. + толықл. + пысықл. + баянл.: ма?: Талаба ўазыйпаларын ўақтында орынлады ма?
5	Не распространенное обще-вопросительное предложение	эга + тўл: Ø **+** ҳол: Ø **+** кес: ми?: Талаба бажардими?	özne + tümleç: Ø **+** zarf tümleci: Ø **+ yüklem**: mı?: Öğrenci yaptı mı?	басл. + толықл: Ø **+** пысықл: Ø **+ баянл.**: ма?: Талаба орынлады ма?
6	Распространенное специально-вопросительное предложение	эга: с.олм + тўл: с.олм + ҳол: с.олм + кес?: Ким нимани қачон топширди?	özne: soru zamiri + nesne: soru zamiri + zarf tümleci: soru zamiri + yüklem?: Kim neyi ne zaman teslim etti**?**	басл.: сораў алмасығы + толықл.: сораў алмасығы + пысықл.: сораў алмасығы + баянл: Ким нени қ ашан тапсырды?
7	Не	эга: с.олм + тўл: Ø **+** ҳол: Ø **+**	özne: soru zamiri + tümleç: Ø	басл.: сораў алмасығы +

	распространенное специально-вопросительное предложение	кес Ким бажарди?	+ zarf tümleci: Ø + yüklem Kim yaptı?	толықл: Ø + пысықл.: Ø + баянлаўыш: Ким орынлады?
8	Распространенное негативное предложение	эга + тўл + ҳол + кес: ма: Талаба вазифаларини вақтида бажар + ма + ди.	özne + tümleç + zarf tümleci + yüklem: ma: Öğrenci ödevlerini zamanında yapma+dı?	басл. + толықл. + пысықл. + баянл.: ма: Талаба ўазыйпаларды ўақтында орынла + ма + ды.
9	Не распространенное негативное предложение	эга + тўл: Ø + ҳол: Ø + кес: ма: Талаба бажармади.	özne + tümleç: Ø + zarf tümleci: Ø + yüklem: ma: Öğrenci yapmadı.	басл. + толықл.: Ø + пысықл.: Ø + баянл.: ма: Талаба орынламады.
10	Распространенное негативное предложение с определением	аниқ. + эга + аниқ. + тўл + аниц + ҳол + кес: ма: Университет талабалари берилган вазифаларни жиддий ёндашиб бажар + ма + дилар.	sıf. + özne + sıf. + tümleç + sif. + zarf tümleci + yüklem: ma: Üniversite öğrencileri verilen ödevleri ciddi yaklaşımla yap+ma+dılar.	анықл. + басл. + анықл. + толықл. + анықл. + пысықл. + баянл.: ма: Университет талабалары берилген ўазыйпаларды жуўапкерли жандасып орынла + ма + ды.
11	Не распространенное негативное предложение с определением	аниқ + эга + тўл: Ø + ҳол: Ø + аниқ. + кес: ма: Университет битирувчилари малакали инженер эмас.	sıf. + özne + tümleç: Ø + zarf tümleci: Ø + sıfat. + yüklem: ma: Üniversite mezunları bilgili mühendis değiller.	анықл. + баянл. + толықл.: Ø + пысықл.: Ø + анықл. + баянл.: ма: Университет питкериўшилери кәнигели инженер емес.
12	Распространенное негативное обще-вопросительное предложение	эга + тўл + ҳол + кес: ма + ми?: Талаба вазифаларини вақтида бажармадими?	özne + tümleç + zarf tümleci + yüklem: ma + mı?: Öğrenci ödevlerini zamanında yapmadı mı?	басл. + толықл. + пысықл. + баянл.: ма + ма?: Талаба ўазыйпаларын ўақтында орынламады ма?
13	Не распространенное негативное обще-вопросительное предложение	эга + тўл: Ø + ҳол: Ø + кес: ма + ми?: Талаба бажармадими?	özne + tümleç: Ø + zarf tümleci: Ø + yüklem: ma + mi?: Öğrenci yapmadı mı?	басл. + толықл.: Ø + пысықл.: Ø + баянл.: ма + ма?: Талаба орынламады ма?
14	Распространенное негативное специально-вопросительное предложение	эга: с.олм + тўл: с.олм + ҳол: с.олм + кес: ма?: Ким нимани қачон бажармади?	özne: soru zamiri + tümleç: soru zamiri + zarf tümleci: soru zamiri + yüklem: ma?: Kim neyi ne zaman yapmadı?	басл.: сораў алмасығы + толықл.: сораў алмасығы + пысықл.: сораў алмасығы + баянл.: ма?: Ким нени қашан орынламады?
15	Не распространенное негативное специально-вопросительное предложение	эга: с.олм + тўл: Ø + ҳол: Ø + кес: ма?: Ким бажармади?	cümle özne: soru zamiri + tümleç: Ø + zarf tümleci: Ø + yüklem: ma?: Kim yapmadı?	басл.: сораў алмасығы + толықл.: Ø + пысықл.: Ø + баянл.: ма?: Ким орынламады?
16	Распространенное специально-вопросительное предложение с определением	эга: с.олм + тўл: Ø + ҳол: с.олм + аниқ + кес?: Ким қаерда малакали инженер?	özne: soru zamiri + yer tümleci: Ø + yer tümleci: soru zamiri + sıfat + yüklem?: Kim nerede bilgili mühendis?	басл.: с. алм + тол: Ø + пысықл.: с. алм + анық + баянл?: Ким қай жерде кәнигели инженер?
17	Не распространенное специально-вопросительное предложение с	эга: с.олм + тўл: Ø + Ҳол: Ø + аниқ + кес?: Ким катта ўқитувчи?	özne: soru zamiri + tümleç: Ø + Zarf tümleci: Ø + sıfat + yüklem?: Kim büyük öğretmen?	басл: с. алм + тол: Ø + пысықл.: Ø + анық + баянл?: Ким үлкен оқытыўшы?

	определением			
18	Распространенное общевопросительное предложение с определением	аниқ + эга + аниқ + тўл + аниқ + ҳол + кес: ми?: Университет талабалари берилган вазифаларни айтилган вақтида бажардиларми?	sıfat + özne + sıfat + tümleç + sıfat + zarf tümleci + yüklem: mi?: Üniversite öğrencileri verilen ödevleri belirtilen sürede yaptılar mı?	анықл. + басл. + анықл. + толықл. + анықл. + пысықл. + баял.: ма?: Университет талабалары берилген тапсырмаларды айтылған ўақытта орынламады ма?
19	Не распространенное общевопросительное предложение с определением	аниқ + эга + тўл: Ø + Ҳол: Ø + кес: ми?: Университет талабалари қатнашадими?	sıfat + özne + tümleç: Ø + Zarf tümleci: Ø + yüklem: mi?: Üniversite öğrencileri katılacaklar mı?	анықл. + басл. + толықл.: Ø + пысықл.: Ø + баянл.: ма?: Университет талабалары қатнаса ма?
20	Распространенное негативное специально-вопросительное предложение с определением	эга: с.олм + тўл: Ø + аниқ + ҳол + аниқ **+ кес: ма**?: Ким катта шаҳарда малакали инженер эмас?	özne: soru zamiri + tümleç: Ø + sıfat + zarf tümleci + sıfat **+ yüklem: ma**?: Kim büyük şehirde bilgili mühendis değil?	басл.: с. алм. + толықл.: Ø + анықл. + пысықл. + анықл. **+ баянл.: ма**?: Ким үлкен шәҳәрде кәнигели инженер емес?
21	Не распространенное негативное специально-вопросительное предложение с определением	эга: с.олм + тўл: Ø + ҳол: Ø + аниқ + кес: ма?: Ким малакали инженер эмас?	özne: soru zamiri + tümleç: Ø + zarf tümleci: Ø + sıfat + yüklem: ma?: Kim bilgili mühendis değil?	басл.: с. алм + толықл.: Ø + пысықл.: Ø + анықл. + баянл.: ма?: Ким кәнигели инженер емес?
22	Распространенное негативное общевопросительное предложение с определением	аниқ + эга + тўл: Ø + аниқ + *ҳол* + аниқ + кес: ма + ми?: Сенинг аканг университетда катта ўқитувчи эмасми?	sıfat + özne + tümleç: Ø + sıfat + *zarf tümleci* + sıfat + yüklem: ma + mi?: Senin ağabeyin üniversitede büyük öğretmen değil mi?	анықл. + басл. + толықл.: Ø + анықл. + ***пысықл.*** + анықл. + баянл.: ма + па/пе (ба/бе)?: Сениң әжағаң университетте үлкен оқытыўшы емес пе?
23	Не распространенное негативное общевопросительное предложение с определением	аниқ + эга + тўл: Ø + ҳол: Ø + аниқ + кес: ма + ми?: Сенинг аканг катта ўқитувчи эмасми?	sıfat + özne + tümleç: Ø + zarf tümleci: Ø + sıfat + yüklem: ma + mi?: Senin ağabeyin büyük öğretmen değil mi?	анықл. + басл. + толықл.: Ø + пысықл.: Ø + анықл. + баял.: ма + па (/пе, ба/бе)?: Сениң әжағаң үлкен оқытыўшы емес пе?
24	Распространенное сложносочиненное предложение	эга + (тўл V Ø) + (ҳол V Ø) + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирдик.	özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem: O (kitabı V Ø) (bugün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirdik.	басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл. бог басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл.: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирдик.
25	Не распространенное сложносочиненное предложение	эга + (тўл V Ø) + (ҳол V Ø) + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирдик. У олди ва биз келтирдик.	özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem: O (kitabı V Ø) (bugün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirdik. O	басл. + (толықл. V Ø) + (пысықл. V Ø) + баянл. дәнекер + басл. + (толықл. V Ø) + (пысықл. V Ø) + баянл.: Ол (китапты V Ø) (бүгин V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелдик. Ол

			aldı ve biz getirdik.	алды ҳәм бизлер әкелдик.
26	Распространенное негативное сложносочиненное предложение	эга + (тўл V Ø) + (ҳол V Ø) + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес + ма: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирмадик.	özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem + ma: O (kitabı V Ø) (bugün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirmedik.	гәп басл. + (толықл. V Ø) + (пысықл. V Ø) + баянл. + дәнекер + басл. + (толықл. V Ø) + (пысықл. V Ø) + баянл. + ма/ме: Ол (китапты V Ø) (бүгин V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелмедик.
27	Не распространенное негативное сложносочиненное предложение	эга + (тўл: Ø) + (ҳол: Ø) + кес бог эга + (тўл: Ø) + (ҳол: Ø) + кес + ма: У олди ва биз келтирмадик.	özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem bog özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem + ma: O aldı ve biz getirmedik.	басл. + (толықл.: Ø) + (пысықл.: Ø) + баянл. + дәнекер + басл. + (толықл.: Ø) + (пысықл.: Ø) + баянл. + ма/ме: Ол алды ҳәм бизлер әкелмедик.
28	Распространенное специально-вопросительное сложносочиненное предложение	гап (эга V эга: с.олм) + (тўл: с.олм V тўл V Ø) + (ҳол: с.олм V ҳол V Ø) + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес?: (У V Ким) (нимани У компьютерни У Ø) (қачон V кеча V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирдик?	(özne V özne: soru zamiri) + (tümleç: soru zamiri V tümleç V Ø) + (zarf tümleci: soru zamiri V zarf tümleci V Ø) + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem?: (O V Kim) (neyi O bilgisayarı O Ø) (ne zaman V dün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirdik?	(басл. V басл.: с. алм.) + (толықл.: с. алм V толықл. V Ø) + (пысықл.: с. алм. V пысықл. V Ø) + баянл. + дәнекер + басл. + (толықл. V Ø) + (пысықл. V Ø) + баянл.?: (Ол V ким) (нени? компьютерди Ø) (қашан V кеше V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелдик?
29	Не распространенное специально-вопросительное сложносочиненное предложение	эга: с.олм + (тўл: Ø) + (ҳол: Ø) + кес бог эга + (тўл: Ø) + (ҳол: Ø) + кес?: Ким олди ва биз келтирдик?	özne: soru zamiri + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem bog özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem?: Kim aldı ve biz getirdik?	басл.: с. алм. + (толықл.: Ø) + (пысықл.: Ø) + баянл. + дәнекер + басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл.?: Ким алды ҳәм бизлер әкелдик?
30	Распространенное негативное специально-вопросительное сложносочиненное предложение	гап (эга V эга: с.олм) + тўл: с.олм + ҳол: с.олм + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес + ма: (У V Ким) (нимани V компьютерни V Ø) (қачон V кеча V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирмадик?	(özne V özne: soru zamiri) + tümleç: soru zamiri + zarf tümleci: soru zamiri + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem + ma: (O V Kim) (neyi V bilgisayarı V Ø) (ne zaman V dün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirmedik?	(басл. V басл.: с. алм) + пысықл.: с. алм + пысықл.: с. алм. + баянл. + дәнекер + басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл. + ма: (Ол V ким) (нени V компьютерди V Ø) (қашан V кеше V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелмедик?
31	Не распространенное негативное специально-вопросительное сложносочиненное предложение	эга: с.олм + тўл: Ø + ҳол: Ø + кес бог эга + (тўл: Ø) + (ҳол: Ø) + кес + ма: Ким олди ва биз келтирмадик?	özne: soru zamiri + tümleç: Ø + zarf tümleci: Ø + yüklem bog özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem + ma: Kim aldı ve biz getirmedik?	басл.: с. алм. + пысықл.: Ø + пысықл.: Ø + баянл. + дәнекер + басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл. + ма: Ким алды ҳәм бизлер әкелмедик?
32	Распространенное	эга + (тўл V Ø) + (ҳол V Ø) +	özne + (tümleç V Ø) + (zarf	басл. + (пысықл. V Ø) +

	общее-вопросительное сложносочиненное предложение	кес боғ эга + (тўл V Ø) + (ҳол V Ø) + кес + ми?: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирдикми?	tümleci V Ø) + yüklem bog‘ özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem + mi?: O (kitabı V Ø) (bugün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirdik mi?	(пысықл. V Ø) + баянл. + дәнекер + басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл. + па/пе (ма/ме)?: Ол (китапты V Ø) (бүгин V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелдик пе?
33	Не распространенное обще-вопросительное сложносочиненное предложение	эга + (тўл: Ø) + (Ҳол: Ø) + кес боғ эга + (тўл: Ø) + (ҳол: Ø) + кес + ми?: У олди ва биз келтирдикми?	özne + (tümleç: Ø) + (Zarf tümleci: Ø) + yüklem bog‘ özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem + mi?: O aldı ve biz getirdik mi?	басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл. + дәнекер + басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл. + па/пе (ма/ме, ба/бе)?: Ол алды ҳәм бизлер әкелдик пе?
34	Распространенное негативное обще-вопросительное сложносочиненное предложение	эга + (тўл V Ø) + (ҳол V Ø) + кес бог эга + (тўл V Ø) + (ҳол V Ø) + кес + ма + ми?: У (китобни V Ø) (бугун V Ø) олди ва биз (уни V Ø) (уйимизга V Ø) келтирмадикми?	özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem bog özne + (tümleç V Ø) + (zarf tümleci V Ø) + yüklem + ma + mi?: O (kitabı V Ø) (bugün V Ø) aldı ve biz (onu V Ø) (evimize V Ø) getirmedik mi?	басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл. + дәнекер + басл. + (пысықл. V Ø) + (пысықл. V Ø) + баянл. + ма + па/пе (ма/ме, ба/бе)?: Ол (китапты V Ø) (бүгин V Ø) алды ҳәм бизлер (оны V Ø) (үйимизге V Ø) әкелмедик пе?
35	Не распространенное негативное обще-вопросительное сложносочиненное предложение	эга + (тўл: Ø) + (ҳол: Ø) + кес бог эга + (тўл: Ø) + (ҳол: Ø) + кес + ма + ми?: У олди ва биз келтирмадикми?	özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem bog özne + (tümleç: Ø) + (zarf tümleci: Ø) + yüklem + ma + mi?: O aldı ve biz getirmedik mi?	басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл. + дәнекер + басл. + (пысықл.: Ø) + (пысықл.: Ø) + баянл. + ма + па/пе (ма/ме, ба/бе)?: Ол алды ҳәм бизлер келтирмедик пе?
36	Повествовательное А: «П».	эга + {тўл V Ø} + {ҳол V Ø} + кес: «эга + {тўл V Ø} + {ҳол V Ø} + кес»: Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди: «Мен {дарсга V Ø} {эртага V Ø} бораман».	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «Ben {derse V Ø} {yarın V Ø} gideceğim».	басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «Мен {сабаққа V Ø} {ертең V Ø} бараман».
37	Негативное А: «П».	эга + {тўл V Ø} + {ҳол V Ø} + кес : «эга + {тўл V Ø} + {ҳол V Ø} + кес + ма»: Ўқитувчи {талабаларга V Ø} {кеча VØ} деди: «У {китобни V Ø} {эртага V Ø} олиб бормайди».	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem : «özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem + ma»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «O {kitabı V Ø} {yarın V Ø} götürmeyecek».	». басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. : «басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. + ма»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «Ол {китапты V Ø} {ертең V Ø} алып бармайды».
38	Общее вопросительное А: «П?».	эга + {тўл V Ø} + {ҳол V Ø} + кес : «эга + {тўл V Ø} {ҳол V Ø} + кес + ми?»: Ўқитувчи	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem : «özne + {tümleç V Ø} {zarf tümleci V	?» басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. : «басл. + {пысықл. V Ø} {пысықл. V

		{талабаларга V Ø} {кеча V Ø} деди: «Сен {дарсга V Ø} {эртага V Ø} борасанми?»	Ø} + yüklem + mi?»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «Sen {derse V Ø} {yarın V Ø} gidecek misin?».	Ø} + баянл. + ба (/бе, ма/ме па/пе)?»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «Сен {сабаққа V Ø} {ертең V Ø} барасаң ба?»
39	Специально вопросительное А: «П?».	эга + {тўл V Ø} + {ҳол V Ø} + кес : «{эга : с.олм V эга} + {тўл : с.олм V тўл V Ø} + {ҳол : с.олм V ҳол V Ø} + кас?»: Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди: «{Ким V У} {нимани V китобни V Ø} {қачон V эртага V Ø} олиб боради?»	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem : {özne : soru zamiri V özne} + {tümleç : soru zamiri V tümleç V Ø} + {zarf tümleci : soru zamiri V zarf tümleci V Ø} + kas?»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «{Kim V Ø} {neyi V kitabı V Ø} {ne zaman V yarın V Ø} götürecek?»	басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. : «{басл. : с. алм. V басл.} + {пысықл. : с. алм. V пысықл. V Ø} + {пысықл. : с. алм. V пысықл. V Ø} + баянл?»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «{Ким V Ол} {нени V китапты V Ø} {қашан V ертең V Ø} алып барады?»
40	Обще – вопросительно - негативное А: «П?».	эга + {тўл V Ø} + {ҳол V Ø} + кес : «эга + {тўл V Ø}{ҳол V Ø} + кес + ма + ми?»: Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди: «Сен {дарсга V Ø} {эртага V Ø} бормайсанми?»	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem : «özne + {tümleç V Ø}{zarf tümleci V Ø} + yüklem + ma + mi?»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «Sen {derse V Ø} {yarın V Ø} gitmeyecek misin?»	басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. : «басл. + {пысықл. V Ø}{пысықл. V Ø} + баянл. + ма + ми?»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «Сен {сабаққа V Ø} {ертең V Ø} бармайсаң ба?»
41	Специально – вопросительно - негативное А: «П?».	эга + {тўл V Ø} + {ҳол V Ø} + кес: «{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V Ҳол V Ø} + кес + ма?»: Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди: «{Ким V У} {нимани V китобни V }{қачон V эртага V} олиб бормайди?»	özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V Zarf tümleci V Ø} + yüklem + ma?»: Öğretmen {öğrencilere V Ø} {dün V Ø} dedi: «{Kim V Ø} {neyi V kitabı V} {ne zaman V yarın V} götürmeyecek?»	басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V Пысықл. V Ø} + баянл. + ма?»: Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты: «{Ким V Ол} {нени V китапты V }{қашан V ертең V} алып бармайды?»
42	Повествовательное «П», - А.	«эга + {тўл V Ø} + {ҳол V Ø} + кес», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «Мен {дарсга V Ø} {эртага V Ø} бора-ман», - Ўқитувчи {талабаларга V Ø} {кеча V Ø} айтди.	«özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem», - özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «Ben {derse V Ø} {yarın V Ø} gideceğim», - Öğretmen {öğrencilere V Ø} {dün V Ø} söyledi.	«басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «Мен {сабаққа V Ø} {ертең V Ø} бара-ман», - Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты.
43	Негативное «П», - А.	«эга + {тўл V Ø} + {ҳол V Ø} + кес + ма», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «Мен {дарсга V Ø} {эртага V Ø} бормайман» - Ўқитувчи	«özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem + ma», - özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «Ben {derse V Ø} {yarın V Ø}	«басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. + ма», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «Мен {сабаққа V Ø} {ертең V Ø}

		{талабаларга V Ø} {кеча V Ø айтди.	gitmeyeceğim», - Öğretmen {öğrencilere V Ø} {dün V Ø} söyledi.	бармайман» - Оқытыўшы {талабаларға V Ø} {кеше V Ø айтты.
44	Обще-вопросительное «П?» -А.	«эга + {тўл V Ø} + {ҳол V Ø} + кес + ми?», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «Сен {дарсга V Ø} {эртага V Ø} борасанми?», - Ўқитувчи {талабаларга V Ø} {кеча V Ø} айтди.	«özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem + mi?», - özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «Sen {derse V Ø} {yarın V Ø} gidecek misin?», - Öğretmen {öğrencilere V Ø} {dün V Ø} söyledi.	«басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл. + ба/бе (па/пе, ма/ме)?», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «Сен {сабаққа V Ø} {ертең V Ø} барасаң ба?», - Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты.
45	Специально – вопросительно «П?» -А.	«{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол V Ø} + кес», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «{Ким V У} {нимани V китобни V Ø} {қачон V эртага V Ø} олиб боради?» - Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди.	.«{zga: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø} + yüklem», - özne + {tümleç V Ø} + {zarf tümleci VØ} + yüklem: «{Kim V U} {neyi V kitabı V Ø} {ne zaman V yarın V Ø} götürecek?» - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi.	«{зга: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø} + баянл.», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «{Ким V Ол} {нени V китапты V Ø} {қашан V ертең V Ø} алып барады?» - Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты.
46	Обще – вопросительно – негативное «П?» -А.	«эга + {тўл V Ø} + {ҳол V Ø} + кес + ма + ми?», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «Сен {дарсга V Ø} {эртага V Ø} бормайсанми?», - Ўқитувчи {талабаларга V Ø} {кеча V Ø} айтди.	«özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem + ma + mi?», - özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «Sen {derse V Ø} {yarın V Ø} gitmeyecek misin?», - Öğretmen {öğrencilere V Ø} {dün V Ø} söyledi.	«басл. + {пысықл. V Ø} + {пысықл. VØ} + баянл. + ма + ми?», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «Сен {сабаққа V Ø} {ертең V Ø} бармайсаң ба?», - Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты.
47	Специально – вопросительно - негативное «П?» -А.	«{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол V Ø} + кес + ма», - эга + {тўл V Ø} + {ҳол V Ø} + кес: «Ким V У} {нимани V китобни V } {қачон V эртага V } олиб бормайди?», - Ўқитувчи {талабаларга V Ø} {кеча V Ø} деди.	«{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø} + yüklem + ma», - özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem: «{Kim V U} {neyi V kitabı V} {ne zaman V yarın V} götürmeyecek?» - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi.	«{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø} + баянл. + ма», - басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл.: «Ким V Ол} {нени V китапты V } {қашан V ертең V } алып бармайды?», - Оқытыўшы {талабаларға V Ø} {кеше V Ø} айтты.
48	Повествовательное А: «П», -А.	эга + {тўл V Ø} + {ҳол V Ø}: «эга + {тўл V Ø} + {ҳол V Ø}: кес», - кес: Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø}: «Улар {вазифаларни V	özne + {tümleç V Ø} + {zarf tümleci V Ø}: «özne + {tümleç V Ø} + {zarf tümleci V Ø}: yüklem», - yüklem: Öğretmen {öğrencilere V	басл. + {пысықл. V Ø} + {пысықл. V Ø}: «басл. + {пысықл. V Ø} + {пысықл. V Ø}: баянл.», - баянл.: Оқытыўшы {оқыўшыларға

		Ø} {эртага V Ø} топширади», - деди.	Ø} {dün V Ø}: «Onlar {ödevleri V Ø} {yarın V Ø} teslim edecek», - dedi.	V Ø} {кеше V Ø}: «Олар {ўазыйпаларды V Ø} {ертең V Ø} тапсырады», - деди.
49	Негативное А: «П», -А.	эга + {тўл V Ø} + {ҳол\/ Ø}: «эга + {тўл V Ø} + {ҳол V Ø}: кес + ма», - кес: Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø}: «Улар {вазифаларни V Ø} {эртага V Ø} топширмайди», -деди.	özne + {tümleç V Ø} + {zarf tümleci\/ Ø}: «özne + {tümleç V Ø} + {zarf tümleci V Ø}: yüklem + ma», - yüklem: Öğretmen {öğrencilere V Ø} {dün V Ø}: «Onlar {ödevleri V Ø} {yarın V Ø} teslim etmeyecek», - dedi	басл. + {пысықл. V Ø} + {пысықл.\/ Ø}: «басл. + {пысықл. V Ø} + {пысықл. V Ø}: баянл. + ма», - баянл.: Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø}: «Олар {ўазыйпаларды V Ø} {ертең V Ø} тапсырмайды», -деди.
50	Обще – вопросительное А: «П?», -А.	эга + {тўл V Ø} + {ҳол V Ø}: «эга + {тўл V Ø} + {ҳол V Ø}: кес + ми?», - кес: Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø}: «Улар {вазифаларни V Ø} {эртага V Ø} топширади-ми?», - деди.	özne + {tümleç V Ø} + {zarf tümleci V Ø}: «özne + {tümleç V Ø} + {zarf tümleci V Ø}: yüklem + mi?», - yüklem: Öğretmen {öğrencilere V Ø} {dün V Ø}: «Onlar {ödevleri V Ø} {yarın V Ø} teslim edecek mi?», - dedi.	басл. + {пысықл. V Ø} + {пысықл. V Ø}: «басл. + {пысықл. V Ø} + {пысықл. V Ø}: баянл. + ма?», - баянл.: Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø}: «Олар {ўазыйпаларды V Ø} {ертең V Ø} тапсыра ма?», - деди.
51	Специально – вопросительное А: «П?», -А.	эга + {тўл V Ø} + {ҳол V Ø}: «{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол V Ø}: + кес», - кес: Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø}: «{Ким V Улар} {нимани V вазифаларни V Ø} {қачон V эртага V Ø}топширади?», -деди.	özne + {tümleç V Ø} + {zarf tümleci V Ø}: «{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø}: + yüklem», - yüklem: Öğretmen {öğrencilere V Ø} {dün V Ø}: {Kim V Onlar} {neyi V ödevleri V Ø} {ne zaman V yarın V Ø} teslim edecek?», - dedi.	басл. + {пысықл. V Ø} + {пысықл. V Ø}: «{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø}: + баянл.», - баянл.: Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø}: «{Ким V Олар} {нени V ўазыйпаларды V Ø} {қашан V ертең V Ø}тапсырады?», - деди.
52	Обще – вопросительно – негативное А: «П?», -А.	эга + {тўл V Ø} + {ҳол V Ø}: «эга + {тўл V Ø} + {ҳол V Ø}: кес + ма + ми?», - кес: Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø}: «Улар {вазифаларни V Ø} {эртага V Ø} топширмайдими?», - деди.	özne + {tümleç V Ø} + {zarf tümleci V Ø}: «özne + {tümleç V Ø} + {zarf tümleci V Ø}: yüklem + ma + mi?», - yüklem: Öğretmen {öğrencilere V Ø} {dün V Ø}: «Onlar {ödevleri V Ø} {yarın V Ø} teslim etmeyecek mi?», - dedi.	басл. + {пысықл. V Ø} + {пысықл. V Ø}: «басл. + {пысықл. V Ø} + {пысықл. V Ø}: баянл. + ма + ма (/ме, ба/бе, па/пе)?», - баянл.: Оқытыўшы {оқыўшы оқыўшыларға V Ø} {кеше V Ø}: «Олар {ўазыйпаларды V Ø} {ертең V Ø} ттапсырмай ма?», - деди.
53	Специально – вопросительно - негативное А: «П?», -А.	эга + {тўл V Ø} + {ҳол V Ø}: «{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол VØ}: + кес + ма», - кес: Ўқитувчи {ўқувчиларга V Ø}{кеча V	özne + {tümleç V Ø} + {zarf tümleci V Ø}: «{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø}: + yüklem +	басл. + {пысықл. V Ø} + {пысықл. V Ø}: «{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø}: + баянл. + ма», - баянл.: Оқытыўшы

		Ø}: «{Ким V Улар} {нимани V вазифаларни V Ø} {қачон V эртага V Ø} топширмайди?» - деди.	ma», - yüklem: Öğretmen {öğrencilere V Ø} {dün V Ø}: {Kim V Onlar} {neyi V ödevleri V Ø} {ne zaman V yarın V Ø} teslim etmeyecek?», - dedi.	{оқыўшыларға V Ø}{кеше V Ø}: «{Ким V Олар} {нени V ўазыйпаларды V Ø} {қашан V ертең V Ø} тапсырмайды?» - деди.
54	Повествовательное «П, -А -П»	«эга + {тўл V Ø} + {ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, -кес».: Улар {вазифаларни V Ø} {эртага V Ø}, - Ўқитувчи {ўқувчиларга V Ø}{кеча V Ø}деди, - топширади».	«özne + {tümleç V Ø} + {zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, -yüklem».: Onlar {ödevleri V Ø} {yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim edecek».	«басл. + {пысықл. V Ø} + {пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл., -баянл.».: Олар {ўазыйпаларды V Ø} {ертең V Ø}, - Оқытыўшы {оқыўшыларға V Ø}{кеше V Ø} айтты, - тапсырады».
55	Негативное «П, -А -П»	«эга + {тўл V Ø} + {ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, - кес + ма»: «Улар {вазифаларни V Ø}{эртага V Ø}, - Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø} деди, - топширмайди».	«özne + {tümleç V Ø} + {zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, - yüklem + ma»: Onlar {ödevleri V Ø} {yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim etmeyecek».	«басл. + {пысықл. V Ø} + {пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл., - баянл. + ма»: «Олар {ўазыйпаларды V Ø}{ертең V Ø}, - Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø} айтты, - тапсырмайды».
56	Обще – вопросительное «П, -А –П?»	«эга + {тўл V Ø} + {ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, - кес + ми?»: «Улар {вазифаларни V Ø} {эртага V Ø}, - Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø} деди, -топширадими?»	«özne + {tümleç V Ø} + {zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, - yüklem + mi?»: «Onlar {ödevleri V Ø} {yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim edecek mi?»	«басл. + {пысықл. V Ø} + {пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл., - баянл. + ма?»: «Олар {ўазыйпаларды V Ø} {ертең V Ø}, - Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø} айтты, -тапсыра ма?»
57	Специально – вопросительное «П, -А –П?»	«{эга: с.олм V эга} + {тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, - кес?»: «{Ким V Улар} {нимани V вазифаларни V Ø}{қ ачон V эртага V Ø}, - Ўқ итувчи {ўқ увчиларга V Ø} {кеча V Ø} деди, - топшираді?»	«{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, - yüklem?»: «{Kim V Onlar} {neyi V ödevleri V Ø} {ne zaman V yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim edecek?»	«{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл., - баянл.?»: «{Ким V Олар} {нени V ўазыйпаларды V Ø}{қашан V ертең V Ø}, - Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø} айтты, - тапсырады?»
58	Обще – вопросительно – негативное «П, -А – П?»	«эга + {тўл V Ø} + {Ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, - кес + ма + ми?»:	«özne + {tümleç V Ø} + {Zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, - yüklem + ma + mi?»:	«басл. + {пысықл. V Ø} + {Пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V

		«Улар {вазифаларни V Ø} {эртага V Ø}, - Ўқитувчи ўқувчиларга V Ø} {кеча V Ø}деди, - топширмайдими?»	«Onlar {ödevleri V Ø} {yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim etmeyecek mi?»	Ø} + баянл., - баянл. + ма + ма (/ме, ба/бе па/пе)?»: «Олар {ўазыйпаларды V Ø} {ертең V Ø}, - Оқытыўшы оқыўшыларға V Ø} {кеше V Ø}айтты, - тапсырмай ма?»
59	Специально – вопросительно - негативное «П, -А – П?»	«{эга: с.олм V эга} +{тўл: с.олм V тўл V Ø} + {ҳол: с.олм V ҳол V Ø}: эга + {тўл V Ø} + {ҳол V Ø} + кес, - кес + ма?»: «{Ким V Улар} {нимани V вазифаларни V Ø} {қачон V эртага V Ø}, - Ўқитувчи {ўқувчиларга V Ø} {кеча V Ø} деди, - топширади?»	«{özne: soru zamiri V özne} + {tümleç: soru zamiri V tümleç V Ø} + {zarf tümleci: soru zamiri V zarf tümleci V Ø}: özne + {tümleç V Ø} + {zarf tümleci V Ø} + yüklem, - yüklem + ma?»: «{Kim V Onlar} {neyi V ödevleri V Ø} {ne zaman V yarın V Ø}, - Öğretmen {öğrencilere V Ø} {dün V Ø} dedi, - teslim edecek?»	«{басл.: с. алм. V басл.} + {пысықл.: с. алм. V пысықл. V Ø} + {пысықл.: с. алм. V пысықл. V Ø}: басл. + {пысықл. V Ø} + {пысықл. V Ø} + баянл., - баянл. + ма?»: «{Ким V Олар} {нени V ўазыйпаларды V Ø} {қашан V ертең V Ø}, - Оқытыўшы {оқыўшыларға V Ø} {кеше V Ø} айтты, - тапсырады?»
60	Косвенная речь	Кўчирма гап	Dolaylı anlatım	Басқаның гәпи
61	Сложноподчиненное предложение	Эргаш гапли қўшма гап	Birleşik tâbilik cümlesi	Бағыныңқы гәпли қоспа гәп
62	Распространенное предложение (аксиома)	ёйиқ гап (аксиома)	Dağınık cümle (aksioma)	Кеңейтилген гәп (аксиома)

На базе конструкций предложений узбекского, турецкого и каракалпакского языков нами построени математические модели этих языков на основе аксиоматической теории.

На базе выше приведенных конструкций предложений узбекских, турецких и каракалпакских языков (узбекский - каракалпакский, каракалпакский - узбекский, узбекский – турецкий, турецкий – узбекский, турецкий – каракалпакский, каракалпакский - турецкий) созданы трехъязычные электронные словари. Разработана программное обеспечение электронного переводчика для высше указанных тюркских языков. Система управления база данных реализована на базе MySQL [2].

Литература

1. Пўлатов А.Қ. Компьютер лингвистикаси. А.Қ.Пўлатов; масъул муҳаррир: А.А.Абдуазизов, М.М.Орипов. –Т.: Akademnashr, 2011. – 520.
2. Hamidov X. Turk tili. Morfologiya, Toshkent, 2011. 188.
3. Shabanov J., Hamidov X., Turk tilining imlo qoidalari. Türkçe Yazım Kuralları, Toshkent, 2010. – 88.
4. Ҳозирги қорақалпоқ адабий тилининг грамматикаси. Сўз ясалиши ва морфология. - Нукус: "Билим", 1994. - 5-94 бб.
5. Қозоқбоева А.Т. Қорақалпоқ тилида қўшимчаларнинг вариантлилиги. Филол.ф.номзоди дис.... автореферати. - Нукус, 2010.
6. Назиров Ш.А., Рахманов К.С., Махмудов А.З. Классификация и построения базы данных словарей по тюркских языков // Международная конференция «Актуальные проблемы прикладной математики, информатики и механики» 17 – 19 сентября 2012 г. г.Воронеж. С. 110-116.

GULILA ALTENBEK

1.College of Information Science and Engineering , Xinjiang University
2. The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center Minority Languages
3.Xinjiang Laboratory of Multi-language Information Technology
Urumqi, Xinjiang , 830046, P.R. China,

IDENTIFICATION OF THE KAZAKH BASIC PHRASES BASED ON THE MAXIMUM ENTROPY MODEL

Abstract: This paper proposed the definition, classification and structure of the Kazakh basic phrases, and established a framework for the classification of it according to their syntactic functions. Meanwhile, the structure of the Kazakh basic phrases were analyzed; and the determination of the Kazakh basic phrases collocation and extraction of the Kazakh basic phrases based on rules were followed. The Maximum Entropy (ME) model uses for the identification of the phrases from texts and achieved a result of automatic identification of Kazakh phrases with an accuracy of 81.58% based on rules System and additional artificial modification. Design feature of this ME model join rely on templates of Kazakh Word, part of speech, affixes. Experimental results show that the accuracy rate reached 91.62**%**.

Key words: Kazakh basic phrase; phrase identification ; maximum entropy; rules.

1 Introduction

Automatic phrase identification is an important task in natural language processing. A phrase is a group of words that work together. Phrase recognition is a grammatical unit agent between words and sentences in natural language processing. Phrase identification Parser has been developed for different languages, and they include the Church's Base NP Recognition for English[1] etc. The rule-based Model and Maximum Entropy Model (ME) are the most commonly used technology for phrase representation and parsing.

Kazakh Language belongs to the Turkish Language group in the Altaic language family, and it is an agglutinative language with word structures formed by adding derivational or inflectional affixes to root words. Phrases identification is an important task in Kazakh information processing, our group put forward Kazakh morphological analysis which contains stem extraction, part of speech tagging(POS), spellchecking, etc. in the past few years .Syntax Parsing, analysis of phrase structure, automatic identification of phrases and depth analysis of structure was recently investigated.. In this paper, we focus on identifying noun phrases, adjective phrase and verb phrases, which are the most difficult aspects of Kazakh phrase recognition analysis using by rules and ME.

2 Related works

There was a variety of techniques used for phrase recognition. Which include rule-based technique, statistical technique, or a combination of rule-based and statistical techniques. For example, Church's English Base NP Recognition(1988) [1]. His approach is through manual or semi-automatic annotation phrase corpus as a training corpus, then any pair of speech tags phrase context information in the statistics of probability, which according to the above probability words in the sentence made between any two adjacent markers, to obtain analysis results. This approach followed in shallow parsing has been widely used.

Moreover, several main approaches or algorithms to phrase recognition was investigated, typically implemented using a Chunk parsing for statistics model to decide the boundary[3] . Chunk

parsing were first introduced by Abney (1991)[2], which is the most widely used syntactic parsing. The main idea of chunk parsing lies in seeking the appropriate breakthrough point, and decomposing the full parsing problems to syntax topology statistics structure and syntactic relation. Jun Zhao and Changning Huang are pioneers in Chinese phrase studies (1998)[4]; Tsinghua University had also completed its TCT (Tsinghua Chinese Trebank) for Chinese (Qiang Zhou,2004)[5]; many language studies of it had been used Kazakh phrase recognition parsing(Gulila.A etc.,2009))[6-7] .

Maximum Entropy was first introduced to NLP area by Berger, et al (1996) and Della Pietra, et al. 1997[8].which is an extremely flexible technique for linguistic modelling , since it can use a virtually unrestricted and rich feature set in the framework of a probability model. It is a conditional, discriminative model and allows for mutually dependent variables[9].

3 Kazakh Phase parsing

3.1 Kazakh Morphology

Morphological analysis is an important task in natural language processing research, which was developed for different languages, included the Porter Stemmer for English[10], PC Kimmo for Finnish[11],Oflazer(1994) and Gülşen,E.(2004)for Turkish[12-13], Beesley,K.R. (1996) [14]for Arabic,

The Kazakh morphological system uses a large number of suffixes and a small number of prefixes. Every word has a root, or a stem[15].

The root is the core of the entire word structure and it conveys its basic meaning.

A stem is a new word generated by adding zero or more various affixes to the root, and it expresses the complete meaning of the word.

Affixes are divided into inflectional affixes and derivational affixes. Inflectional affixes, when they are added to a word, they do not cause grammatical changes, and do not lead to meaning changes . Derivational affixes change the meaning of the word when added to a root word.

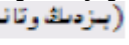
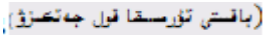
3.2 The Categories of Kazakh Phrase

Parsing is one of the most basic and fundamental component in natural language processing. Much research on parsing focused on their languages. Chunk parsing (or called shallow parsing) intends to obtain a fragment without thinking deeply.

According to certain rules, a Kazakh phrase is composed of two or more words in the vocabulary and grammatical meaning of the word structure associated with the language unit. In the Kazakh language, word and its structures are compounded according to certain rules of combination on the other hand certain structural rules are decided by certain grammatical relations. Furthermore, relationships between words are constrained by the syntax.

From the central parts of speech to points, a Kazakh phrase is a syntactic unit composed of two or more content words. In each phrase, there is only one head word, since other ingredients or adjuncts complement the role of a phrase as the head word .

In the case of Kazakh, Kazakh phrases can be divided into fixed phrases and temporary phrases by the meaning of phrase. Fixed phrases were formed in history, and are used as a word in the sentence. A fixed phrase includes the fixed phrase idioms, such as ات ۇستى قاراڭ (not to think highly of), temporary phrases, such as a noun phrase

 (our Motherland), and verb phrases such as  (get a happy life) etc.

According to computational linguistics, the basic phrase, is a non-nested phrase which does not intersect the structure of a word. It can only belong to one phrase and each phrase inside can no longer contain other phrases. According to syntax function, the Kazakh phrase divides noun phrases and verb phrases. Which can further divide them by the function of phrases as shown in table 1.

Table 1 Part of Kazakh phrase categories

NO	Category	Explanation	Example

1	NP	noun phrase	التىن كۇز
2	VP	verb phrase	مۇراتقا جەتۋ
3	AP	adjective phrase	تاپ- تازا
4	RP	pronouns phrase	ونىڭ مۇراتى
5	MCP	numeral phrases	سەگىز توعىز مىڭ
6	MP	a quantifier phrase	جيىرما جاس

Kazakh language is rich in the external morphology, and this forms the most prominent manifestations of phrase structure. Kazakh phrase can be divided into parallel structure (سالالاس قۇرىلىمى) , consistent with the structure (قيىسۋ قۇرىلىمى) , the dominant structure (مەڭگەرۋ قۇرىلىمى) , genitive structure (ماتاسۋ قۇرىلىمى) , additional structure (قابىسۋ قۇرىلىمى) , adjacent structures (جاناسۋ قۇرىلىمى) . Such as : بيىل ءبىزدىڭ اۋىلىمىز استىقتان مول ءونىم الدى. (we got grain crops harvest at hometown in this year.) show in fig. 1

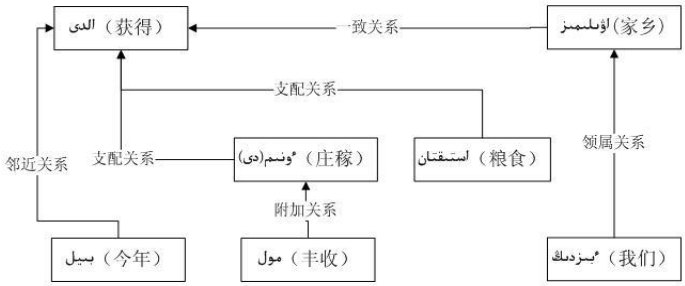


Figure 1: Example of Kazakh phrase structure

4 Statistics and Analysis of Kazakh phrase structure

Referring to modern Kazakh grammar[15-16], the basic rules of phrase structure of Kazakh language was summed up, which extracted the structure of phrases from the corpus, and created a set of rules.

In the representation of basic phrase structures, the part of speech tagging symbols in XML documents of Kazakh corpus was used which are v (verb), n. (noun), adj. (adjective) , ono. (onomatopoeia) , pron. (pronoun) , exc.(exclamation), num. (number), adv. (adverb), au.(auxiliary).

4.1 Kazakh Verb phrase structure

A Verb phrase has a verb as the center and is made up of more than one verb. A verb phrase contains a main verb and one or more helping verbs. The helping verbs help the main verb to show the action. The verb in a sentence expresses the tense or time type, person and number of grammatical categories. Using the same verb phrase often acts as a predicate in a sentence, the subject, the attribute and so on. The Kazakh noun phrases divided by the function of phrases in our system are shown below.

1) n+v; 2) v+v; 3) adv+v; 4) n+vc; 5) n+n+v; 6) n+va; 7) vc+v ; 8) pron+v;

9) pron+va; 10) va+v ;11)n+vd; 12)adv+vd; 13)n+v+v; 14)vb+v; 15)adj+va;

16) num+"سە"+"v; 17) adj+va; 18)v+v+v

For the purpose of testing,1000 sentences are used for analysis in "Xinjiang Daily" (Kazakh version corpus) , which are as follows:1000 sentences ; 8871 words in the text; 416 verb phrases ; The average sentence length is 9 words(see attached 1).

4.2 Kazakh noun phrase structure

A noun is a word that names a person, place, thing, or idea. noun phrases are related to several areas, including the plural form, case, possessive person, predicate person in Kazakh. . These

factors are the difficulties in Kazakh phrase extraction. The Kazakh noun phrases divided by the function of phrase in our system are shown below and in attachment 2.

1) n+n; 2) n+conj+n; 3) pron+conj+pron; 4) pron+n; 5) adj+conj+adj;

6) adj+n; 7) adj+adv+n ; 8) num+n ; 9) v+n; 10) [ ]+n

4.3 Kazakh adjective phrase structure

An adjective is a word that describes a noun or pronoun.The Kazakh adjective phrase divided by the function of phrase like follow and attached 2 .

1)adj+n; 2)adj + v; 3) adj+n+v; 4)pron+adj; 5) adv+adj+n; 6) adj+adj+n; 7) num+adv+n ;

5 Rule-based verb phrase recognition algorithm

Kazakh has two characteristics that have to be taken into account: agglutinative morphology, and rather free word order with explicit case marking.

Input：word segmentation（extraction stem and affix） and POS tagged corpus (test.xml);

Output：First：Phrase tagged file(result.xml)；Second：Phrase file(resultP.txt);

Rule-based phrase recognition algorithm as follow:

（1）i=1;

（2）while (!(test.xml))

①From right to left match rule in rule base;

②if match then put phrase boundary and phrase POS tag.

③i=i+1 (move right)

（3）Output recognition phrase and phrase file.

Based on the basic rules of phrase, we have done extraction of phrases from POS tagged Kazakh corpus. The extraction process is as follows:

(a) First roughly segmented XML corpus. The common segmentation marks include semicolon, comma, full stop, exclamation mark, question mark.

(b) For the segmented data, we extract the three elements of basic phrase: part of speech (POS), affix, and the word.

(c) Look for the matched rule in the rule set. If found, save the basic phrase. Otherwise go back step 1.

6 Based on ME for Kazakh phrase Identification

The Kazakh phrase recognition task is that x represents the environmental context words to be marked and y is the output. Achieve task: the instance or context condition x, construct a model can accurately estimate the category marker appears the result y probability, as: p(y/x)。

Model input:

Labeled training data from the training sample set extracting T = {(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)}, (x_i, y_i) that appear in the corpus when it yi context information for the x_i. Feature function in that f is between x and y refers to a particular relationship exists, a binary function that:

F(a,b)= { 1 If x, y condition

0 otherwise

The entropy model P： $$H(p) \equiv -\sum_{x,y} p(x,y)\log(x,y)$$

Maximum Entropy Model：Such a model can be shown to have the following form:

$$p^* = \arg\max_{p \in C} H(p)$$

Goal: select a distribution p from a set of allowed distributions that maximizes H(Y|X). compute:

$$p^*(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_i \lambda_i f(x,y)\right]$$

$$Z_\lambda(x) = \sum_y \exp\left[\sum_i \lambda_i f(x,y)\right]$$

Where the λ_i are the model parameters and the f are the features of the model.

6.2 Feature extraction

6.2.1 Feature defined

According to own characteristics of a Kazakh basic verb , this feature space is defined as:

(1) *the word*, including the current word, the right and left sides of a word. (2) *part of speech*, including the current word speech, about the two parts of speech information. (3) *Affix ingredients*, including the current word and the word about the additional ingredient information.

(4) Phrase tag that contains the current word and the words to the right and the left two words Phrase marker.

This rule-based approach applied to generate the maximum entropy model training corpus, based on Kazakh Linguistics, the feature space show as table 2.

Table 2. Feature templates

Feature tag	Meaning	Feature tag	meaning
w（-1）	previous one word	POS（+1）	POS of next one word
w（0）	the current word	POS（+2）	POS of next two word
w（+1）	next one word	affix（-1）	affix of previous word
pos（-2）	POS of previous two word	affix（0）	affix of current word
pos（-1）	POS of previous one word	affix（1）	affix of next one word
pos（0）	POS of the current word		

6.2.2 Feature selection

There are two general feature selection methods: incremental feature selection and feature selection of based on frequency threshold. Appear relatively large frequency characteristic was selected, the frequency is greater than a threshold value equal to a characterristick. Through repeating them many times, the frequency threshold value was characterized k = 5, characterized in that the use of the frequency characteristic is greater than 5.

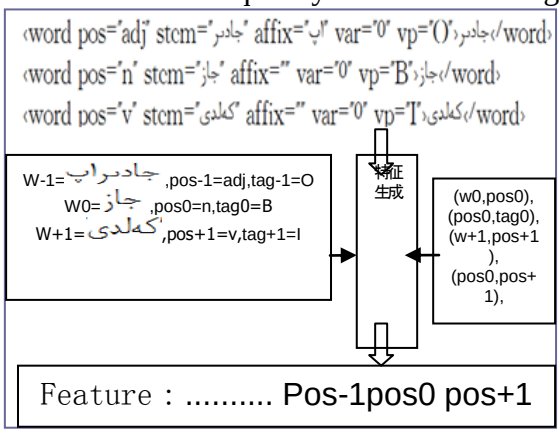


Figure 2. Feature generate process

7 Phrase disambiguation

7.1 Analysis of Kazakh phrase structure ambiguity

Ambiguity computer analysis of language structure has been one of the difficulties problems faced by the earliest. problems and eliminate ambiguity effective structural policy research has Hindle, Rooth of computational linguistics research and Brill of rule-based approach eliminate ambiguity of the phrase matching.

This article from the delimitation ambiguity and structural relationship is to study two aspects of phrase structure ambiguity.

one of the difficulties in Kazakh phrase research is the phrase disambiguation problem. Ambiguous reasons is word POS ambiguity, phrase boundaries is not easy to determine, POS with the same sequence, there are five ambiguous forms.

（1）VD form（v+adv）

Eg.1a：قابىلداۋى/v تومەندەۋ/adv is verb phrase.

Eg.1b：قابىلدۇى/v مۇەكشە/adv is adverb phrase.

（2）ND form（n+adv, pron+adv)

Eg.2a：كيىمىن/n جاڭالاۋ/adv is verb phrase.

Eg.2b：ناتيجەسى/n كەرەمەت/adv is adverb phrase.

（3）NPV form（n+prep+v, pron+prep+v)

Eg.3a：بىنتىماق/n تۇرالى/prep ۇيرەنۋ/v is verb phrase.

Eg.3b：اسان/n عانا/prepەدى/v is noun phrase.

（4）VPV form（v+prep+v)

Eg.4a：كەلدى/v دە/prep كەتتى/v is verb phrase.

Eg.4b：زەرتتەۋ/v تۇارلى/prepتۇسنۋ/v is adverb phrase.

（5）VP form（v+prep)

Eg.5a：سويلەۋدەن/v بۇرىنداۋ/prep is verb phrase.

Eg.5b：جونىندە/prep زەتتەۋ/v is verb phrase.

For these ambiguities, we can not simply use the rules to match ways to eliminate, but rather to use maximum entropy model to solve the problem.

8 Kazakh phrase system

Kazakh phrase recognition system consists of four modules, for example, training module, identification module, test module and auxiliary module. By following a comprehensive analysis of Kazakh words, the following is the Kazakh shallow parsing process:

（1）Sentence：

قوڭىر كۇز كەلىپتى، قامبار سول جەرگە، قوي باعىپ كەلسە، ازىناعان كۇزدىڭ جەلى سوعىپ تۇرىپتى.

（2） POS:

قوڭىر/n كۇز/n كەلىپتى/v، قامبار/n سول/pron جەرگە/n، قوي/n باعىپ/v كەلسە/v، ازىناعان/adj كۇزدىڭ/n جەلى/n سوعىپ/v تۇرىپتى/v.

（3）Phrase POS:

[[قوڭىر/n كۇز/n NP[كەلىپتى/v]، VP[[قامبار/n سول/pron جەرگە/n،]RP NP[[قوي/n باعىپ/v VP[كەلسە/v]]، VP[[[[ازىناعان/adj كۇزدىڭ/n جەلى/n]NP]AP سوعىپ/v]VP تۇرىپتى/v].VP

（4）Tree bank:

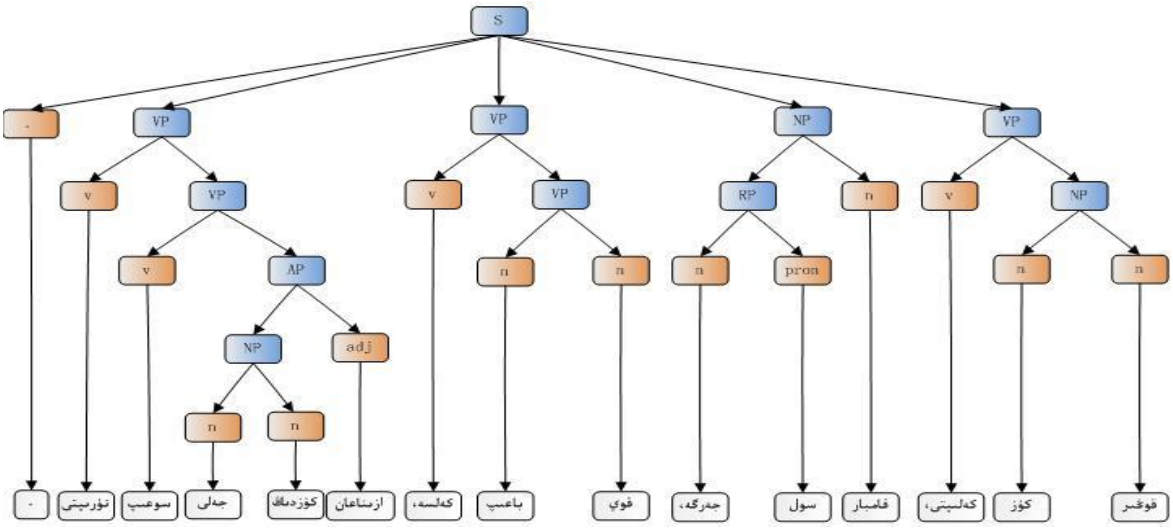


Fig.3 Kazakh Tree Bank

Fig.4 Kazakh verb phrase identify system

9 . Experiment results and analysis

9.1 Dataset

In this paper, as the data set we are using is the data of 31 days of January 2008 of the Xinjiang Daily corpus. The corpus consists of the raw texts and the POS tagged XML format texts. Experiments were done for phrase extraction .

```
<paragraph id="1">
  <word pos="n" stem="ءتتلتىنز" affix="/تىنز" var="0" vp="O">ءتتلتىنز</word>
  <word pos="n" stem="كۇاش" affix="" var="3" vp="O">كۇاش</word>
  <word pos="n" stem="جيان" affix="" var="5" vp="O">جيان</word>
  <word pos="n" stem="ءنرلەمجەن" affix="/دەن" var="0" vp="B">ءنرلەمجەندەن</word>
  <word pos="v" stem="حابارلايدى" affix="" var="0" vp="I">حابارلايدى</word>
  <punction>.</punction>
```

Fig. 5 Verb phrase Annotated corpus

9.2 Experiment results

The experiments of the accuracy rates are evaluated using as follow standard evaluation measures:

recall=a/(a+b)*100% ;

precision= a/(a+c)*100% ;

leakage=b/(a+b)*100% ;

error=c/(a+c)*100% ;

Note: recall +leakage=1 ; precision +error=1 ; a is number of correctly identified phrases. b is number of missed phrases. c is number of wrongly identified phrases.

In the test corpus, there are 3000 correct tagged sentences as training data for the close test, and other 1000 sentences for the open test.

Table 3.phrase indentify test

method	Test type	precision（%）	recall（%）	error（%）	leakage（%）
rule	Close test	81.58	72.51	18.42	27.49
rule	Open test	78.22	70.01	21.78	29.99
ME	Close test	91.62	87.33	8.81	15.67
ME	Open test	87.89	83.13	12.11	16.87

10 Conclusion

This paper identified Kazakh phrases based on rules and the maximum entropy method. It used the Kazakh word, part of speech, affixes context information to design template of features by maximum entropy model. GIS algorithm was investigated to the feature set of parameter estimation, and the final output of the optimal recognition results of the phrase. Based on statistical methods, we can obtain higher accuracy in the close test, but were unable to get a good result in the open test, which requires training more and more corpora.

Acknowledgments

This work is supported by National Natural Science Foundation of China（NSFC） under Grant No. 61063025.

Reference

[1] Church K.A stochastic parts program and noun phrase parser for unrestricted text[J]. In Proceedings of the Second Conference on Applied Natural Language Processing. Texas, USA. 1988,19(8):136-143.

[2] Steven Abney. Parsing by chunks[M]. Dordrecht: Kluwer Academic Publishers,1991:257-278

[3]Rob Koeling. Chunking with Maximum Entropy Models[J]. Proceedings of CoNLL-2000 and LLL-2000,2000,109(15):139-141

[4] Zhao Jun and Huang Changning,. Chinese basic noun phrase structure analysis model, Computer sinence[J].,1999，22(2):141-146.

[5]Qiang Zhou,2004,Annotatiion scheme for Chinese Treebank, Journal of Chinese Information Processing, Vol 18(4),Pages 1-8.

[6] Gulila.Altenbek,Ruina-Sun,Kazakh Noun Phrase Extraction based on N-gram and Rules,2010 International Conference on Asian Language Processing (IALP2010),Harbin,China,2010, Pages 305-308.

[7] Gulila, A. and Dawel,A. and Muheyat,N.(2009).A Study of Word Tagging Corpus for the Modern Kazakh Language, Journal of Xinjiang University[J]., 26(4), Pages 394-401.

[8] Adam Berger, Stephen Della Pietra, and Vincent Della Pietra(1996),A Maximum Entropy Approach to Natural Language ,Processing Computational Linguistics, 22(1), Pages 39-71.

[9]Adwait Ratnaparkhi. Learning to parse natural language with maximum entropy models[J].Machine Learning,1999,341(3):151-176

[10]Porter,M.F.(1980)..An algorithm for suffix stripping, Program, 14(3)：130−137.

[11]Karttunen,Lauri(1983). KIMMO: A general morphological processor. Texas Linguistic Forum, 22:163–186.

[12]Gülşen,E. and Eşref,A.(2004).An affix stripping morphological analyzer for Turkish, Proceedings of the International Conference on Artificial Intelligence and Application, Austria, 299-304.

[13]Kemal Oflazer(1994).Two-level description of Turkish morphology. Literary and Linguistic Computing,9(2):137-148.

[14]Beesley,K.R.(1996).Arabic finite-state morphological analysis and generation. In COLING-96, Copenhagen,pages 89-94.

[15]Milat,A.(2003).Modern Kazakh language, Xinjiang People's press, China.

[16]Dingjing Zhong. Practical Grammar of Modern Kazakh Language. Beijing: Central University for Nationalities Press,2004.

Attachment 1 : Part of Speech match statistics of verb phrase

POS match	Example	Number of VP	Percentage of VP
n+v	جەردى قورعاۋ	65	15.62%
v+v	باسا زەرتتەۋ	49	11.78%
adv+v	جوعارى بولۋ	38	9.13%
N+vc	ساعاتقا سوزىلعان	36	8.65%
n+n+v	سايا ساتتەلەدى ەكەن	35	8.41%
N+va	باجىنى ازايتۋعا	28	6.73%
vc+v	ورناتىلىپ بولدى	27	6.49%
pron+v	مىنانى داربيتەۋ	24	5.77%
pron+va	ودان پايدالانۋ	23	5.53%
va+v	بوساتۋدى قولداۋ	21	5.05%
n+vd	نىنتماقتى كۇشەيتىپ	16	3.85%
Adv+vd	كوككە جەتكىزىپ	12	2.88%
n+v+v	اۋىرۇدىن باسا قويدى	12	2.88
vb+v	شعىسقا جەتكىزۇ	11	2.64%
Adj+v	كوپ كورۋ	9	2.16%
num+"ءسە+ "v	ءبىر ەسە ورلەۋ	6	1.44%
Adj+va	رەتتى تۇزەۋىنە	2	0.48%
v+v+v	الىپ قاشىپ كەتۋ	2	0.48%

Attachment 2 : Part of Speech match rules of noun phrase

rule	Type	Example	Rule	type	example
Rule B	n+مۇن+n	شينجياڭ مۇن گانسۋ	Rule A	n+n	قىزىل بالا
	n+بۇن+n	قاعاز بۇن اعاش		adj +n	جاقسىز بالا
	n+پۇن+n	ورىندىق پۇن الما		num+n	ءۇش كىتاپ
	n+ءارىز+n	مۇعالىم ءارىز وقۋشىز		pron+n	بارلىق وقۋشىز
	n+جانە+n	ادىل جانە اسقە		v+n	قۇلاعان تام
	pron+مۇن+ pron	ولار مۇن ءبىز		adj+adv+n	بۇگىن سۇنبىز
	pron +بۇن+ pron	ءسىز بۇنءبىز	Rule C	adj+ سۇپىلىك +n جالعاۋ	بازار تۇتىنۋشىلىق بازار
	pron +پۇن+ pron	كوپشىلىك پۇن			

		تئثدارمان
	pron +ءارئ+ pron	ءبئز ءارئ سةندةر
	pron +جانة+ pron	ءبئز جانة ولار
	N or pron+ تاؤةلدئك ئلئك سةپتئك orجالعاؤ +nounجالعاؤ+ تاؤةلدئك جالعاؤ	مةنئث كئتابئم

Attachment 3 : Part of Speech match rules of adjective phrase

type	Example	Type	example
adj+v	**تەز جۇز**	adj +n +adj	**كومىس شاشتى اۋليەدەي**
adj+n	**اقبوز ات**	adj +conj +adj	**كەڭ دە جايلى**
Pron+adj	**بارلىعىمىز تىلەكتەس**	pron +adj +pron	**دشكىم اقماق دەمەس**
adv+adj	**وتە ناشار**	adj +n +n	**سۇرى ايۇدىڭ تارىسى**
num+adj+n	**بەس بۇرىشتى ساراي**	n +adj +v	**بويى بيىك دكەن**
adv+adj+n	**دەڭ تاڭدامالى شعارما**	n +adv +adj	**بويى دەڭ بيىك**
adj+adj+n	**قويۇ قارا شاش**	adv +n +adj	**بۇگىن اۋارايى جاقسى**
adj+n+v	**بولىنگەندى ٴبورى جەيدى**	adv +adj +v	**نەداۋىر جاقسى وتتى**
adj +مەن+adj	**اقىلدى مەن اقىلسىز**	adv +adj +pron	**ونشا كەرەمەت دەمەس**
adj +بەن+ adj	**كوپەس پەن سىعان**	pron +adj +v	**ولار جان-جاقتىلى ويلاستىردى**
adj +پەن+ adj	**جاراتىلىستىق پەن قوعامدىق**	pron +n +adj	**بارلىعىنڭ رۋحى كوتەرىڭكى**
adj +ٴارى+ adj	**كورىكتى ٴارى اقىلدى**	pron +adj +adj	**بارلىعى ويلاعانداي سالتاناتتى**
adj +جانە+ adj	**قىزىل مەن جاسىل**	adj +pron +v	**ٴوزىمشىل دەمەس ەدى**
n + adj + n	**ولكە دارەجەلى قامسىزداندىرۋ**	adj +n +adj	**ۇيالشاق ادام تارتىنشاق**
adj + v + v	**تەز جۇرسەڭ بولاما؟**	n +adv +adj	**جايلاۋ قانداي تاماشا**
pron + adj + adv	**سول قىرىقتى كەزدەر**	n +adj +adj	**ٴدامى كەرەمەت جاقسى**
pron + adj +n	**مناۋ جاقسى ات**		

З.А. СИРАЗИТДИНОВ, Б.З.СИРАЗИТДИНОВ

Институт истории, языка и литературы Уфимского научного центра РАН, Уфа, Республика Башкортостан

КОРПУСНЫЕ ПРОЕКТЫ В БАШКИРСКОМ ЯЗЫКОЗНАНИИ

В докладе рассматривается общее состояние корпусной лингвистик в зарубежной и отечественной лингвистике и вопросы разработки корпусов в Институте истории, языка и литературы УНЦ РАН. Автором анализируется деятельность лаборатории лингвистики и информационных технологий в рассматриваемой области. Описываются предлагаемые методы создания корпусов прозаических и публицистических текстов башкирского языка, ставится задача на перспективу.

Ключевые слова: корпусная лингвистика, башкирский язык, информационные системы, прикладная лингстистика.

The article discusses the state of corpus linguistics in the domestic and foreign linguistics and design issues of corpus at the Institute of History, Language and Literature, Ufa Science. The author analyzes the work of the laboratory of linguistics and information technology in this area. We describe the proposed methods of creating of corpus of the Bashkir language, analyzes the results obtained, the task for the future.

Keywords: corpus linguistics, the Bashkir language, information systems, applied lingstistika

Статья подготовлена при поддержке гранта РФФИ 11-06-97001-р_поволжье_а “Разработка корпуса прозаических текстов башкирского языка”.

Зародившееся в 60-х годах прошлого века направление в зарубежном языкознании, связанное с компьютерной обработкой больших объемов текстов, сформировалось в новое быстро растущее направление филологии - корпусная лингвистика – “со своими традициями, признанными авторитетами, научными центрами, методами и проблематикой” [1]. Данному направлению сегодня во всем мире уделяется значительное внимание. Объектом нового филологического направления являются речевые материалы, реализованные в виде как письменных текстов, так и устных (фонетических) массивов данных. Корпусная лингвистика занимается созданием общих унифицированных принципов представления таких сверх-больших массивов языковых данных (корпусов), непосредственным созданием самих корпусов и выполнением конкретных экспериментальных лингвистических исследований на базе этих данных [2;3]. Данное направление лингвистики является приоритетным и в отечественной филологии. Так, если в “Плане фундаментальных исследований Российской академии наук на период 2006-2010 гг.” был раздел 9.2.3., касающийся создания электронного корпуса текстов русского языка, то в “Плане фундаментальных исследований Российской академии наук на период 2011-2025 гг.” в разделе 9.(б) ставится научная задача создания электронных корпусов текстов языков народов Российской Федерации [3]. Научный фонд РФФИ отдельно выделил корпусные исследования в своем классификаторе (06.4.20, Корпусно-ориентированные исследования) [4].

На сегодня в мире насчитываются более тысячи корпусов, количество их растет экспоненциально. Первый корпус был разработан в 60-х годах. Это Брауновский корпус американского варианта современного английского языка, создававшийся в Брауновском университете в 1962—1963 гг. Объем корпуса около 1 млн словоупотреблений. В начале 2000-х был создан корпус русского языка, на сегодня его объем составляет более 500 млн. словоупотреблений.

Вся совокупность имеющихся корпусов весьма различна, поскольку, как было отмечено выше, объектом самой корпусной лингвистики являются многообразие речевых и письменных материалов языка. Так по английскому, немецкому, китайскому, японскому, турецкому, эстонскому, русскому, польскому языкам реализованы речевые корпусы, содержащие как мультимедийные данные, так и транскрипции речи [5-11]. На стадии создания корпусы и по другим языкам [12-13].

Но наибольшее количество корпусов составлены по письменным текстам. От поставленных целей и задач создания эти корпусы можно по разному классифицировать. Если корпус создается по текстам одного языка, то такой корпус является одноязычным. По объему привлеченных текстовых материалов среди них выделяются корпусы немецкого (DeReKo, 5,4 млрд. слов) [14], английского (BNC, 100 млн. слов) [15], американского варианта английского (450 млн. слов) [16], китайского (LIVAC Synchronous Corpus, 1 млр. слов) [17], венгерского (100 млн. слов) [18], испанского (100 млн. слов) [19], итальянского (100 млн. слов) [20], чешского (200 млн. слов) [21], русского (НКРЯ, 500 тыс. слов) [22] языков. Если же создаются корпусы текстов переведенных на разные языки, то возникают многоязычные или по другому параллельные корпусы. Примерами таких корпусов являются польско-украинский, польско-русский, черногорско-английский, нидерландско-французский, японско-английский и другие параллельные корпусы [23-27]. Такие корпусы используются для сравнительно-соспоставительных исследований. Но в последнее время параллельные корпусы нашли практическое применение в разработках систем статистического перевода, зачинателем которого является компания Google. Одним из ярких примеров такого использования является параллельный корпус слушаний Европарламента, включающий тексты на 21 европейском языке [28].

В зависимости от стилистической принадлежности тестов выделяются художественные, научные [29-30], публицистические [31-33], драматургические, поэтические корпусы [34].

Текстовые корпусы также различаются по принципу отбора материала: выделяются полнотекстовые, когда в корпус попадают полные варианты печатных текстов, и фрагментнотекстовые. В последнем случае в корпус отбираются выборки из текстов. Объемы выборок и место расположения их в текстах каждый составитель определяет произвольно. Так Брауновский корпус построен на базе выборок из 500 текстов, каждый из которых включает 2 000 словоупотреблений. Бирмингемский корпус английского языка и Основной корпус Национального корпуса русского языка являются представителями полнотекстового корпуса [35: 66; 22].

Для решения различных лингвистических задач мало лишь наличия массива текстов. Требуется также, чтобы сами тексты содержали в себе дополнительную лингвистическую информацию в виде специальных разметок, позволяющую использовать их для разных исследовательских и иных целей. В этой связи известный отечественный специалист в области составления корпусов, руководитель проекта Национального корпуса русского языка член-корр. РАН В.А.Плунгян даже подчеркивает, что “собственно, наука о корпусах … — это прежде всего наука о том, как сделать хорошую разметку корпуса” [36: 6].

Составители корпусов по разному подходят к определению состава разметок, но большинство сходится в том, что разметки должны быть двух типов: экстралингвистические (метатекстовые) и лингвистические [37: 175-176]. К экстралингвистическим относится информация, которая паспортизирует сами тексты в целом и дает сведения об авторе (фио, год рождения автора, пол, образование и т.д.), информацию о тексте: (название, год создания, год издания, жанр, тип текста, носитель текста: книга, журнал, электронное издание) и другие. Лингвистические разметки включают морфологические, синтаксические и семантические характеристики, относятся ко всем словоупотреблениям текста, поэтому некоторые авторы называют их лексическими разметками.

Для работы с размеченными текстами необходимо соответствующее программное сопровождение, позволяющее производить разнообразный поиск по корпусу, получать

статистические данные. Размеченные тексты вместе с программным сопровождением образуют корпус в его полном понимании.

В создании корпуса трудоемким и сложным являются следующие этапы:

1) Подготовка электронных текстов. На данном этапе существующие печатные варианты книг сканируются, редактируются и вводятся на электронные носители. Современные зарубежные корпусы создаются при поддержке крупных издательств, которые на безвозмездной основе передают предпечатные варианты текстов разработчикам корпусов.

2) Проведение разметки текстов. Степень трудоемкости данного этапа определяется уровнем развития таких разделов конкретного языка как компьютерная и математическая лингвистика. Если в языке проведены соответствующие исследования и составлена компьютерная модель, то возможны разработки средств автоматизации процесса. Первостепенной задачей в этом процессе является разработка автоматического морфологического анализатора языка. Далее следуют программы автоматического снятия омонимии, синтаксического и семантического анализа. Но даже в этом случае остается значительная доля ручной работы, поскольку не все языковые явления однозначно могут быть идентифицированы программными средствами.

Сейчас все крупные языки обзавелись своими национальными корпусами. К созданию корпусов приступили все остальные языки мира. Ведутся корпусные разработки и по языкам народов России: бурятского [38-39], калмыцкого [40-41], лезгинского [42] осетинского [43] и др. Отдельно отметим научные разработки и корпусные проекты по языкам тюркской группы, родственным башкирскому языку: казахский [44], татарский [45-46], тувинский [47-48], турецкий[491], шорский [50], хакасский [51].

Лингвистику 21 века называют корпусной лингвистикой. При этом данное направление лингвистики активно влияет на все остальные направления языкознания, изменяет теоретические приоритеты и создает новые идеологии в понимании того, что же представляет собой язык [52; 7-8].

Иследователями также отмечается, что корпусы открывают перспективу для новых исследований не только в области лингвистики, но и в смежных областях: в литературоведении (для стилеметрических исследований, определения нормативности употребления языковых реалий), в общественных науках (изучение социальных объектов через язык, используя такие параметры текстов, как период, автор или жанр, семантический контент текстов), в информационно-технических разработках (создание автоматизированных систем машинного перевода, распознавание речи, информационный поиск).

Сегодня в Институте истории, языка и литературы УНЦ РАН активно осваиваются новые направления лингвистики прикладного характера, основывающиеся на накоплении лингвистических баз данных и компьютерной обработке. Есть первые результаты по экспериментальной фонетике, выполненные Ишкильдиной Л.К. [53]. Каримовой Р.Н. накапливается диалектная текстологическая и речевая база [54, 55], разработан машинный фонд башкирского языка [56]. Сиразитдиновым З.А. и Мигрrunовой Л.Г. составляется база терминологических данных [57], полным ходом идет работа и по корпусной лингвистике.

Работа по корпусу башкирского языка осуществляется сотрудниками лаборатории лингвистики и информационных технологий ИИЯЛ УНЦ РАН (Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г., Полянин А.И.) в двух направлениях: а) корпус прозаических текстов; б) корпус публицистических текстов.

Первое направление разрабатывается по гранту РФФИ “Разработка корпуса прозаических текстов башкирского языка”, № 11-06-97001-р_поволжье_а. Начало работы 2011 г., конец — 2013 г.

Второе направление осуществляется в рамках программы Президиума РАН “Корпусная лингвистика. Создание и развитие корпусных ресурсов по языкам народов России”. Сроки реализации 2012—2014 гг. [58].

На сегодня по корпусу прозаических текстов разработаны системы экстралингвистических и лингвистических помет для разметки, создана программа

автоматического морфологического анализа, подготовлены и автоматически размечены тексты 773 произведений более 70 авторов общим объемом порядка 10 миллионов (10829086) словоформ, запущен проект поисковой системы в сети [59]. Сейчас идет отладка и оптимизация работы корпуса, ведется работа по оцифровке новых текстов. К концу года намечается доведение объема корпуса до 20 миллионов словоформ и запуск самого корпуса в сети Интернет на сервере Института со своим доменным именем. Проект корпуса прозаических текстов полностью разработан на базе СУБД Оракл на платформе Unicode [http://mfbl.ru/bashkorp/korpusp]. Для работы с корпусом пользователь может установить башкирскую раскладку клавиатуры средствами системы (ОС Vista, Seven), установить программу Хамелеон 8.0 (для ОС 98, ME, 2000, XP) или воспользоваться виртуальной клавиатурой самого корпуса.

По второму направлению подготовлены тексты республиканских газет и журналов общим объемом в 5 миллионов словоформ. Идет работа по автоматической морфологической разметке. Корпус будет выставлен к концу года.

Система экстралингвистических разметок публицистического корпуса включает название прессы, год, месяц и день выхода, название статьи, автора. Все тексты размечены по тематике и жанру. Для рассматриваемого корпуса выделены следующие тематики и жанры:

Тематика: политическая и социальная жизнь (политика, право, философия); экономика (производство, строительство, бизнес, финансы, коммерция); сельское хозяйство; искусство, культура и литература; наука и техника; образование; природа, путешествие; частная жизнь; спорт; религия; психология; медицина; красота и здоровье.

Жанры текстов: интервью, беседа, статья, очерк, репортаж, обозрение, советы, письма, обзор печати (новости из других источников), поздравления, художественно-публицистические жанры (эссе, фельетон, рассказ, стихи, эпиграммы), рецензия.

По корпусу же прозаических текстов нами выделяются только авторы, названия произведений, год издания/завершения работы над произведением.

Разрабатывемые корпусы текстов башкирского языка по классификации Захарова В.П. [2 12-13] относятся к следующим типам:

по типу языковых данных	письменный
по параллельности	одноязычный
по критерию литературности	литературный
по жанру	литературный, публицистический
по доступности	свободный доступ
по разметке	размеченный
по характеру разметки	морфологический, семантический
объем текстов	полнотекстовый

Система морфологической разметки обоих корпусов ориентирована на представление всех регулярных словоизменительных грамматических форм, не всегда отражаемых и совпадающих с формами, принятыми в академической грамматике. Морфологическая информация башкирской словоформы в корпусе включает: а) частеречную характеристику; б) совокупность морфологических признаков по типу агглютинативных аффиксов словоизменения, которые подразделяются на именные и глагольные формы[*].

Выделяются 12 частей речи: имена существительные, числительные, прилагательные, наречия, глаголы, местоимения, подражательные слова, междометия, модальные слова, союзы, частицы, послелоги. Эти характеристики указываются в словаре основ.

Именные морфологические признаки включают показатели следующих 15 категорий: числа, падежа, принадлежности, сказуемости, вопросительности, неопределенности, усиления, притяжательности, уменьшительно-ласкательности, уподобления, атрибутивный

[*] Авторы выражают благодарность член-корреспонденту РАН А.В.Дыбо за ценные советы в разработке системы морфологических разметок башкирского языка.

локатив (дағы/тағы), обладательности, лишительности, предельности, сравнительной степени.

Глагольные морфологические признаки включают показатели следующих 11 категорий: вопросительности, неопределенности, усиления, отрицания, наклонения, деепричастия, причастия, имени действия, инфинитива, хабитуалиса (**сан/-сән**: барыусан, үсегеүсән), образования абстрактных субстантивов (**-лыҡ/-лек**: етерлек, алырлыҡ).

В корпусе размечаются следующие подкатегории для глагольных форм: 1) времена (настоящее время, будущее время: будущее неопределенное время, будущее определенное время, прошедшее время: прошедшее неопределенное время, прошедшее определенное время, предпрошедшее определенное время –**ғайным/-гәйнем**); 2) подкатегория лица (1-3); 3) подкатегория числа (ед., мн.).

Для именных форм выделяются следующие подкатегории: 1) подкатегория лица (1-3); 2) подкатегория числа (ед., мн.).

Морфологический анализатор корпуса реализован на основе алгоритма последовательного вычленения из словоформы букв и сравнения остатка словоформы и вычлененного фрагмента со словарями основ и аффиксов башкирского языка.

Для правильной идентификации основы и аффиксов используются грамматические фильтры: 1. Фильтр соответствия фонетической структуры аффикса фонетической структуре основы 2. Фильтр соответствия сочетаний аффиксов нормативным правилам. Данный фильтр основывается на списках возможных моделей сочетания словоизменительных аффиксов башкирского языка, которые были нами ранее рассмотрены в одной из наших работ [60]. 3. Фильтр графической передачи на стыках фонем.

Словарь основ включает нарицательные и собственные слова. Нарицательная часть словаря основ состоит из 60 тыс. единиц, включает лексику литературного башкирского языка. Часть имен собственных словаря включает имена, фамилии, отчества, клички животных и людей, географические названия башкирского и русского языков, имеет объем порядка 20 тыс. единиц.

В словарях основ указаны части речи, типы нарушений сингармонизма и возможные остатки основ при словоизменительных процессах и прочие варианты.

Прект национального корпуса башкирского языка художественной прозы позволяет производить следующие операции:

- поиск словоформы,
- поиск леммы,
- поиск грамматических категорий словоизменений,
- поиск грамматических подкатегорий,
- поиск сочетаний грамматических категорий,
- поиск сочетаний грамматических подкатегорий,
- поиск сочетаний словоформ,
- поиск сочетаний лемм,
- выдача списка небашкирской лексики (вкраплений по языкам источникам),
- построение частотного словаря словоформ,
- построение частотного словаря лемм.

Сегодня проект корпуса прозаических текстов активно используется сотрудниками отдела языкознания при составлении многотомного академического толкового словаря башкирского языка.

Перед коллективом лаборатории линвгистики и информационных технологий ИИЯЛ УНЦ РАН в 2013 г. стоят следующие задачи:

1) доведение объема корпуса до 20 миллиона словоупотреблений;

2) разработка системы выдачи статистических распределений по любому заданному пользователем подкорпусу;

3) разработка системы выдачи графических представлений статистических распределений.

Литература

1. Рыков В.В. Прагматически ориентированный корпус текстов//Тверской лингвистический меридиан, Тверь, 1999 (http://rykov-cl.narod.ru/t.html, дата обращения: 17.06.2013).

2. Захаров В.П. Корпусная лингвистика: Учебно-методическое пособие. – СПб., 2005. – 48 с.

3. План фундаментальных исследований Российской академии наук на период 2011-2025 гг. URL:http://www.ras.ru/scientificactivity/plan2025.aspx (дата обращения: 17.06.2013).

4. Классификатор РФФИ. URL:http://scs.viniti.ru/rubtree/main.aspx?tree=RFFI&cod=06 (дата обращения: 17.06.2013).

5. LDC Top Ten Corpora (мультимедийные корпусы английского языка). URL: http://www.ldc.upenn.edu/Catalog/topten.jsp (дата обращения: 17.06.2013).

6. Chinese Broadcast Conversation Speech (мультимедийный корпус китайского языка). URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId= LDC2013S04 (дата обращения: 17.06.2013).

7. The *Corpus* of Spontaneous Japanese (мультимедийный корпус японского языка). URL: http://www.ninjal.ac. jp/products-k/katsudo/seika/corpus/public/index.html (дата обращения 17.06.2013).

8. The Spoken Turkish Corpus (vультемидийный корпус разговорного турецкого языка). URL:http://stc.org.tr (дата обращения: 17.06.2013).

9. Фонетический корпус спонтанной эстонской речи. URL:http://www.murre.ut.ee/phonetic-corpus (дата обращения: 17.06.2013).

10. Фонетический немецкого разговорного языка URL:http://dsav-wiss.ids-mannheim.de/korpora/pf/pf _doku.htm (дата обращения: 17.06.2013).

11. Фонетические корпусы русского и польского языков URL:http://www.voicemethods.com/new/databases/corpuses.php3 (дата обращения: 17.06.2013).

12. Людовик Т.В., Робейко В.В., Пилипенко В.В. Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи)// Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 - 29 мая 2011 г.). Вып. 10 (17).- М.: Изд-во РГГУ, 2011. С.478-489.

13. Крючкова О. Ю., Гольдин В. Е. Корпус русской диалектной речи: концепция и параметры оценки/ Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). — М.: Изд-во РГГУ, 2010. С..359-368.

14. Das Deutsche Referenzkorpus (DeReKo) URL:http://www.ids-mannheim.de/kl/projekte/korpora (дата обращения: 17.06.2013).

15. British National Corpus (BNC). URL:http://www.natcorp.ox.ac.uk (дата обращения: 17.06.2013).

16. The corpus of contemporary american english (COCA). URL:http://corpus.byu.edu/coca (дата обращения: 17.06.2013).

17. Корпус китайского языка. (LIVAC Synchronous Corpus). URL:http://www.rcl.cityu.edu.hk/livac (дата обращения: 17.06.2013).

18. Magyar Nemzeti Szövegtár (корпус венгерского языка). URL:http://corpus.nytud.hu/mnsz (дата обращения: 17.06.2013).

19. Corpus del español (корпус испанского языка). URL:http://www.corpusdelespanol.org (дата обращения: 17.06.2013).

20. Corpus di riferimento della lingua italiana scritta contemporanea (“CoLFIS”) (корпус итальянского языка) URL:http://www.ge.ilc.cnr.it/dizionari.php (дата обращения: 17.06.2013).

21. Český národní korpus (ČNK) (чешский национальный корпус). URL:http://ucnk.ff.cuni.cz (дата обращения: 17.06.2013).

22. Национальный корпус русского языка. URL:http://www.ruscorpora.ru (дата обращения: 17.06.2013).

23. Польско-украинский параллельный корпус. URL:http://www.domeczek.pl/~polukr/index.php?option=welcome (дата обращения: 17.06.2013).

24. Польско-русский параллельный корпус. URL:http://pol-ros.polon.uw.edu.pl (дата обращения: 17.06.2013).

25. Englesko-crnogorski paralelni korpus (черногорско-английский параллельный корпус). URL:http://www.eiprevod.gov.me/korpus (дата обращения: 17.06.2013).

26. *Dutch Parallel Corpus* (**DPC)** (нидерландско-французский параллельный корпус). URL:http://dpc.inl.nl/indexd.php (дата обращения: 17.06.2013).

27. Japanese-English Parallel Corpus (японско-английский параллельный корпус). URL:http://www.manythings.org/corpus (дата обращения: 17.06.2013).

28. European Parliament Proceedings Parallel Corpus 1996-2011 (параллельный корпус слушаний Европарламента). URL:http://www.statmt.org/europarl дата обращения: 17.06.2013).

29. Corpus Albaruthenicum (корпус научных белорусских текстов). URL:http://grid.bntu.by/corpus/ (дата обращения: 17.06.2013).

30. *Zientzia eta Teknologiaren Corpusa (н*аучно-технический баскский корпус). URL:http://www.ztcorpusa.net/cgi-bin/kontsulta.py (дата обращения: 17.06.2013).

31. Корпус русских публицистических текстов второй половины XIX века. URL:http://smalt.karelia.ru/corpus/index.phtml (дата обращения: 17.06.2013).

32. Компьютерный корпус текстов русских газет конца XX века. URL:http://www.philol.msu.ru/~lex/corpus/ (дата обращения: 17.06.2013).

33. Romanian corpus (корпус румынской прессы). URL:http://corp.hum.sdu.dk/cqp.ro.html (дата обращения: 17.06.2013).

34. Поэтический подкорпус НКРЯ. URL:http://www.ruscorpora.ru/search-poetic.html (дата обращения: 17.06.2013).

35. Баранов Анатолий Николаевич Введение в прикладную лингвистику: Учебное пособие. — М.: Эдиториал УРСС, 2001. — 360 с.

36. Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, С. 6—20.

37. Поляков А. Е. Технология подготовки информации в Национальном корпусе русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 175—192.

38. Бадмаева Л.Д., Бадагаров Ж.Б., Цыдыпов Б.З. Общие проблемы формирования корпуса бурятского языка с. 24-30/Труды международной конференции «Корпусная лингвистика – 2008» 6–10 октября 2008 г., Санкт-Петербург. Санкт-Петербург, 2008.

39. Корпус бурятского языка. URL:http://web-corpora.net/BuryatCorpus/search/?interface_language=ru (дата обращения: 17.06.2013).

40. Куканова В. В. Архитектура метаописания в Национальном корпусе калмыцкого языка // Вестник Калмыцкого института гуманитарных исследований РАН. 2011. № 1. С. 139–145.

41. Корпус калмыцкого языка. URL:http://web-corpora.net/KalmykCorpus/search/?interface_language=ru (дата обращения: 17.06.2013).

42. Корпус лезгинского языка. URL:http://www.dag-languages.org/LezgianCorpus/search/ (дата обращения: 17.06.2013).

43. Корпус осетинского языка. URL:http://www.ossetic-studies.org/iron-corpus/search/index.php?interface_language=ru. (дата обращения: 17.06.2013).

44. Жұбанов А.Қ. Қазақ тілінің аннотацияланған мәтіндер корпусындағы етесті сөздерге лексик-морфологиялық белгі-код (белгіленім) қоюдың алғышарттары/”Тілтаным”, 2012. № 1, 18-25 б. (Журнал Института языкознания им. А.Байтурсынова, Казахстан, Алматы).

45. Сулейманов Д.Ш., Хакимов Б.Э., Гильмуллин Р.А.Корпус татарского языка: концептуальные и лингвистические аспекты// Вестник Татарского государственного гуманитарно-педагогического университета. № 4(26), 2011. С. 211-216.

46. Письменный корпус татарского языка. URL:http://corpus.tatfolk.ru (дата обращения: 17.06.2013).

47. Салчак А. Я. Электронный корпус текстов тувинского языка // Новые исследования Тувы. 2012, № 3. (Электронный журнал). URL:http://www.new-tuva.info/journal/issue_15/5231-salchak.html (дата обращения: 17.06.2013).

48. Проект тувинского корпуса. URL:http://www.tuvancorpus.ru (дата обращения: 17.06.2013).

49. *Sözlü Türkçe Derlemi (*корпус разговорного турецкого языка). URL: http://std.metu.edu.tr (дата обращения: 17.06.2013).

50. Электронный корпус шорских текстов. URL:http://shoriya.ngpi.rdtc.ru (дата обращения: 17.06.2013).

51. Шеймович, А. В. (2011) Морфологическая разметка корпуса хакасского языка //Российская тюркология. № 2(5). С. 48–61.

52. Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики // Русский язык в научном освещении, 2008, № 16 (2), С. 7—20.

53. Ишкильдина Л.К., Уртегешев Н.С. Фонема [w] башкирского языка: функционирование, история развития, артикуляторные характеристики (по данным томографирования) // Тумашевские чтения: актуальные проблемы тюркологии. Материалы IV Всероссийской научно-практической конференции. – Тюмень: изд.-во «Печатник», 2010. С. 442-446.

54. Каримова Р.Н. Текстологический электронный корпус башкирских говоров //Урал—Алтай: через века в будущее: Материалы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир. Уфа, 2010. С. 189-191. (на башк. яз.).

55. Каримова Р.Н. Электронный фонд экспедиционных аудиозаписей // Урал—Алтай: через века в будущее: Материалы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир. Уфа, 2010. С. 162-163.

56. Сиразитдинов З.А., Максутов А.Д., Полянин А.И., Бускунбаева Л.А. Информационная лингвистическая система “Машинный фонд башкирского языка”// Урал-Алтай: через века в будущее: Материалы IV Всероссийской научной конференции, посвященной III Всемирному курултаю башкир (25-27 марта 2010 г.). Уфа, 2010. I том. С.286-290.

57. Сиразитдинов З.А., Мигранова Л.Г., Ишмухаметова А.Ш., Ибрагимова А.Д., Бускунбаева Л.А. К созданию терминологического банка данных башкирского языка//Урал-Алтай: через века в будущее: Материалы V Всероссийской конференции, посвященной 80-летию Учреждения РАН ИИЯЛ УНЦ РАН (21-22 июня, 2012г), ИИЯЛ УНЦ РАН, 2012, Уфа, 2012, С.111-114.

58. Сиразитдинов З.А., Бускунбаева Л.А., Ишмухаметова А.Ш., Ибрагимова А.Д., Мигранова Л.Г. Корпус текстов периодической печати на башкирском языке/ Актуальные проблемы диалектологии языков народов России: материалы XII региональной конференции. — Уфа, 2012.//С. 139-141.

59. Сиразитдинов З.А., Ибрагимова А.Д., Ишмухаметова А.Ш., Полянин А.И. О пилотном проекте национального корпуса прозаических текстов башкирского языка// Урал-Алтай: через века в будущее: Материалы V Всероссийской конференции, посвященной 80-летию Учреждения РАН ИИЯЛ УНЦ РАН (21-22 июня, 2012г), ИИЯЛ УНЦ РАН, 2012, Уфа, 2012, С.108-111.

60. Сиразитдинов З.А. Моделирование грамматики башкирского языка. Словоизменительная система. Уфа: Гилем, 2006. 160 с.

А.Қ.ЖҰБАНОВ, А.Ә.ЖАҢАБЕКОВА, С.ҚҰЛМАНОВ

А. Байтұрсынұлы атындағы Тіл білімі институты, Алматы, Қазақстан

ФИЛОЛОГТАР ҚАУЫМДАСТЫҒЫ МЕН КОРПУСТЫҚ ЛИНГВИСТИКА ОРТАЛЫҒЫН ҚҰРУ – ҚАЗАҚ ТІЛІНІҢ ҰЛТТЫҚ КОРПУСЫН ЖАСАУДЫҢ АЛҒЫШАРТЫ

Тәуелсіз қазақ елінің экономикалық әлеуеті қарыштап өсіп, саяси аренада тұрақтылық пен толеранттылықтың үлгісі ретінде әлемге танымал болып отырғаны ақиқат. Айналасы бар болғаны жиырма шақты жылдың ішінде адам сенбес жетістіктерге жеткен, әлем елдерінің арасында дағдарысты экономикасына сызат түсірмей еңсере білген мемлекетіміз рухани мәдениетіміздің де өркендеуіне ерекше көңіл аударып келеді. Әсіресе ұлтымыздың ұлы тірегі саналатын қазақ тіліне мемлекеттік тіл мәртебесі беріліп, оны қоғам мен ғылымның барлық салаларында дамыту мен қолданудың өрісін кеңейту мақсатында жасалып жатқан жұмыстар ұшан-теңіз. Мемлекеттік тілді оқытудан бастап, іскағаздарын жүргізу, қазақ тілін ғылым дәрежесіне көтеру бағытында бірнеше бағдарлама қабылданып, тіл біліміндегі зерттеулерді жандандыру, тілші ғалымдардың қызметін белсендіре түсу үшін іс-шаралар ұйымдастырылуда. Осындай іс-шараның бірі ретінде 2013 жылғы 17 мамыр күні тәуелсіз Қазақстан тарихында алғаш рет Филологтардың I съезінің өтуі егемен еліміздің мәдени, рухани өміріндегі ерекше оқиға болғандығын айтуға болады.

Осы съезде тіл біліміне, білім беру ісіне қатысты сан алуан мәселелер көтерілгені, оларды шешу жолындағы іс-әрекеттердің қолға алынып жатқандығынан көзі қарақты оқырман хабардар деп ойлаймыз. Сондықтан біз бұл мақаламызда аталған Съезде баяндама жасаған ҚР Білім және ғылым министрі Б.Т. Жұмағұловтың сөйлеген сөзін негіз етіп алып [1], қазақ тіл білімінің корпустық лингвистика саласының проблемаларына ғана, дәлірек айтқанда, «Қазақ тілінің Ұлттық корпусын» құру мәселелеріне тоқталмақпыз.

ҚР БҒМ Б. Жұмағұлов өз сөзінде қазақ филологтарының «ерекше міндеттерінің бірі – Ұлттық қазақ тілі корпусын қалыптастыру», – деп қадап айтты. Осыған орай министрдің Ұлттық корпус құру туралы ойын оқырман есіне салуды жөн көрдік: «... Тілдің Ұлттық корпусы, бұл – нақты тілде ақпараттың барлық типтері мен түрлерін ауқымды түрде жинақтау. Оны өңдеу, жіктеу және талдау жөніндегі IT-технологиялар. Яғни, тіл білімінің жаңа деңгейін жетілдіру.

... Бұл – біздің еліміз үшін өте өзекті болып табылады.

Мұндай жұмысты күшейту үшін Тіл білімі институты базасында Корпустық лингвистика орталығын құруды орынды деп есептейміз.

Осы орталық арқылы Қазақстан филологтарының қызметін үйлестіруге және жүйелендіруге болады.

Филологтар рөлін арттыру мақсатында **Филологтар қауымдастығын құруды ұсынамын.**

Жақсы тәжірибе – өміршең. Кезінде математиктер, биологтар, тарихшылар қауымдастығы құрылған болатын. Олар уақыт пен заман талабына сәйкес ұсыныстар беріп, осы салалардың дамуына ықпал етуде. Филологтар қауымдастығы да осындай талап пен талғам биігінде болады деген ойдамыз.

Съезд жұмысы Елбасы Н.Ә. Назарбаевтың стратегиялық бағытын іске асыруға және филологиялық ғылым мен білімді одан әрі дамытуға өз үлесін қосады деп сенемін» [1].

Министрдің Филологтар қауымдастығын және Корпустық лингвистика орталығын құру жөніндегі бастамасы осы сала мамандарының, оның ішінде қазақ тілін компьютерлендіру бағытында әртүрлі бағдарламалар жасап, сөздіктер құрастырып жұмыс жасап келе жатқан шағын топтың әрі қарай үйлесімді әрі нәтижелі жұмыс істеуіне серпін береді деп ойлаймыз.

Осындай үміт ұялатқан идеяның негізі жалпы қоғамды автоманттандыру, оның ішінде қазақ тілін компьютерлендіру мәселесінде жатқандықтан, қазақ компьютерлік лингвистикасының пайда болу тарихына тоқталмақпыз.

ХХ ғасырда басталған ғылыми-техникалық «революция» әлемнің кез келген мемлекетінің ішкі-сыртқы саясатына, әсіресе экономикалық әлеуетіне ерекше серпін беріп қана қоймай, Тәуелсіз Қазақстан Республикасы сияқты дамушы елдердің жас мемлекет ретінде қалыптасуында айрықша рөл атқарды. Қоғамдық қызметтің қай саласында да қолданбалы бағыт басымдық алды. Осы орайда лингвистиканың қолданбалы саласы да қалыптасып, дәстүрлі тіл білімінің бағыттарын өз әдіс-тәсілдерімен зерттеуге кірісті.

Қазіргі жаһандану кезеңінде әртүрлі саяси-әлеуметтік, экономикалық қарым-қатынастарға байланысты ақпарат ағыны бұрын-соңды болмаған қарқынмен өршуде. Ал қоғам өміріндегі мұндай ақпарат ағымының таралуы табиғи тілде жүзеге асатындықтан, тіл білімінің қызметі күннен-күнге кеңеюде. Осыған байланысты ұшы-қиырсыз ақпарат ағынын игеру мақсатында шетел және орыс тіл білімінде орасан зор нәтиже беріп отырған тілдік корпустарды қазақ тіл білімінің материалдары негізінде жасау бүгінде үлкен сұранысқа ие болып отыр.

Сондықтан тіл білімінің осындай аса қызығушылық туғызып отырған жаңа саласы – корпустық лингвистиканың зерттеу нысанына нелер жатады, тілдік корпус дегеніміз не, мәтіндер корпусын құрастыру не үшін қажет және ол қандай ғылыми-теориялық мәселелерді шешуге септігін тигізеді деген мәселелерге арнайы тоқталмақпыз.

Соңғы жылдары «Корпустық лингвистика» ғылымның бір саласы ретінде айқын басымдық алып отыр. Өйткені осы саланың зерттеу нәтижесі – мәтіндік корпустарды пайдаланбай тілдік зерттеулерде тәжірибе жүргізудің, әсіресе сөздік құрастырудың, неше түрлі грамматикалар дайындаудың мүмкін еместігі айқындалып отыр. Қазіргі кезде корпустық лингвистиканың мәселелері кейбір оқу құралдарының да арнайы тақырыбына айналуда [2; 3].

Корпустық лингвистика 1963 жылы АҚШ-та пайда болып, Браун корпусынан (The Brown Standard Corpus of American English) бастама алады. Бастапқыда бұл корпустың көлемі 1 млн. сөзқолданыстан тұрып, оның құрамында әрбіреуі 2 мың сөзқолданысқа тең 500 мәтін қамтылған. Браундық корпус осыған ұқсас корпустар құруға қатысты зерттеулердің кеңінен тараған нысаны мен стандартына айналды. Ғалымдар көптеген лингвистикалық зерттеулерді сапалы жүргізу тек ауқымды тілдік материалдар негізінде ғана жүзеге асатындығын ұғына бастады. Осы айтылғандардың барлығы мәтіндерді корпус түрінде ұйымдастыру ережелері мен оларға талдау жүргізу әдіснамасын зерттейтін бағыттың пайда болуына себепші болып, ғалымдарды одан әрі ынталандыра түсті. Сонымен корпустық лингвистика осы әдіс-тәсілдердің әдіснамасы ретінде туындап отыр деуге әбден болады.

Корпустық лингвистиканың ағылшын тіл білімінде кең етек алуын ғалымдар АҚШ-та компьютерлік техника мен ХХ ғасырдың 60-80 жылдары британ лингвистикасындағы интеллектілік ахуалдың белсенді дамуымен түсіндіреді. Осы кездерде тілдік зерттеулердің ең көп бөлігі компьютерленген мәтіндік корпустарға лингвистикалық талдау жүргізуге бағытталғандығы мәлім. Мұндай зерттеулердің нәтижесі 2001 жылы Бирмингем университетінде корпустық лингвистика кафедрасын ашуға және International Journal of Corpus Linguistics журналын баспадан шығарып тұруға мүмкіндік туғызды. Бірнеше тілдің материалдары бойынша құрастырылған, түрі мен қызметі жағынан ерекшеленетін корпустар және солардың негізінде неше түрлі сөздіктер түзіліп, грамматикалар жазылды.

Корпус дегеніміз – әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін мәтіндер жинағы. В.П. Захаров «Корпусная лингвистика» атты оқу құралында былай дейді: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размечанный, филологический компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [3; 4]. Демек, корпус дегеніміз әр тілдегі электронды пішінге келтірілген, яғни бір басқару орталығынан автоматты түрде жұмыс істейтін мәтіндер жинағы. Ал оның қызметі

осы мәтіндер жинағын адамның қарым-қатынас барысында пайдалануына мүмкіндік беруге бағытталады.

Корпуста сақталатын бірлік – ол пәндік саланың қандай да бір жиынтығы. Мысалы, оларға жататындар: сөз, сөзтіркес, сөйлем немесе толық мәтіндер жиынтығы. Мәселен, корпусқа енгізілген мәтіндердегі әрбір тілдік бірлікке, ол не жеке сөз не қосымша болсын, лингвистикалық ақпарат беріледі. Тілде мұндай лингвистикалық ақпарат беруді аннотациялау деп атайды. Аннотация дегеніміздің өзі «қысқаша мазмұн» дегенді білдіреді. Соған сәйкес корпусқа енгізілген тілдік бірліктердің аннотациясы да шартты белгілер арқылы қысқаша беріледі. Орыс тілінде «разметка» деп аталатын мұндай шартты белгілер қазақ тілінде «белгі-код», «белгіленім» деген терминдермен аталып жүр. Мәтіндер бірліктеріне берілген лингвистикалық ақпараттардың толық сипатта болуы осы белгіленімдердің әртүрлілігіне байланысты. Белгіленімдер тілдегі жекелеген деңгейлерге тән тілдік мәліметтерді қамтиды. Ондай белгілер семантикалық (лексика-семантикалық), морфологиялық, синтаксистік сипатта болуы мүмкін. Мәселен, орыс тілінің Ұлттық корпусында орын алған морфологиялық белгілер барлық сөздерді сөз табына, септелу категориясына, тегіне (род – муж., жен.), жіктелуіне, жанды-жансыздығына, етіс және етістік түріне, салыстырмалы шырай және т.б. морфологиялық сипаттамаларына қарай арнайы белгі қою жүргізілген. Сол сияқты семантикалық талдау арқылы да мәтін ішіне қажетті белгілердің қойылатынын айтуға болар еді. Мысалы, орыс тіліндегі үстеу сөздерге «Таксономияға», «бағалауға», «сөзжасамдыққа» қатысты белгілеулер орын алған. Мұндағы «Таксономия» белгісі: орынды, бағытты, қашықтықты, уақытты, жылдамдықты, санына және т.б. жүйелеулерге қатысты семантикалық сипаттамаларды білдіреді.

Мәтіндер корпусы туралы жазылған ғылыми еңбектерде фонетикалық, морфологиялық, семантикалық, синтаксистік белгіленімдердің енгізілетіндігі туралы айтылады. Бірақ корпус жасау барысында аталған белгіленімдердің барлығын бір уақытта енгізу қиындық тудырады. Осы орайда шетел, орыс тіл біліміндегі мәтіндер корпусында лингвистикалық белгіленімдер енгізу ісі кезең-кезеңмен жүзеге асырылған. Толық лингвистикалық ақпарат берілген корпустарды «терең аннотацияланған» (глубоко аннотированный) деп атайды.

Екіншіден, корпус мазмұнының күрделілігі терең аннотациялануымен қатар әртүрлі стильді қамтуына да байланысты. Әдетте, корпус құрастыруда, сондай-ақ жиілік сөздіктер жасауда да көбінесе төрт түрлі стиль қамтылады. Олар: көркем стиль (проза, поэзия), драматургия, газет-журнал (публицистикалық), ғылыми-техникалық стильдер. Бұлардан басқа ауызекі стильден де корпус мәтіндері жинақталады. Үшіншіден, корпустардың сапасы ондағы қамтылған сөзқолданыс мөлшерімен де өлшенеді. Жалпы тіл білімінде алғашқы корпустардың кемінде 1 млн. сөзқолданыстан бастап жасалғандығы айтылады. Қазіргі кездері сол алғаш 1 млн. сөзқолданыстан жасалған мәтіндер корпусының көлемі 20 миллионнан 100 млн.-ға дейін жетеді екен. Демек, корпустың құрастырылуы туралы мәселе сөз болғанда, ең алдымен оның көлемі туралы нақты деректер беріледі.

Корпустар құрастыруда оның түріне қарай (Ұлттық, стильдік, кезеңдік) мәтіндер таңдалып алынады. Оларды таңдама мәтіндер деп атайды. Корпусқа енгізілетін мәтіндер көбінесе ақын-жазушылар шығармаларынан алынады. Корпустар құрастыру тәжірибесінде әсіресе, проза жанры басым. Сондықтан стильдік жағынан алғанда корпустарды «проза жанрына орталықтандырылған» (литературацентричный) деуге болады.

Корпустарға қойылатыны талаптардың негізгісі – репрезентативтілік (тұлғалылық), яғни оны сол корпустың пәндік аяның барлық қасиетін бейнелей алу мүмкіндігі немесе сол лингвистикалық зерттеу типіне қатысты пәндік аядағы құбылыстың кездесу жиілігінің тілдік бірліктерді бір-бірінен ажырата алатындай мәнде болу қажеттігі деуге болады [2].

Корпус түрлерінің ішінде қатар тілдер (параллель) корпустары бір тілден екінші тілге аударма жасауға қатысты талдау жұмыстарын жүргізуге аса қолайлы болып келеді. Мысалы, «Орыс тілінің Ұлттық корпусында» параллель мәтіндердің (қатар тілдер мәтіндері) корпустары да орын алған. Мұндай корпустар ерекше корпустар қатарына жатады. Себебі,

орыс тіліндегі мәтінге оның басқа тілге аударылған үлгісі және, керісінше, шет тілдеріндегі мәтіндерге орысша аудармасы сәйкестендірілген.

Түпкі және аударма мәтіндердің бірліктері арасында «теңестіру» нәтижесінде арнайы қарастырылған шаралар бойынша сәйкестік жүзеге асады. Теңестірілген паралельді корпус – ол ғылыми зерттеулердің, әсіресе, аударма жасаудың теориясы мен практикасының аса тиімді құралы.

Корпустық лингвистиканың жетістіктерін өзіне сақтаған аса дамыған корпус түрі – Ұлттық корпус. Мұндай корпус белгілі дәрежеде Ұлттық тілді толық түрде бейнелейді. Ұлттық корпустың репрезентативтілігі (тұлғалылығы) – сол тілдің жазба және сөйлеу түріндегі мәтіндерінің барлық типтерінің бейнеленуі. Ұлттық корпустың айтарлықтай дәрежеде көлемді (ондаған, жүздеген миллион сөзқолданыс) болуы репрезентативтілікке жетудің қажетті шарты болып саналады. Ұлттық корпустың ажыратылмас бөлігі оның белгіленген (аннотацияланған, мазмұндалған) бейнесі.

Теориялық және практикалық маңызы:

Ұлттық корпус, ең бірінші кезекте, тілші-ғалымдарға сол тілдің лексикасы мен грамматикасын жан-жақты зерттеуге мүмкіндік тудырады. Ал корпустың келесі міндеті – тілдің ішкі салалық (лексика, грамматика, тіл тарихы және т.б.) аясына қатысты әртүрлі анықтағыштық рөл атқару.

Егер Ұлттық корпуста тілдік бірліктердің статистикалық сипаты да берілетін болса, ондай деректермен әдебиетшілер, тарихшылар және басқа да қоғамдық ғылымдардың сала мамандары пайдалана алады.

Әрине, Ұлттық корпустың қолдану аясы тілдерді ана тілі немесе шет тілі ретінде оқыту кезінде көбірек байқалады. Сондықтан қазіргі кезде көптеген оқулықтар мен оқу бағдарламалары мәтіндік корпустарға бағышталып құрастырылуда. Мәселен, мағынасы күңгірт сөздер мен грамматикалық формалардың қолдану ерекшеліктерін белгілі авторлардың шығармалары бойынша электрондық корпус көмегімен әрі тез, әрі ұтымды тексеруді шетелдік азамат та, оқушы да, оқытушы да, журналист те және жазушы да жүзеге асыра алады.

Ұлттық корпус сол тілдің өмір сүрген белгілі кезеңіндегі сан алуан жанрын, стилін, аймақтық, әлеуметтік нұсқасын және т.б. да түрлерін қамтиды.

Корпустық лингвистика тіл білімінің жеке саласы ретінде өзімен іргелес жатқан тіл ғылымы пәндерімен жанасып жатады, яғни математикалық лингвистика, дискурстік анализ және лексикография салаларымен жақын жатады. Корпустық лингвистиканың басқа тіл ғылымы пәндерімен қарым-қатынаста болу ерекшелігі, бір жағына алғанда, мәтіндер корпусының корпустық лингвистика қызметінің нәтижесі ретінде болса, ал екіншіден – лингвистикалық пәндердің басқа түрлеріне де бастапқы эмпирикалық материал болу мүмкіндігінде. Міне, дәл осы жағдай корпустық лингвистиканың фонетика, лексикология, грамматика және стилистика салаларымен тығыз байланыста болуының дәлелі десек те болады.

Ұлттық корпус ғылыми зерттеулердің түр-түрін жүргізуді қамтамасыз етеді: лексикографияға, жасанды интеллектіге, әдебиеттануға, сөйлеу тілін талдау мен жинақтауға және лингвистиканың барлық салаларына қатысты зерттеу түрлері. Сонымен бірге беделді академиялық сөздіктер мен ғылыми грамматикаларды құрастыру да корпустар негізінде жүзеге асады. Ұлттық корпусты пайдаланушылар – әртүрлі саладағы тілшілер, әдебиеттанушылар, тарихшылар және гуманитарлық білім салаларының өкілдері. Ұлттық корпустың ана тілі мен шет тілін оқытуда, оқулықтар мен бағдарламалар құрастыруда маңыздылығы аса зор деуге болады [4; 5; 6].

Корпустық лингвистика қазақ тіл білімінің де ерекше саласы ретінде қалыптасатын болса, қазақ тілші-ғалымдарына көлемді тәжірибелік материалдарды пайдалануға, қажетті деген тілдік деректерді тауып алуға және оларға тиісті деген өңдеулер жүргізуге мүмкіндік туындатады. Осының бәрі қазақ тіліне қатысты зерттеулердің шынайылыққа (ақиқаттыққа)

жетудің эмпирикалық тәсілдеріне жаңаша көзқараспен қарауға және ғылыми айналым аясына аса маңызды тілдік материалдарды енгізуге жағдай жасайды.

Орыс тілі корпусы туралы:

Интернет желісіндегі 2003 жылдан бері өзіне жүктелген қызметті ойдағыдай атқарып келе жатқан «Орыс тілінің Ұлттық корпусын» атауға болады. Қазіргі кезде орыс тілінің Ұлттық корпусының жалпы көлемі 230 млн. сөзқолданыстан тұратын әртүрлі мәтіндер бөлігін қамтиды.

Орыс тілінің Ұлттық корпусы басқа да тілдердің Ұлттық корпустары сияқты мынадай екі маңызды ерекшеліктерге ие:

- біріншіден, корпусқа аса көлемді және әр кезеңдер бойынша шамалас көлемдегі мәтіндердің алынуы. Дәлірек айтсақ, біріншіден, орыс тілінің Ұлттық корпусына барлық жазбаша және ауызша мәтіндер (көркем әдебиеттің әртүрлі жанрлары, көсемсөз жанры, оқу, ғылыми, іскағаздары, сөйлеу тілі, аймақтық тіл және т.б.) қамтылып, олар әр кезең бойынша сәйкес көлемде алынады;
- екіншіден, корпус құрамындағы мәтіндер айрықша сипаттағы қосымша ақпаратқа ие болады. Мұндай ақпарат шартты түрдегі белгіленімдер арқылы көрініс табады (орысша аталуы – «разметка» немесе «аннотация»).

Орыс тілінің Ұлттық корпусының даму барысын сөз етсек, ол ең алдымен XIX ғ. басынан XXI ғ. бастапқы кезеңін қамтиды деуге болады. Бұл кезең орыс тілінің әртүрлі әлеуметтік-лингвистикалық тұстарын бейнелейді – әдеби тіл, сөйлеу тілі, тұрмыстық тіл, ішінара диалектілік тіл. Корпустық қорға көркем әдебиеттің (проза мен драматургия, поэзия) мәдени маңыздылығы мол және тілдік тұрғыда тілші-ғалымдардың қызығушылығын тудыратын түпнұсқа түріндегі шығармалары енгізіледі. Бірақ Ұлттық корпус тек көркем әдебиетке қатысты мәтіндерден ғана тұрмайды, ол сонымен бірге мәтін үлгілерінің басқа да жазба нұсқаларын (қазіргі кезеңде – ауызша нұсқасын да) қамтиды. Оған жататындар: мемуарлар, эсселер, көсемсөз стильдері, ғылыми-көпшілікке арналған және ғылыми әдебиеттер, жұрт алдында сөйлеген сөздер, жеке адамдар арасындағы хат алысу, күнделіктер, құжаттар және т.б. мәтіндер болуы мүмкін.

«Орыс тілінің Ұлттық корпусын» құрастыру үшін Ресейлік ғылым академиясының тек В. В. Виноградов атындағы Орыс тілі институты ғалымдарының күшімен ғана емес, оған Ресейдегі аса ірі ғылыми топтардың қатысқанын, дәлірек айтсақ, Мәскеу, Санкт-Петербург, Казань, Воронеж, Саратов және басқа да Ресейлік ғылыми орталықтардың көптеген ғалымдардың қауымдастығымен орындағаны мәлім болып отыр. Шындығында, 2003-2010 жылдары «Орыс тілінің Ұлттық корпусын» құрастыруға қолғабыс еткен ғылыми мекемелер:

1) Ресей ғылым академиясының тарихи-филологиялық («Филология және ақпараттану») бөлімі;

2) Ресейлік қоғамдық ғылыми қор;

3) Білім берудің федералды агенттіктері бойынша «Орыс тілі» федералдық мақсатты бағдарлама.

Сонымен бірге, В. В. Виноградов атындағы Орыс тілі институтының мамандарымен бірге жобаға басқа да мекемелер қатысқан:

1) РҒА-ның Тіл білімі институты [ИЯз РАН], РҒА-ның Ақпарат тарату мәселелері институты;

2) РҒА-ның Бүкілресейлік ғылыми және техникалық ақпарат институты [ВИНИТИ РАН];

3) Санкт-Петербургтегі РҒА-ның лингвистикалық зерттеулер институты;

4) Казань (Приволжский) федералды университеті;

5) Воронеж мемлекеттік университеті;

6) Саратов мемлекеттік университеті.

Осыншама мекемелер мен бірнеше ғылыми топтарының аталған жобаға қатысуына себеп, ол орыс тілінің жазба және сөйлеу тілі мәтіндерінің негізгі корпусын құруда көптеген мәселелерді қарастыру қажеттігінде. Қысқаша айтқанда олардың бір тобы мыналар:

1) XVIII ғасырдағы жазба мәтіндерінің тұлғалы корпусын құру;

2) XIX-XX ғасырдың бірінші жартысы аралығындағы жазба мәтіндерінің тұлғалы корпусын құру;

3) Қазіргі кезеңнің (XX ғ. ортасы – XXI ғ. басы) жазба мәтіндерінің тұлғалы корпусын құру;

4) Жазба мәтіндерінің корпустарын теңгеру (баланстау) үшін морфологиялық және сөзтудырушы-семантикалық белгіленім қағидаттарына (принциптеріне) зерттеме жүргізу және компьютерлік бағдарламалар мен корпустарға белгіленім жасауды қамтамасыз ету.

Осы аталған мәселелер Ресей елінің бірнеше ғылыми мекемелерінің бірнеше мамандары қатысып, шешімін тапқаны мәлім. Мысалы, мәтіндерге морфологиялық белгіленім жүргізудің ортақ принциптерін айқындау мәселесі үшін ғана 5 ірі ғалымдар атсалысқан (В. А. Плунгян, Г. И. Кустова, А. Е. Полякова және Д. В. Сичинава).

Сол сияқты, орыс тілі корпусына қажетті морфологиялық белгіленімді автоматтандыруды компьютерлік бағдарламамен қамтамасыз ету мәселесімен **Mystem** (Яндекс бірлестігі) және **Dialing** атты программалық қорын құрастырушы 10-нан аса прогаммист-ғалымдар ұжымы айналысқаны белгілі (Д. В. Панкратов, А. Е. Поляков, В. А. Титов, Т. А. Архангельский, А. И. Зобнин, А. В. Сокирко және т.б.). Ал осы компьютерлік бағдарламаларға қатысты морфологиялық талдаудың теориялық қағидаттарын зерттеуді Л.Л.Иомдин, В.З.Санников (Mystem), Н.Н.Леонтьева (Dialing) сияқты белгілі филолог-ғалымдар өз міндеттеріне алған болатын.

Сонымен, орыс тілінің Ұлттық корпусын құру мен оны жетілдіру ісін қажетті компьютерлік бағдарламалармен қамтамасыз ету мәселесіне, яғни іздестіру жүйесін, метамәтіндік белгіленімді, морфологиялық, синтаксистік, семантикалық белгіленімді және т.б. жетілдіру ісінің әр кезеңінде және алынған нәтижелерді эксперттен өткізу мәселелеріне көптеген ірі ғалымдармен бірге Мәскеудің жоғары оқу орындарындағы филолог-студенттер мен аспиранттар, магистранттар қатысып, өз үлестерін қосып отырғаны мәлім.

Орыс тілінің Ұлттық корпусын құрастыруда әр салаға қатысты белгіленім түрлерімен айналысатын орындаушылар тобы да сан жағынан түрліше. Мысалы, корпусқа сөзжасаушы-семантикалық белгіленім енгізу мәселесін зерттейтін ғылыми тобы 9 орындаушыдан тұрса, семантикалық белгіленімнің компьютерлік бағдарламалық құрамдау тобы 2 ғалымнан (А. Е. Поляков, А. И. Зобнин) тұрады екен. Келесі ғылыми топ метамәтіндік белгеленім мен мәтіндерді таңдаудың жалпы қағидаттарын зерделеумен шұғылданған ғалымдар саны 7-ге тең. Ал әр ғылыми топтардың орыс тілінің Ұлттық корпусы мәтіндеріне жүргізіп жатқан метамәтіндік белгіленімдерді бірізділігін координациялайтын ғалымдар тобы 20 шақты орындаушылардан тұратынын айта кетпекпіз.

Орыс тілінің кезеңдік корпустарын құрастыру үшін, мысалы, XIX ғасырдағы мәтіндер корпусы, XX ғасырдағы мәтіндер корпусын құрастыру кезіндегі өңдеу мен метабелгіленім жүргізуді де жеке ғылыми топтар жүзеге асырады. Сонымен, аталған міндеттерді орындайтын ғалымдар тобының саны мен әр топтағы ғалымдар саны да өне бойы өсіп отыратынын байқауға болады.

Орыс тіл білімінің тәжірибесіне сүйенсек, оларда орыс тілінің Ұлттық корпустарын құрастыру ісіне көптеген ғылыми-лингвистикалық, техникалық орталықтар, баспасөз, баспа, жоғары оқу орындары т.б. атсалысып, бірігіп атқарып отыр. 2003-2010 жылдарғы Ресейлік ғалымдардың «Орыс тілінің Ұлттық корпусы» жобасының қандай ғылыми күшпен орындалғанынан байқауға болады. Өйткені олар корпус құрастырудың маңызын өз кезінде жақсы түсініп, ауқымды істі бірігіп атқаруға жұмылдырылған. Нәтижесінде түрлі-түрлі лингвистикалық аннотациялар жасап, сонымен қатар мәтін көлемі жағынан да ұтып отыр.

Осындай қазақ тілінің Ұлттық сипаттағы «тұлғалы» тілдік корпустарын құрастыру мәселесі қазіргі кезде Қазақстанның бірнеше ғылыми-қолданбалы бағыттағы орталықтарында қолға алынып, дербес жұмыс істеп жатуы мүмкін. Олардың барлығы да орыс тілі тәжірибесіне сүйеніп, корпус құрастыру мәселесін өзінше шешемін деп талап қылып жатқанымен, ауқымды мәтіндерді компьютер жадына енгізу, лингвистикалық белгіленімдер талдамасын жасау ісінде шашыраңқылық танытатыны белгілі. Өйткені, әр

мекемеде жасалып жатқан корпустардағы лингвистикалық белгіленімдер мен олардың моделі, шартты белгілері бірізді емес.

Екіншіден, автоматты түрде лингвистикалық белгіленім қою мәселесі әлі де болса толық шешімін таппаған. Яғни тілдік талдаулардың өзінде де даулы мәселелер баршылық. Сондықтан аннотацияланған тілдік корпустарды құрастыру ісіне көптеген аса білімді практик лингвистерді тарту қажеттігі туындап отыр.

Үшіншіден, жоғарыда сөз болғандай, миллиондаған сөзқолданыстан тұратын корпустар құрастыру үшін аса көлемді мәтіндердің электронды варианты керек болады. Ал оларды «қолдан» енгізу көп уақытты қажет ететіні белгілі. Осы орайда бұл мәселе Қазақстан аумағындағы кітап, газет-журнал шығаратын баспалармен келісімге келе отырып шешілетін мәселе. Бұл мәселенің шешімін табу айтарлықтай оңай еместігі жоғарыда аталған «Қазақ тілінің Ұлттық корпусын» жасау кезінен таныс деуге болады. Сондықтан бұл мәселе тек ҚР БҒМ ҒК тұрғысынан ғана шешімін табуы мүмкін.

Аталған мәселе «Орыс тілінің Ұлттық корпусын» құрастыру жағдайында Ресей баспаларымен келісе жасау арқылы шешімін тапқан тәрізді. Оған негіз болып отырған Интернеттегі корпус құрастырушыларның Ресейлік 21 баспа орнына өз алғыстарын білдіргендігі: *«Разработчики Корпуса приносят благодарность следующим издательским коллективам и фондам, предоставившим для архива Корпуса электронные версии находящихся в их распоряжении текстов»* [6] (баспа аттарын келтірмеуді жөн санадық).

ҚР БҒМ ҒК А. Байтұрсынұлы атындағы Тіл білімі институтында қазақ тілінің корпусын құрастыру мәселесі «Мәдени құндылықтар ретіндегі қазақ тіліндегі мәтіндер корпусы және сөздіктердің «Тіл – қазына» атты Ұлттық компьютерлік қоры» атты тақырыпқа қатысты зерттеу жұмыстарынан бастама алған болатын. Аталған зерттеу жұмысының негізгі мақсаты – қазақ тілінің мәдени құндылығы болып саналатын толық мәтіндеріне, қажеттілікке сай, грамматикалық белгі-кодтар енгізіп, оның дербес түрдегі «Тіл – қазына» атты мәтіндер корпустарының компьютерлік базасын құру. Алғашында (2009-2011 ж.ж.) толық мәтіндердің компьютерлік қорының нысандары ретінде М. Әуезовтің, Ә. Кекілбаевтың, М. Мақатаевтың, М. Мағауинның толық шығармаларынан тек таңдама мәтіндер ғана алынды. Ал басқа қазақ классиктерінің, ғылыми мәтіндердің, публицистикалық шығармалардың мәтіндер қорпусын жасау Қолданбалы лингвистика бөлімдегі шағын ғана топтың қолынан келер нәрсе емес, әрине. Егер Институтымыздың қолға алған «Қазақ тілінің аннотацияланған Ұлттық корпусын» жасаушы ғалымдар саны жеткілікті болғанда мынадай корпустар түрлерін де жасауымызға болар еді:

1) Қазақ тілінің қазіргі кездегі (немесе кезеңдік) бұқаралық ақпарат құралдары (газет, журнал бетіндегі) мәтіндерінің жеке корпусы;

2) Қазақша сөйлеу тілі жазбасының (мәтінінің) жеке корпусы (орыс тілінің «Корпус живой русской речи» тәріздес);

3) Қазақ тілінің мультимедиалық корпусы (корпустың негізін мәтіндердің видео- және аудиожазбалары құрайды);

4) Қазақ тілімен параллель тілдердің жеке корпусы (түркітілдес және үндіеуропа тілдері), мысалы, қазақ-қырғыз, қырғыз-қазақ, қазақ-өзбек, өзбек-қазақ және т.б., сол сияқты, қазақ-орыс, орыс-қазақ, қазақ-украин, украин-қазақ және т.б. қатар тілдер корпусы;

5) Қазақ тілінің диалектілік мәтіндерінің жеке корпусы (орыс тілінің «Корпус русских диалектных текстов» тәріздес);

6) Қазақ тілінің поэтикалық мәтіндерінің жеке корпусы (орыс тілінің «Корпус русских поэтических текстов» тәріздес);

7) Қазақ тілінің білім беру корпусы (орыс тілінің «Обучающий корпус русского языка» тәріздес).

Бір айта кететін жайт – «Орыс тілінің Ұлттық корпусы» бойынша ақпарат іздеу жүйесін құру әрекетіне «Яндекс» компаниясы қолдау көрсеткені мәлім. Сол сияқты «Қазақ тілінің Ұлттық корпусынан» ақпарат іздестіруге және оның интернеттегі сайтының дизайнына да қолдау көрсететін компаниялар табылып жатса нұр үстіне нұр болар еді.

Қорыта келе айтарымыз: Ұлттық тіл мәтіндерінің компьютерлік корпусын құру жобасы бір ғана ғылыми ұйымның шешетін мәселесі емес және ол зерттеу жұмысы 3-5 жылда аяқтала қояды деуге де болмайды. Себебі бұл аса күрделі және оның нәтижелері әлемдік дәрежедегі аса маңызды ғылыми жұмыс болып саналады. Зерттеу жұмысының мақсатына сай орындалатын міндеттері де сала-салаға, кезең-кезеңге бөлініп, тек қана ғалымдар қауымдастығын құру арқылы ғана ауқымды нәтижеге ие боларымыз сөзсіз. Мемлекеттік тілдің өз деңгейінде қызмет етуін шындап мақсат етсек, тілімізді компьютердіру ісімен айналысатын жеке институт құрсақ та артықтық етпес еді.

Бір сөзбен айтқанда, қазақ тілінің тілдік корпустарын жасау – көп болып жұмылып атқаратын ұлттық құндылығымыз. Сондықтан Қазақстанның әр жерінде бір-бірінен дербес атқарылып жатқан корпус жасау ісін орталықтандыру керек немесе БҒМ Б. Жұмағұловтың сөзімен айтсақ, «Филологтар қауымдастығын құру» аса қажет демекпіз. Бұл ретте:

- оған Қазақстанның әр жерінде ғылыми-педагогикалық қызмет атқарып жүрген ғалымдардан арнайы **лингвистикалық топ** құру керек. Өйткені тілдік бірліктерді модельдеу – өте күрделі мәселе. Сондай-ақ тілдік корпустар құрастырудың өзі ең алдымен лингвистикалық белгіленім талдамасын жасауға тіреледі;

- осы кезге дейін жасалып жатқан корпус жасау тәжірибесіндегі **нәтижелерді** бір орталыққа жинақтау керек;

- кітап, газет-журнал шығаратын **баспалармен** шартқа отырып, олардан мәтіндердің электронды нұсқаларын алу қажет.

Сонымен қазақ тілінің Ұлттық корпусын жасау үшін А. Байтұрсынұлы атындағы Тіл білімі институтында арнайы орталық құрудың қажеттігі мен оның алғышарттарын атап көрсеттік. Ал мұндай ауқымды іске мемлекет тарапынан қолдау көрсетілсе, біртұтас қазақ тілінің Ұлттық корпусын құрастыру ісі алға басатыны сөзсіз. Мұндай Ұлттық құндылықты жасап шығару бүгінгі қазақ тіл білімінің ғана емес, қоғамның болашаққа қояр талаптарының бірі деп білеміз.

Әдебиеттер

1.http://www.edu.gov.kz/baspasz_yzmeti/silegen_szderi/silegen_sz/?tx_ttnews%5Btt_news%5D=5126&cHash=d3c4dcba878d7195a36e094ff8023dfe

2. *Баранов А.Н.* Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. –М.: Едиториал УРСС, 2003. С. 112-137.

3. *Захаров В.П.* Корпусная лингвистика: Учебн.-метод. пособие. –СПб., 2005. –48 с.

4. *Вербицкая Л.А., Казанский Н.Н., Касевич В.Б.* Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Серия 2. 2003. №6. –С.2-8.

5. *Шаров С.А.* Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2003. №6. –С.9-18.

6. Национальный корпус русского языка // http://www.ruscorpora.ru

ТҮРІК ТІЛДЕРІН ЛАТЫНДАНДЫРУ: СТАНДАРТТАР ЖӘНЕ ТЕХНОЛОГИЯЛАР
ЛАТИНИЗАЦИЯ ТЮРКСКИХ ЯЗЫКОВ: СТАНДАРТЫ И ТЕХНОЛОГИИ
LATINIZATION OF TURKIC WRITING : STANDARDS AND TECHNOLOGY

С.Н. БАЙБЕКОВ

Қазақ технология және бизнес университеті, Астана, Қазақстан

ҚАЗАҚ АЛФАВИТІНІҢ ЛАТЫН-АҒЫЛШЫН ГРАФИКАСЫНДАҒЫ ЖАҢА ЖОБАСЫ

Қазіргі таңда біреулер қазақ тілін дамытамыз десе, біреулер оған немқұрайлы қарайды, ал кейбіреулер, шынын айту керек, қарсы. Тоқетерін айтсақ, бұл мәселе – қазақ тілінің кәзіргі жағдайы, оның қоғамдағы орны туралы. Демек, қазақ тілінің дәл қазіргі жағдайы мәз емес. Неге? Оның бір ғана себебі бар. Ол - қазақ тілі техника тіліне келгенде шорқақ, ал қазір техника заманы. Яғни қазақ тілін дамытудың негізгі жолдарының бірі ол – техниканы қазақша сөйлету, немесе қазақ тілін техникаша сөйлету керек. Шындығын айтсақ, техника қазақ тіліне ешқашан бейімделген емес, бейімделмейді де. Осы себепті қазақ тілін техникаға - әлемдегі қолданыста жүрген қаптаған техникалық дүниеге бейімдеу қажет. Ол үшін алдымен текхниканың басты құралы компьютерді қазақша меңгеруіміз керек. Алғашқы мәселе компьютер клавиатурасының түймелерінде (клавишаларында) орналасқан әріптер мен таңбаларда. Себебі компьютер құдіреті осы таңбалардан басталады. Демек, қазақ алфавитін кириллицадан латын графикасы негізінде құралған 26 әріптен тұратын ағылшын алфавитіне (яғни, компьютер алфавитіне) ауыстыру керек. Бұл кезде қазақ әріптенінің саны 26-дан аспау керек, сонда қазақ тілі компьютер дүниесінде еркін өмір сүре алатын болады да, компьютерді қазақша сөйлетуге үлкен мүмкіншілік туады. Бұл өте маңызды мәселе. Ол баршамызға айтпай-ақ түсінікті. Осы жерде тағы да айта кететін мәселе – егер қазақ әріптерінің саны 26-дан бір ғана әріпке артық болып кетсе, онда қазақ алфавитінің компьютердегі мүмкіншілігі белгілі себептерге сәйкес күрт төмендеп кетеді, ондай реформадан пайда шамалы болады. Бұл реформаға қойылар *бірінші талап.*

Екінші шарт-талап. Реформа кезінде ұсынылмақ әріптер мен фонемаларды жаңаша белгілеу үлгілері оңай жатталып есте қаларлықтай ыңғайлы, әрі үйреншікті және қазақ тілі заңдылығына толығымен сәйкес болу керек. Егер олар әлемде қабылданған жүйеге мейлінше бейімді әрі үйлесімді болса – оған құба-құп.

Үшінші шарт-талап. Жаңа алфавит жүйесі техника дүниесіне ғана бейімделмей, ол осы күнге дейін қолданып келе жатқан ескі де жаңа ғылыми, тарихи, әдеби және басқа да мұрамызға жат болмай, олармен табиғи тығыз байланыста болу керек. Әсіресе ол кириллица негізінде жазылған жазба құндылықтарымыз осы жаңа алфавит негізіне оңай да тез аударылуына икемді болу керек.

Енді осы латын-ағылшын алфавитіне көшу тәсілін қарастырайық.

Алдымен қазіргі таңдағы 42 әріптен тұратын қазақ алфавитінен 12 төл әріптерді бөліп алып, оларды төмендегідей етіп латын-ағылшын алфавитіне ауыстырудың жобасын ұсынсақ, бұған ешкім қарсы болмас еді: ***Аа=Aa, Бб=Bb, Гг=Gg, Дд=Dd, Ее=Ee, Лл=Ll, Мм=Mm, Нн=Nn, Оо=Oo, Пп=Pp, Рр=Rr, Тт=Tt.*** Себебі бұл ұсыныс осы бағытта ізденісіп жүрген барлық авторлардың ойымен сәйкес келеді.

Осыдан кейін қазақтың басқа төл әріптері ішінен алдымен ***«у»*** әріпін қарастырсақ, онда көптеген авторлар арасында келісілген тоқтам жоқ. Себебі ***«у»*** дыбысы орыс тілінде дауысты болып табылады. Ал қазақ тіліне келсек кейбір оқулықтрда ол дауысты, ал басқа бір оқулықтарда дауыссыз деп саналады. Осы мәселе шешімінің бір варианты ретінде ұсынбағымыз – егер осы ***«у»*** дыбысы сөздің басында, немесе дауыссыз дыбыстан кейін орналасса, онда ол дауысты болады, ал егер дауысты дыбыстан кейін келетін болса, онда ол дауыссыз болады. Мәселен, *ауа, тауық, қауын ...*

Осы дыбысты кейбір авторлар ***«w»*** әріпі ретінде беріп жүр. Егер қазақтың төл сөздерін жазсақ, онда бұған келісуге болады. Ал егер қазақ тіліне еніп төлсөздей болып кеткен

сөздерді жазатын кезде бұған келісуге мүмкіншілік болмайды. Мәселен, *«университет, уран, ультра, утопия»* т.с.с. сөздерді алайық. Бұл сөздерді қазақ тілі үндестігіне сәйкес етіп *«үніберсітет, ұран, ултіре, ыутопійə»* деп жаз немесе айт деп кәзіргі таңда ешкімді көндіре алмаймыз. Осы тәрізді сөздер тілімізді байытады, оларды тіпті қазақ тіліне аудара да алмаймыз, оның қажеті де жоқ. Бұл сөздер әлдеқашан қазақша болып кеткен. Енді осы сөздерді ***«w»*** әріпі ақылы жазсақ (мәселен, «*wniversitet, wran, wltra»*), онда олар ерсі, үйлесімсіз, әрі қисынсыз болатыны көрініп тұрады. Демек ***«у»*** дыбысын ағылшынның өзіне тән ***«и»*** әріпімен белгілеу керек, яғни ***Уу=Uи***. Осы кезде жоғарыда келтірілген сөздер жазылуы әлемде қабылданған жүйеге дәлме дәл болады.

Енді ***«и,й»*** әріптеріне келейік. Бұл жерде де әртүрлі ұсыныстар бар. Әсіресе *«и»* дыбысы, бұл да *«у»* дыбысы сияқты біресе дауысты, біресе дауыссыз. Кейбір авторлар бұл дыбысты мүлдем алып тастайық деп ұсыныс береді. Сонда жоғарыда айтылғандай *«имам, импорт, импульс, индекс, инерция, интеграл, ион, изотоп»* сияқты қазақы болып кеткен сөздерді қалай айтамыз және қалай жазамыз?

Ал кейбір авторлар *«и»* дыбысын *«й», «і», «ы»* әріптерінің комбинациясы арқылы белгілеуді ұсынады. Сонда олар бұл дыбысты латын графикасында *«іу»* түрінде, ал кейбір кезде *«уі»* әріптерімен жазуды ұсынады. Бұндай ұсыныс жарамсыз. Себебі жазба тексте ағылшындық *«і»* және *«у»* әріптері қаптап кетеді. Екінші жағынан кейбір сөздерді жазғанда әріптер және буындар арасында көптеген үйлесімсіздектер пайда болып, оқырман шатасуға душар болады. Сондықтан ***«и»*** және ***«й»*** әріптерінің екеуінде латынның ***«і»*** әріпімен белгілеуді ұсынамыз, демек ***Ии(й)=Іі.*** Сонда жоғарыда айтылған және сол сияқты сөздер ешқандай шатасымсыз әлемдік бейімге сәйкес дәлме дәл жазылады, мәселен *«import»,* немесе *«ion».* Бұл жерде айта кететін тағы да бір жағдай - егер осы ***«и»*** әріпі сөздің басында, немесе дауыссыз дыбыстан кейін орналасса онда ол дауысты болады, ал егер дауысты дыбыстан кейін келетін болса, онда ол дауыссыз болады.

Енді ***«к»*** және ***«қ»*** әріптерін қарастырайық. Кейбір авторлар *«к»* әріпін латынның *«k»* әріпімен, ал *«қ»* әріпін *«q»* әріпімен белгілеуді ұсынады. Бұның бір ыңғайсыз жері *«Қазақстан»* атауын *«Qazaqstan»* деп жазсақ, онда бұл атау әлемде қабылданған жүйеге сәйкес болмай оғаштау көрінеді. Сондықтан біз алғашқыда осы атауды әдеттегідей етіп *«Kazakstan»* деп жазуды ұсындық. Демек, *Ққ=Kk, Кк=Qq.*

Қазіргі таңда қазақ тілі заңдылықтарын зерттей отыра, ***«Ққ»*** және ***«Кк»*** әріптерін латын графикасында ***«Kk»*** әріпі арқылы белгілеуді ұсынамыз. Бұл жерде аздап түсінбеушілік туындау мүмкін. Оны шешу үшін мынадай ереже ұсынамыз:

- Егер латынша жазылған ***«Кк»*** әріпі қазақтың латын баламасында жазылған «***а, о, ы, ұ***» әріптерімен көрші орналасса, онда ол ***«Ққ»*** әріпі болып оқылсын, демек ***«k(а, о, ы, ұ)= қ».***

- Егер латынша жазылған ***«Кк»*** әріпі қазақтың латын баламасында жазылған «***ə, ө, і, ү*** және ***е***» әріптерімен көрші орналасса, онда ол ***«Кк»*** әріпі болып оқылсын, демек ***«k(ə, ө, і, ү,е)= к».***

Бұл жерде осы қағидаға ***«и»*** және ***«у»*** әріптері сәйкес келмей, өзгешелік танытады. Мысалы, *«қиын»* және *«киім»* сияқты сөздер осы қағидаға бағынбайды. Сондықтан қазақ тілі үндестік заңдылықтарын пайдалана отырып, осы сөздерді келесідей етіп жазсақ *«қиын»=«қыйын»* және *«киім»=«кійім»*, онда қазақ тілі заңдылықтары жазба заңдылықтарына сәйкес және үйлесімді болып, бәрі орын орнына келеді де, жоғарыда айтылған ереже мүлтіксіз орындалады. Осыған сәйкес ***«у»*** әріпінің заңдылықтарын зерттей отырып, *«қуат»* және *«куə»* тәрізді сөздерді қарастырсақ, онда бұл сөздер *«қыуат»* және *«кууə»* болып жазылу керек те, қазақ тілінің үндестік заңдылықтары толығымен сақталып, жоғарыда айтылған ереже мүлтіксіз орындалады. Демек,

***«Kk**(а, о, ы, ұ)»= «Ққ».*

***«Kk**(ə, ө, і, ү, е)»= «Кк».*

Келесі мәселе ***«с»*** және ***«з»*** әріптері туралы. Бұл әріптер ағылшынның *«с», «s»* және *«z»* әріптеріне жақындау екені баршамызға мәлім. Сондықтан қазақтың *«с»* әріпін ағылшынның *«с»* әріпі арқылы белгілеу туралы да ұсыныс бар. Бұл біздің ойымызша дәлме-дәл балама

емес. Себебі бізге сіңіп кеткен *«сайт, сантиметр, сатира, сейф, синтетика, спорт, спираль, спирт, стадион, синус, секунд*» сияқты көптеген сөздер ағылшын тілі мәтінінде *«s»* әріпі арқылы жазылады. Тіпті өзіміздің төл сөзіміз *«сазан»* ағылшын тілінде аудармасыз осылай аталып, *«sazan»* деп жазылады. Ал *«з»* әріпіне келсек, онда *«зона, зоология, зебра, Зевс»* тәрізді сөздердің бәрі ағылшын тілінде *«z»* әріпі арқылы жазылады. Сондықтан қазақтың **«с»** әріпін ***«s»*** әріпі арқылы, ал **«з»** әріпін ***«z»*** әріпі арқылы белгілеуді дұрыс деп ойлаймыз (***Сс=Ss, Зз=Zz***).

Енді ***«ы»*** әріпін ағылшынның ***«y»*** әріпімен белгілесек, оған ешқандай қарсылық болмас. Мәселен, Almaty, Atyrau - бұл қабылданып қойған қағида тәріздес. Ал егер ***«ұ»*** әріпіне келсек, кезегін күтіп латын-ағылшын алфавитінің жеті әріпі қалды: *j, h, v, f, c, x, w*. Бұлардың алғашқы алтауы әлі де өз орындарын табатын болғандықтан, ***«ұ»*** әріпін ***«w»*** әріпімен белгілейік, бір жағынан түрлері де ұқсастау. Яғни ***Ыы=Yy*** және ***Ұұ=Ww.***

Енді ***«ә», «ө», «ү»*** және ***«і»*** әріптерін қарастырайық. Осы жерде профессор А.Шәріпбаев еңбегіне, дәлірек айтсақ оның «пирамидасына» сүйенеміз. Ол өзінің тілзерттеу бағытындағы еңбегінде дыбыстарды компьютер арқылы зерттей отырып ***ә, ө, ү, і*** аллофондары осыларға сәйкес дауысты дыбыстарға ***е*** фонемасын қосу арқылы шығатынын, яғни келесі төрт тепе-теңдікті ***ә=а+е, ө=о+е, ү=ұ+е, і=ы+е*** дәлелдеп, оны латын транскрициясына қолданды. Шәріпбаев ұсынған алфавитте ***ә, ө, ү, і*** аллофондарына әріп берілмеген. Оларды апостроф (') көмегімен белгілейді де, оны қажетті әріптен кейін орналастырады. Мәселен, «*өмір*» және *«қаракөз»* деген сөздерді былай жазады: *«o'my'r», «karaqo's»* (бұл жерде апостроф «*о»* және «*у»* әріптеріне әсерін тигізіп, оларды қазақтың *«ө»* және *«і»* әріптеріне айналдырып тұр). Біз осы жобаны қолдаймыз, демек ***Әә=A'a', Үү=W'w', Өө=O'o'*** және ***Іі=Y'y'***.

Бұл жерде айта кететін қағида – ұсынылмақ жаңа алфавит жүйесінде апостроф (') әсерін тек қана ***«а», «о», «ұ»*** және ***«ы»*** (яғни, ұсынылмақ жаңа алфавиттегі ***«а», «w», «о»*** және ***«у»***) әріптеріне ғана әсерін тигізе алады да, оларды төмендетіп ***«ә», «ү», «ө»*** және ***«і»*** дыбыстарына айналдырады.

Осы жерде тағы да айта кететін жағдай - қазақ тілінде осы келтірілген төрт дыбыстар бір сөздің ішіндеде бірнеше рет кездеседі, мәселен ***«көзілдірікті адам»=«ko'sy'ldy'ry'qty' adam».*** Осы мысалдан шығатын қорытынды - егер апострофты тура осылай қолданатын болсақ, онда жазуымыздың бәрі шүпірлеген апостроф-ноқаттарға толып кетіп, жазба мәтіннің көркемдігі төмендейді. Сондықтан жаңа алфавитте жазылатын мәтіндерде апострофтарды азайту үшін келесі ережені ұсынамыз: **егер апостроф белгілі бір әріптен кейін (мәселен, *«а», «о», «ұ»* немесе *«ы»*) орналаса, онда ол әсерін тек қана сол әріпке тигізеді, ал егер апостроф осы әріптердің біреуінің алдында орналасса, онда ол әсерін жазылып отырған сөздегі келесі әріптердің бәріне тигізеді.** Мысалы, ***«k'osyldyryqty adam», «сайгүлік=saig'wkyq»***

Енді ***«ж»*** әріпін қарастырайық. Көптеген авторлар ***«ж»*** дыбысын ***«j»*** әріпімен белгілеп жүр. Кейбір авторлар бұл әріпті «***zh»*** комбинациясы арқылы белгілеуді ұсынады. Біздің қолдауымыз ***Жж=Jj***. Себебі қазақ тіліне еніп, төл сөздей болып кеткен көптеген атаулар мен терминдер, мысалы *«жапон», «жаргон», «жихад»* және т.с.с. сөздер ағылшын және басқа тілдерде ***«j»*** әріпі арқылы жазылады. Егер біз оларды «***zh»*** арқылы жазсақ, онда ерсілеу болатыны анық. Демек *«japon», «jargon», «jihad».*

Келесі қаратырылмақ дыбыстар ***«ғ»*** және ***«ш».*** Бұл дыбыстар қазақтың *«г»* және *«з»* дыбыстарына ұқсас екенін ескере отырып, оларды ***ғ=gh*** және ***ш=sh*** арқылы белгілесік ешқандай қарсылық туа қоймас. Бұл жерлерде «***h»*** әріпі ***<u>күшейткіш</u>*** немесе ***<u>жуандатқыш</u>*** таңбасы ретінде қолданылып тұр. Осы жерде ***«щ»***-ны да қарастыра кетейік. Бұл қазақ тілінде өте аз қолданылатын дыбыс. Оны біраз авторлар қазақ тіліне тән емес деп жүр. Дегенменде *«ащы», «тұщы»* және сол сияқты сөздерді жазу кезінде біраз қиындықтар туады. Егер оларды *«ашшы»* және *«тұшшы»* деп жазсақ (яғни «*ashshy», «twshshy*»), онда біраз жазу көркемділігінен айырыламыз. Сондықтан оны ***щ=sch*** түрінде белгілесек, онда ол өз орнын табар еді. Демек***, Ғг=Ghgh, Шш=Shsh, Щщ=Schsch.***

Енді қазақ тілінде көп кездесетін ***«ң»*** дыбысын қарастырайық. Көптеген авторлар әртүрлі ережелерді ұсына отыра оны ***«ng»*** комбинациясымен жазуды ұсынып жүр. Бұл біз тарапынан қолдау таппайды. Себебі көптеген түсінеспеушіліктер пайда болады. Мәселен *«күнгей», «майдангер»* тағы сол сияқты қаптаған сөздер дұрыс оқылмай қалады, олардың жазылуын ережеден шығарып ерекшелік (исключение) ретінде жаттап алу мүмкін емес. Бір мысал: «*еріңе* (күйеуіңе)» және *«ерінге»* сияқты сөздер бірдей жазылып қалады*: «eringe»*. Тағы бір мысал: *«Сәкенге қара»* немесе осы сөздің сыйластық түрі *«Сәкеңе қара»* сөздері де бірдей жазылып, мәнісі жоғалады. Немесе *«Көкенге* (адам аты) *ұқса*» немесе *«көкеңе* (әкеңе) *ұқса*» сөздері де осындай толық шатасуға түседі. Сол сияқты *«qw'ngj»* деген сөз екі түрлі оқылып кетіп екі түрлі мағана береді: бірі – «*күңі*» (малайы), екіншісі – «*күнгі»* (әр күнгі тірлігі). Тағы бір мысал: «сол *qelinge* ризамын» деп жазсақ, шатасу пайда болады – *келіңе (*яғни *келіге)* риза ма, әлде *келінге* риза ма? Бұны ешқандай ережемен ажырата алмайсыз. Осындай мысалдарды жеткілікті түрде көптеп келтіруге болады. Мәселен, *«zhenggenge»* деп жазсақ ол екі вариантта оқылыды. Біріншісі – *«жеңгенге»* (жеңімпазға), екіншісі – *«жеңгеңе».* Сондықтан бұл дыбысты жоғарыда көрсетілген *«г, ш»* дыбыстарына ұқсастау етіп ***nh*** комбинациясымен белгілеуді ұсынамыз, демек ***ң=nh.***

Сонымен қазіргі қазақ алфавиті ішіндегі барлық төл әріптерді қарастырып біттік. Енді қалған кірме әріптерге тоқталайық. Алдымен ***«в»*** әріпін қарастырсақ, оны біраз авторлар қазақ тіліне жат деп, оны қолданыстан мүлдеп алып тастайық дейді. Онда қазақ алфавитін латын-ағылшын вариантына көшірген кезде осы әріпсіз *«болт»* және «*вольт»* сөздері бірдей (мәселен, ***«bolt»)*** жазылып, кездейсоқ қиыншылық туады. Осындай мысалдарды көптеп келтіруге болады. Демек, бұл дыбысты қазақ дыбыстарының қатарына қосып, оған да жеке әріп беру керек: ***Вв=Vv.*** Бұдан қазақ тіліне пайда келмесе, ешқандай зияндық келмейді, қазақ тілі заңдылықтары бұзылмайды. Осындай қорытындыны ***«ф»*** дыбысы үшін де толық айтуға болады. Шын мәнінде қазақ тіліне еніп, етене төл сөз болып кеткен аударылымы жоқ көптеген сөздерді (мәселен *«фильтр, финал, фотон, фокус, футбол, формат, формула, функция, фильм, факс*») бұрмалай жазсақ, одан бізге зиян болмаса пайдасы жоқ. Сондықтан осы дыбысты өз қатарымызға тартып, өзі сұранып тұрғандай оны ***«f»*** әріпімен белгілесек өте дұрыс болар деген ойдамыз, яғни ***Фф=Ff.***

Енді қазақ тіліне мүлдем жат ***«ц»*** әріпіне тоқталайық. Бұл жерде айтарымыз - *«ф», «ц»* әріптері орыс тілінде де ең аз қолданылатын әріптер қатарына жатады. Кейбір авторлардың айтуынша олар орыс тіліне де жат. Бірақ кезінде оларды орыс тілін техника тіліне <u>бейімдеу</u> үшін кіргізілгені анық. Ендеше бұл дыбыстарды жатырқамайық, біз де қазақ тілін техникаға бейімдейік. Жоғарыда айтылғандай қазақ тіліне аудармасыз кіріп, төл сөздер қатарына еніп кеткен көптеген сөздер (*циклоида, цилиндр, цензор, центр, церомония, цельсий, цемент, цент, целлофан, цикл* т.с.с) өзінің ағылшын вариантында *«с»* әріпі арқылы жазылады, мысалы, «*cement, cent, censor, cylinder*». Демек, *«ц»*-ны осы осы әріп алқылы белгілесек орынды болады, яғни ***Цц=Cc.***

Енді әдеби қазақ тілінде өте сирек кездесетін ***«ч»*** дыбысына келсек, көп зерттеушілер оны қазақ тіліне жат деп, қолданыстан алып тастауды ұсынады. Келісуге де болады, бірақ Алматы обылысының едәуір көпшілігі *«ч»*-деп сөйлейтінін еске алсақ, бұл дыбыс та сұранып тұр. Шын мәнінде бұл дыбыс қазіргі техника тілінің едәуір белсенді дыбыстарының бірі, мысалы «*чек, чемпион, чипс, чартер* (рейс)» және т.с.с. сөздер осыған куә және бұл сөздер ағылшын нұсқасында *«ch»* комбинациясы арқылы жазылады. Демек егер *«ч»* дыбысын қажет деп тапсақ, онда ***Чч=Chch.***

Егер ***«х»*** әріпін қарастырсақ, ол қазақ тілінде сирек кездесетін әріптердің қатарына жатады. Ол аздаған төл сөздерімізде ғана кездеседі, мысалы «*хабар, халық, хош иісті, хал-жағдай, хан, хат»* және т.с.с. Ал шет тілінен енген сөздерді алсақ «*Ханой, Хельсинки, хинди, хит, хот дог, хулиган»* және т.с.с. онда олар ағылшын алфавитінде ***«h»*** әріпі арқылы беріледі. Сондықтан көптеген авторлар осы дыбысты жатсынса да, біз оны өз алдына жеке әріппен белгілеуді жөн көрдік. Демек, ***«Хх=Hh»***. Бұл ешқандай қарсылық тудыра қоймас, яғни *«Hanoi», «Helsinki», «hindi», «hit», «hot dog»* және т.с.с

Осы жерде мынадай сұрақтың туындауы мүмкін. Жоғарыда қазақтың ***ғ, ш, щ, ң, ч*** әріптерін ***h***-әріпі көмегімен ***gh, sh, sch, nh, ch*** деп белгіледік. Ал енді қазақтың ***х*** әріпін ағылшынның ***h***-әріпімен белгілеп, яғни ***«х=h»*** деп отырмыз. Сонда бұл әріптер бір-бірімен шатасып кетпей ме деген сұрақ туады. Бұл сұраққа біздің бере жауабымыз. Біріншіден қазақ тілі заңдылықтары бойынша ***«ғ»*** әріпі мен ***«х»*** әріпі ешқашан қатар орналаспайды, сол сияқты ***«с»*** әріпі мен ***«х»*** әріпі, ***«с», «ц»*** әріптері және ***«х»*** әріпі, және де ***«н»*** әріпі мен ***«ц»*** әріпі ***«х»*** әріпімен ешқашан қатар орналаспайды. Яғни «***gh, sh, sch, nh, ch»*** комбинациялары таза қазақ сөздері үшін ешқандай қарама қайшылық тудырмайды. Екіншіден, егер ағылшын тілінен қазақ тіліне өткен терминдерді қарастырсақ, онда ***g*** әріпі ***h*** әріпімен өте сирек кездеседі де**,** олар қазақтың ***«ғ»*** әріпіне жақын дыбыс береді, мысалы мемлекет атауы *Гана=Ghana=/'ga:nə/*. Ал ***«sh»*** комбинациясына келсек**,** онда бұл екі әріп ағылшын тілінде ***«ш»*** болып оқылады. Сол сияқты «***ch»*** комбинациясы ***«ч»*** болып оқылады, мысалы *«матч=match»*. Енді ***nh*** комбинациясына келсек ағылшын сөздірінде бұл екі әріп ешқашан қатар орналаспайды**.** Ал ***sch*** комбинациясын қарастырсақ**,** онда ағылшын тілінен қазақ тіліне өткен терминдердің ешқайсында бұл үш әріп қатар кездеспейді. Бұдан шығатын қорытынды – ешқандай қарама-қайшылықтар болмайды. Тағы да айта кетеріміз - ***h***-әріпі көмегімен белгілеп отырған осы дыбыстар қазақ тілінде сирек кездесетін дыбыстар қатарына жатады. Демек ***h***-әріпі ұсынылып отырған алфавит бойынша өзіндік мөлшерде (аз да емес, көп те емес) қолданылады да, оның қолданыс жиілігі үйлесімдік шамадан асып кетпейді.

Тағы да тоқтала кететін жағдай. Жоғарыда айтылған *«в», «ф», «ц», «ч»* дыбыстарына ерекше мән беруіміздің негізгі себебі – олар кәзіргі заманның техникалық дыбыстары болып табылады. Олардың күннен күнге барлық тілдерге еніп жатқаны баршамызға мәлім. Бұл қарқын әсіресе ғарыш пен микроәлем игеріліп, нанотехнология дамыған сайын күшей береді. Олар дүние жүзілік жаңа терминдерді сипаттайтын әріптер болмақ. Ендеше оларды жатсынбайық.

Қалған дыбыстарды ***(h, э, ю, я, ё, ь, ъ)*** қазақ тіліне жат екенін ескере отырып және оларды пайдалану ешқандай нәтиже бермейтіндіктен, оларды қолданыстан шеттетуді ұсынамыз. Бірақ, қазіргі кириллица мәтінінде жазылған қазақы әдеби, ғылыми және басқа құндылықтарды ұсынылмақ жаңа алфавит мәтініне көшіру (аудару) үшін келесі баламаны ұсынамыз: *** Һ(қ)=k; э(е)=е; ю(йу)=iu, я(йа)=ia , ё(е)=е.*** Мұнда аударма кезінде *«я»* және *«ю»* әріптерінің салдарынан *«і»* екі рет кездесе қалса оның орнына бір-ақ *«і»* жазуды ұсынамыз. Мәселен, «*партия = partiia = partia*». Тағы да айта кетелік – бұл балама тек біржақты, яғни бұрынғы кириллицадағы текстерді латын алфавитетіне көшіру кезінде ғана қолданылады. Ал келесі белгілерді ***ь, ъ*** қолданыстан мүлдем алып тастауды ұсынамыз.

Оқырмандар тарапынан «Дүние жүзі *Canada, Coca Cola* деп жазып жүрген терминдерді және сол сияқты басқа сөздерді біз жаңа алфавит бойынша ерекшеленіп *Qanada, Qoqa qola* деп жазсақ ерсі болмай ма?» деген сұрақ болуы мүмкін. Осы мәселені шешу үшін төмендегідей ереже қабылдасақ бұның да шешімі табылар еді. Ағылшын жазба дүниесінде егер ағылшынның ***«с»*** әріпі ағылшынның ***«е»*** және ***«і»*** әріптері алдында орналасса, онда ол біздің ***«с»*** және ***«ц»*** дыбыстарына ұқсас үн береді. Ал осы ***«с»*** әріпі басқа кезкелген әріптердің алдында орналасса, онда ол қазақтың ***«к»*** дыбысындай дыбыс береді. Егер осы қағиданы қабылдасақ, онда біздің ұсынып отырған алфавитіміз әлемдік жүйеге сайма сай болмақ. Онда *«кафе=café», «камера=camera», «класс=class», «клуб=club», «компьютер=computer», «консультант=consultant», «контакт=contact»* тағы сол сияқты. Демек бұл жолы жаңа алфавит бойынша ***«с»*** әріпі ***«к»*** дыбысы болып оқылады. Бұл қағида тек қана қазақ тіліне енген териминдер үшін ғана қолданылады. Ал қазақи сөздер үшін ***«к»*** дыбысы ***«q»*** әріпімен белгілену керек.

Оқырмандар тарапынан тағы да *«Оксфорд», «такси», «максимум», «ксерокс»* сияқты қазақ тіліне еніп етене болып кеткен териминдердің жазылуы туралы бір сауал туындауы мүмкін. Әрине біз оларды елден ерекше етіп *«Oqsford», «taqsi»* деп жазбаймыз. Сондықтан бұл жерде ағылшын әріпінің ішінен әлі біздің алфавитке енбеген ***«х»*** әріпін пайданаланып, оны ***«кс»*** дыбысы үшін қолданамыз. Сонда әлемдік үйлесім де, жазба көркемділік те орын

орнына келе қалады. Онда *«Oxford», «taxi», «maximum», «Xerox»* тағы сол сияқты. Бұл қағида тағы қайталап пысықтайық, тек қана қазақ тіліне енген терминдер үшін ғана қолданылмақ.

Енді жаңа алфавит бойынша *«тойота», «символ», «камри», «дисплей», «Йорк», «Йемен», «йога», «йогурт»* тәріздес термин-сөздерінің жазылуын қарастырайық. Әрине бұл жолы да ерекше етіп *«toiota», «simvol» «camri»* деп жазбаймыз. Мүмкіншілігінше әлемдік жүйеге сәйкестікті қарастырамыз. Сол үшін тағы да бір ереже ұсынамыз. Ағылшын үндестігінде қазақы ***«ы»*** және ***«і»*** дыбыстары мүлдем кездеспейді. Сондықтан осы дыбыстарды белгілеу үшін ұсынылған ***«у»*** әріпіне қосымша мүмкіндік берейік – егер осы әріп қазақ тіліне енген териминдерде кездессе, онда ол ***«и»*** немесе ***«й»*** болып оқылсын. Егер осы қағиданы қабылдасақ бәрі орнына келеді де, жоғарыда келтірілген сөздер *«toyota», «symvol», «camry», «displey», «York», «Yemen», «yoga», «yogurt»* болып жазылар еді. Бұл қағида тек қана шет тілдерден ауысқан терминдер үшін қолданылмақ.

Тағы да әлемдік жүйеге сәйкес әрі үйлесімді болу үшін қазақ тіліне еніп үйреншікті болып кеткен *«квота, кварц, квант, квалификация, квартет»* және осы сияқты көптеген сөздерді қарастырайық. Бұл сөздердің алғашқы әріпі ағылшын тілінде *«q»* әріпі арқылы жазылатынын ескере отырып және халықаралық мәтінге сәйкес болу үшін оларды *«quota, quartz, quant, qualification, quartet»* деп жазсақ және оны қабылдасақ, онда шет тілінен ауысқан кірме териминдер үшін ***Qq=Кк.*** Осы жерде тағы да аңғаратын жағдай: ағылшынның ***«Qq»*** әріпі бұл жолы әрдайым ағылшынның ***«и»*** әріпінің (былайша айтқанда кирилицадағы ***«в»*** әріпінің) алдында орналасады.

Енді ұсынылмақ жаңа қазақ алфавитін кесте түрінде көрсетейік.

	1	2	3	4	5	6	7	8	9	10	11
Қазақ тілі фонемалары	**Аа**	**Бб**	**Гг**	**Дд**	**Ее**	**Лл**	**Мм**	**Нн**	**Оо**	**Пп**	**Рр**
Латын графикасындағы қазақ алфавиті	**Aa**	**Bb**	**Gg**	**Dd**	**Ee**	**Ll**	**Mm**	**Nn**	**Oo**	**Pp**	**Rr**

12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Тт	**Ққ**	**Кк**	**Ии,й**	**Зз**	**Сс**	**Жж**	**Уу**	**Вв**	**Фф**	**Цц**	**Ыы**	**Ұұ**	**Хх**	***[кс]***
Tt	**Kk**	**Kk(Cc,Qq)**	**Ii (Yy)**	**Zz**	**Ss**	**Jj**	**Uu**	**Vv**	**Ff**	**Cc**	**Yy**	**Ww**	**Hh**	**Xx**

Бұл кестеде ***ә, ө, ү, і*** әріптері келтірілмеген. Себебі олар жоғарыда айтылғандай ***a, o, w, y*** әріптері және апостроф арқылы берілмек. Сол сияқты ***ғ, ч, ш, щ, ң*** әріптері де жаңа алфавитке кірмейді. Бұл фонемалар үндестігіне сәйкес ***g, c, s, sc, n*** әріптері және жуандатқыш ***h*** белгісі арқылы беріледі. Олар, айта кету керек, ағылшын жазу үлгілеріне сәйкес келеді. Бұл олардың тәжірибе жүзінде қолданысын анағұрлым тиімді және жеңіл етеді. Ал ***ю, я, ё, э, h*** әріптері де жаңа алфавитке енгізілмейді. Олардың жоғарыда көрсетілген жаңа таңбалану үлгілері кириллица негізінде жазылған мәтіндерді ұсынылмақ алфавитке көшіру кезінде ғана пайдаланылады да, басқа қолданыстан шығарылып тасталады. Осы кестеге кірмеген келесі белгілер ***ь, ъ*** қолданыстан мүлдем алынып тасталады.

Ал ***«с», «у», «х», «q»*** әріптеріне келсек, онда олар жоғарыда айтылған ережелерге сәйкес кейбір кездерде қазақ тіліне енген терминдер үшін

«к», «и,й», «кс», «к» болып оқылады

Тағы айтарымыз – бұрынғы және қазіргі жазба құндылықтарды ұсынылмақ жаңа алфавит мәтініне көшіруді (аударуды) ешқандай қиындықсыз автоматтандыруға болады. Ол үшін осы

құндылықтар мәтінін компьютер жадына енгізу керек, яғни олардың электрондық варианты керек. Ал бұл болса кез келген кітапханаларда істеліниіп қойылған (немесе істеліниіп жатқан) нәрсе. Бар болғаны осы. Осыдан кейін осы электронды мәтінді жаңа жазба мәтініне лезде көшіруге болады. Бұның ешқандай қиыншылығы жоқ.

Осы айтылғандар жоғарыда берілген үш талапқа түгелдей сәйкес келеді.

Сөз соңында айтарымыз – қазақ алфавитін латын-ағылшын графикасына ауыстырудың осы ұсынылып отырған нұсқасы кемшіліксіз ең дұрыс нұсқа деп айтудан біз аулақпыз. Осындай нұсқалар неғұрлым көп болса - соғұрлым дұрыс болады. Онда олардың тиімді жағын қабылдап, осалдау жерлері болса, оларды бір-бірімен салыстыра отыра жетілдіріп, тиянақты бір шешімге келер едік. Бұл ел үшін, халық үшін және біздің болашағымыз үшін өте қажет те керек дүние. Ендеше іске сәт!

Д.Ш. СУЛЕЙМАНОВ

Академия наук Республики Татарстан и Казанский федеральный университет

ОБ АДЕКВАТНОМ АЛФАВИТЕ ДЛЯ ТАТАРСКОГО ЯЗЫКА: ЛАТИНИЦА ИЛИ КИРИЛЛИЦА

Аннотация

Татарский алфавит на основе латинской графики, который разрабатывался и активно обсуждался учеными и широкой общественностью в 90-е годы, был принят к реализации на уровне соответствующих правительственных решений и Закона Республики Татарстан об алфавите татарского языка латинской графике, в силу разных причин, так и не был допущен до практического применения. Являясь одним из разработчиков татарского алфавита и базовых программ национальной татарской локализации на основе латинской графики, автор в данной статье анализирует лингвистические и технологические аспекты разработки адекватного алфавита татарского языка на основе латинской графики и предлагает новый вариант татарского алфавита на основе латинской графики, состоящий из 26 букв.

Введение

Очевидно, что простая смена алфавита, его упорядочение и приведение в соответствие с базовыми закономерностями языка не является гарантией дальнейшего развития языка и его широкого применения. Привлекательность и живучесть языка - в его активности, и прежде всего в том, насколько язык является языком наук, информационных технологий, языком общения на официальном государственном уровне, языком передачи и усвоения знаний, языком межнационального и международного общения. Вместе с тем, адекватный алфавит, представляющий собой систему базовых элементов языка на стыке слышимого и видимого, является важным фактором, упорядочивающим вербальное проявление языка. Для укрепления фонетического иммунитета языка, для обеспечения его устойчивости важно, чтобы графемы соответствовали фонемам языка и в алфавите не было «мертвых» графем, не имеющих соответствующих фонем в языке.

Построение адекватного алфавита воспринимается нами как упорядочение основ орфоэпии и орфографии языка, приведение в соответствие фонетического и графемного рядов, очистка от «неработающих» букв.

1. Технологический аспект

Рассмотрим ряд аргументов в пользу татарского алфавита на основе латинской графики.

Во-первых, и это, возможно, самое важное, особенно в аспекте обучения языкам, переход на латиницу приводит к четкому графическому разграничению фонетических систем татарского и русского языков.

Очевидно, графическое разграничение фонетических систем татарского и русского языков полезно как для татарского, так и для русского языков. Непосредственное и постоянное соседство двух сильно различающихся фонетических систем в рамках одного алфавита приводит к диффузии, размыванию фонем, соответственно, к разрушению обоих языков. Чуть ли не все буквы кириллического алфавита, имеющие одинаковые начертания, по-разному звучат в русском и татарском языках. Например, как в следующих словах: акын, Казан, тын, чик, кол, кот, товар, карга, кит, тир, кара, имеющих определенные значения на русском и на татарском языках. В зависимости от того, к какому языку они относятся, татарскому или русскому, эти слова читаются совершенно по-разному.

Необходимо отметить следующее: здесь важно даже не столько переход на латиницу – сколько разграничение с кириллицей, т.е. переход на алфавит, отличный от кириллицы. То есть проблема разграничения татарского и русского алфавитов стоит даже независимо от того, к каким графемам переходить. Если бы татарский язык и английский были в такой же близости как русский и татарский, а русский был бы языком географически далекого народа как народы с английским языком, то с теми же обоснованиями в интересах татарского языка можно юыло бы переходить на кириллицу или на любой другой, отличный от латиницы, алфавит.

Но реальная ситуация такова, что стоит проблема разграничения именно татарского и русского алфавитов, базирующихся на кириллической графике. Латиница в данном случае устраивает по следующим причинам: 1) латиницу используют практически все народы, формирующие сегодня мировую науку, 2) на латинице базируются также и информационные технологии, 3) нет активного коммуникативного пересечения, близкого соседства, и активного и широкого взаимовлияния татарского и английского языков.

Дифференциация алфавитов по принципу: разные графемы - разные фонемы является удобным и при изучении языков. Переход на латиницу является выигрышным при изучении татарского языка как для татар, для которых русский язык, соответственно, и русский алфавит являются первым языком, первым алфавитом, так и для русских, желающих изучать татарский язык – легче будут усваиваться новые фонемы.

Во-вторых, для того чтобы выжить и развиваться в настоящее время (а в скором будущем тем более) татарский язык, должен войти в компьютерные технологии как язык накопления, обработки и передачи информации.

Очевидно, чем ближе алфавит татарского языка к алфавиту языка информационных технологий, каковым сегодня является латиница, и чем меньше промежуточных конвертаций (программных транслитераций - переводов с одного алфавита на другой) – тем он эффективнее, т.к. это приводит к экономии памяти для хранения и сокращению времени обработки – а значит, и экономически выгодно. Уменьшение промежуточных модулей (конверторов, таблиц перехода) приводит к увеличению надежности системы (меньше элементов – больше надежность). Как известно, в технических системах, время, память, надежность – это одни из самых критичных и важных показателей. Для татарского языка на основе латиницы – меньше проблем при использовании программ, обрабатывающих чисто латинические тексты. Меньше проблем с совместимостью и адаптацией данных и программ для обработки языковых блоков, с сортировкой, отображением на экране информации, с конвертацией текстов. Меньше проблем с татарской локализацией новых пакетов программ и операционных систем.

В-третьих, сокращение количества графем позволяет разработать более подходящую раскладку клавиатуры для татарского языка.

В-четвертых, упрощается общение в компьютерных сетях на татарском языке. Тексты на латинице, в отличие от кириллицы, читаются всегда корректно. В принципе, эта проблема

для кириллического текста разрешима и разрешается, однако требует дополнительных усилий по поддержке корректного отображения кириллических букв.

С переходом на латиницу татарские тексты будут читаться стабильно. Во всяком случае, если даже и не будет соответствующих драйверов и шрифтов, и не все буквы будут отображаться корректно, текст будет понятен, а это при переписке, а также и при ознакомлении с некоторым текстом, зачастую, самое главное: понять, что же там написано.

В-пятых, в настоящее время подготовлена необходимая документальная и программная база, регламентирующая использование татарской латиницы в компьютерных технологиях.

Разработана опытно-эксплуатационная версия пакета драйверов и шрифтового обеспечения для татарского языка на основе латиницы. Академией наук Татарстана подготовлены соответствующие материалы и принято Постановление Кабинета Министров РТ «О стандартах кодировки символов татарского алфавита на основе латинской графики и базовых программах для компьютерных применений» N 625 от 27 сентября 2000 года.

Разработана и реализована базовая версия конвертера татарского текста с кириллицы на латиницу и наоборот.

В-шестых, практически все языки развитых информационных технологий используют латиницу.

Татарский язык, в силу исключительной регулярности грамматики и агглютинативности, является удобным и весьма перспективным языком для компактного хранения и эффективной обработки языковой информации [1, 2, 3]. Следовательно, переход татарского языка на латиницу обеспечит естественное вхождение его в компьютерную среду, упростит использование татарского языка в качестве программного инструментария при создании интеллектуальных систем.

В-седьмых, почему нельзя просто модифицировать кириллический (русский) алфавит, адаптируя его к татарскому фонемному ряду?

Одна из причин, почему нельзя обойтись простым изменением кириллицы, «подгонкой» кириллического алфавита, заключается в следующем. Анализ кириллического алфавита с точки зрения соответствия русских букв татарскому фонемному ряду показывает, что для приведения кириллического алфавита в соответствие с татарской фонемной системой необходимо внести в алфавит не менее 20 изменений, т.е., добавить в русский алфавит не менее 20 новых символов, причем, удалив оттуда ряд «чужих» для татарского языка (таких как, «е, ё, я, ю, ц, щ, э, ь, ъ»). Изменению подлежат даже такие буквы как «а», «о», «в», «ч», «ы». Очевидно, это уже будет не измененный кириллический алфавит, а полная эклектика, т.е. новый алфавит, далекий как от кириллицы, так и от латиницы.

2. Лингвистический аспект

В данном разделе статьи осуществляется анализ татарского алфавита на основе латинской графики из 34 букв, утвержденного Указом Президента РТ и признанного в настоящее время экспериментальным вариантом, и предлагается новый вариант алфавита, состоящий из 26 букв.

В ряде писем в Академию наук Республики Татарстан и Комиссию по реализации языковой политики в компьютерных технологиях Комитета по реализации Закона «О языках народов РТ» при Кабинете Министров Республики Татарстан их авторами предлагалось использовать для отображения татарских букв только латинские буквы английского алфавита, а специфические татарские фонемы обозначать в виде сочетания латинских букв (например, как немецкие ‘Ш’ –‘ Sch’ и ‘Ч’ – ‘Tsch’, английские ‘Ч’ – ‘Ch’, ‘Х’ – ‘Kh’). Некоторые специалисты, учитывая закон сингармонизма гласных в татарском языке предложили специфические мягкие гласные отображать комбинацией букв - через твердые гласные плюс какая-либо буква или символ, обозначающие смягчение твердой гласной (например, ‘Av’ или ‘Ae’, или ‘Ah’, или ‘A`’ вместо татарской буквы ‘Ә’). Считая такое предложение достойным внимания, мы провели анализ и пришли к следующему выводу. Специфические согласные буквы из нового алфавита: ‘Ч’, ‘Ң’, ‘Ш’, ‘Ж’, действительно,

возможно отображать в виде комбинации латинских букв. Причем, для их обозначения второй буквой в качестве оператора можно использовать английскую букву ‘h’, по аналогии с принятой нормой в ряде языков на латинической основе (немецкий, английский) и достаточно редко встречающуюся в сочетаниях с согласными в татарских словах. Таким образом мы получаем следующие обозначения фонем: ‘Ч’ –‘Ch’, ‘Ң’ –‘Nh’,’Ш’ –‘Sh’, ‘Ж’ – ‘Zh’, являющиеся достаточно естественными и с точки зрения общепринятых норм написания. В этом ряду вне рассмотрения остается буква ‘Җ’, которую можно было бы оставить в том же виде, как и в принятом алфавите – ‘C’. Вместе с тем, для фонемы ‘Җ’ возможно ввести следующую комбинацию букв: ‘Jh’. Тогда буквой ‘C’ можно обозначить фонему ‘Ц’, которая в новом алфавите передается сочетанием букв ‘Ts’, или фонему ‘Ч’, которая гораздо более частотная для татарского языка (сравните: коэффициент частотности ‘Ч’ – 0,015, а фонемы ‘Ц’ – 0,0004). В том случае, когда ‘Jh’ передает звук ‘Җ’, буква 'J', во изменение нового алфавита, обозначает фонему ‘Й’ (‘и’ краткое). Соответственно, освобождается буква Y, которая традиционно обозначает букву ‘Ы’, например, как в слове ‘KRYM’, тем самым появляется хорошая возможность освободиться от целого узла неприятностей написания татарских букв ‘Ы’ и ‘И’, вернув для ‘И’ обычное написание, принятое в классическом латинском алфавите – ‘I’, а ‘Ы’ обозначив как ‘Y’. Эти буквы в новом алфавите отображены ‘İ ’ (‘I’ с точкой или «палка с точкой») и ‘I’ (‘I’ без точки или «палка без точки»), соответственно, и порождают массу неприятностей (особенно при разработке шрифтов и их распознавании, путаница с заглавной буквой ‘И’ в татарском и английском алфавитах, сложности в компьютерных технологиях и др.).

Далее рассмотрим такие одиозные буквы, за которые чуть ли не сердцем цепляются некоторые филологи – это твердые варианты букв ‘Г’ и ‘К’. Они введены в новый алфавит и обозначены буквами Ǧ - ‘G’ с ижицей (“галочкой”) и ‘Q’. Во-первых, создается прецедент, когда в алфавит вводятся пары согласных по принципу: «мягкий-твердый». Из истории хорошо известно, что у татар уже был такой алфавит – рунический, где все согласные были парные: «твердые-мягкие». Однако в настоящее время такой возврат, тем более, в нескольких буквах, не является обоснованным, так как твердость и мягкость татарских согласных в настоящее время однозначно определяется по закону сингармонизма, по мягкости-твердости контекста из гласных букв. Вместо того чтобы установить приоритет закона сингармонизма родного языка, вводя мягкие и твердые пары для букв ‘Г’ и ‘К’, мы сознательно идем на его нарушение, то есть отвергаем правила татарского языка только ради закрепления произношения, пришедшего из другого языка с его фонетическим рядом и ставшего близким нашему слуху (читай: слуху некоторых филологов). Введение твердых ‘Г’ и ‘К’ многократно усложняет разработку автоматического конвертора татарских текстов с татарской кириллицы на латиницу, требуя различать в тексте заимствованные и родные для татар слова (что, я думаю, будет не просто не только для машины, но и для носителей языка), а в некоторых случаях делает ее практически невозможной без обращения к смыслу предложения. Скажем, как можно было бы, даже прекрасно зная татарский язык, перевести с кириллицы на латиницу следующий текст: канатка канат сукты? Возможны 4 варианта: kanatqa kanat suqtı (‘канат ударил по канату’), qanatqa kanat suqtı (‘по крылу ударил канат’), qanatqa qanat suqtı (‘по крылу ударило крыло’), kanatqa qanat suqtı (‘по канату ударило крыло’), и все они верны с точки зрения языка.

Резюмируя сказанное, нами предлагается в новом алфавите оставить только варианты графем K, G, и пусть закон сингармонизма решает их морфонологию. Путаницы будет меньше, чем когда мы искусственно пытаемся внести порядок.

Такая путаница ожидает нас и с буквами ‘V’ и ‘W’. Фамилия 'Василов’ должна конвертироваться как ‘Wasilov’, а 'Васильев' – как ‘Vasil’ev’. Пытаясь уйти от одной сложности, мы порождаем другую, когда изучающий татарский язык, или даже носитель языка, всегда должен помнить, не является ли слово заимствованным. Это слово уже никогда не станет ему родным, за этим будет следить специальное правило. Лучшим выходом из этой ситуации было бы обогащение татарского языка новым словом через его

фонетическую ассимиляцию, отменив странное правило, когда заимствованные слова склоняются иначе, чем «родные», и они должны «помнить», что они «чужие».

Буква ‘W’ одна вполне справится с отображением фонемы ‘В’. Даже использование только буквы ‘V’ – прямого фонетического аналога кириллической буквы ‘В’ будет менее проблематично (во всяком случае, с точки зрения татарского языка). Во всяком случае, сложностей будет гораздо меньше, чем при одновременном использовании ‘V ‘и ‘W’.

Такая же ситуация с буквами ‘Х’ и ‘Н’. Даже носитель языка, хорошо владеющий татарским языком, зачастую не знает когда писать ту, и когда другую буквы – настолько в речи они схожи. ‘Х’, как правило, смягчается, а ‘Н’ произносится чуть тверже, чем при отдельном алфавитном произношении. В речи они практически сливаются.

Очевидный выход – положиться на тот же закон сингармонизма, оставить только ‘Н’, исключив ‘Х’. Тогда, например, получаем следующее написание слов: ‘hat’, ‘hərhəldə’ (сегодня литературные варианты: хат, һәрхәлдә). Именно вариант ‘h’ и выбран нами для включения в алфавит на основе латиницы, предлагаемого в данной статье. Вместе с тем, имеется очевидный аргумент против использования в латиническом татарском алфавите буквы ‘Х’. На латинице эта буква используется для обозначения фонемы ‘KS’, а не 'XA’, как в новом алфавите, принятом Законом Татарстана. То есть, включая букву ‘Х’ в алфавит, заранее готовим себе те же грабли - за одной и той же графемой уже в новом алфавите закрепляются две совершенно разные фонемы. В то же время известно, что в латинических алфавитах, как правило, фонема ‘Х’ обозначается комбинацией двух букв: ‘KH’, скажем, ‘Bukharaev’,‘Sakhibullin’, ‘Khayrullov’. Таким образом, получается, что фонема ‘Х’ из татарского кириллического алфавита переносится в новый алфавит на основе латиницы под той же «крышей», закрепив новую путаницу с фонемами. В такой ситуации очевидно, что букву ‘Х’ лучше не использовать в новом алфавите, а при необходимости (скажем, если даже и отображать данный аллофон в языке графически) для отображения фонемы ‘Х’ в латиническом написании использовать комбинацию букв: ‘Kh’

Переходя к гласным буквам, можно с уверенностью утверждать, что принцип отображения их комбинацией букв, или с использованием оператора мягкости или твердости здесь не проходит. Это прежде всего потому, что гласные буквы в татарском языке очень мобильны и они сами одновременно играют роль операторов мягкости или твердости. В отличие английского языка, имеющего главным образом корневое изменение и очень слабое аффиксальное, татарский язык, является агглютинативным языком, в котором к корню могут присоединяться аффиксальные морфемы (потенциально - без ограничений), в которых очень активны гласные фонемы. Словоформа, образованная присоединением к корню даже двух-трех аффиксов с фонемами (а это практически норма для языка), составленными по формуле «буква+оператор», становиться практически нечитабельной. В этом легко убедиться на простом примере: ‘kuhbahlahklahrgah’. Это простое слово ‘күбәләкләргә - бабочкам’ образовано с использованием смягчающего оператора ‘h’. Использование других символов тоже не разрешает ситуацию: ‘kuvbavlavklavrgav’,’ kuebaelaeklaergae’,’ ku`ba`la`kla`rga`’. Очевидно, что с татарскими гласными ни вариант комбинации букв, ни применение оператора не являются приемлемыми. В английском алфавите, хотя и имеются схожие фонемы, они передаются не графемой, а сочетанием букв, причем, не всегда одинаковым. Таким образом, использование специфических графем оказывается единственно приемлемым вариантом. Начертание мягких гласных можно оставить такими, как они приведены в новом алфавите ('a’ – ’ə’, ’u’ – ’ü’ (‘u’ умляут), ’o’ – ’ө’), однако более прогрессивным и удобным с точки зрения технологий представляется унифицированный вариант пар гласных букв: ‘a’ – ‘ä’ (‘a’ умляут), ‘u’ – ‘ü’ (‘u’ умляут), ‘o’ – ‘ö’ (‘o’ умляут). Вполне приемлемо для татарского языка (как это имеет место и в алфавитах других языков, скажем, буква I в английском алфавите), когда рукописные и наборные начертания мягких гласных различаются, поддерживая в рукописном варианте “прошлый” привычный образ гласных фонем (привычный - без “рожков и двоеточий”). В этом консерватизме, похоже, есть нечто стабилизирующее, связующее и обеспечивающее

некое смягчение перехода на новый алфавит, возможно, связанное с тонким телом, с кодом нации. В таком случае в алфавите в качестве примера рукописного варианта будут фигурировать пары мягких гласных: ‘Əə’, ‘Ää’; ‘Өө’,’Öö’;‘Yү’, ‘Üü’.

Ниже представлены таблицы с вариантами татарского алфавита на основе латиницы, состоящего из 26 букв, предлагаемыми автором статьи. Нами разработаны 3 варианта алфавита, различающиеся только лесико-графической упорядоченностью (ЛГУ) на основе различных принципов.

Вариант 1 - порядок графем скоординирован с порядком соответствующих графем алфавита английского языка (то есть на основе ЛГУ английского алфавита).

Вариант 2 - тот же алфавит по варианту 1 упорядочен (уточнен) дополнительно таким образом, что графемы сгруппированы по принципу фонемной близости.

Вариант 3 - порядок графем соответствует фонемному порядку алфавита русского языка. Данный вариант предложен в целях соотнесения татарского и русского алфавитов для удобства составления параллельных словарей, что особенно актуально в условиях функционирования двух государственных языков (то есть татарский алфавит на осное латиницы упорядочен на основе ЛГУ русского алфавита по фонемному принципу).

Вариант 4 – порядок графем определен на основе статистических данных частотности фонем (На статистическом материале, приведенном Бухараевым Р.Г., Ибрагимовым Т.И., Еникеевым А.И. в сборнике трудов [4]).

Вариант 5 – тот же алфавит (вариант 4) упорядочен по принципу фонемной близости графем.

Кроме того, для каждого варианта 1-5 разработаны по 3 уточненные таблицы (A, B, C), отображающие те же алфавиты для случаев, когда буква C используется для отображения фонем ‘Ц’,‘Җ’ или ’Ч’, соответственно.

На первый взгляд, букву ‘Ч’ целесообразно отображать привычной комбинацией ‘Ch’, что, как правило, и предлагается сторонниками передачи фонем в виде комбинации букв. Однако как показывает изучение частотности этих трех фонем и букв, фонема ‘Ч’ (коэффициент частотности: 0,015) в три раза чаще встречается в татарских текстах, нежели фонема ‘Җ’ (коэффициент частотности: 0,005) (фонема ‘Ц’, как считается, не присуща для литературного татарского языка), а буква ‘Ч’ даже еще более частотна, чем буквы ‘Җ’ и ‘Ц’. Коэффициент частотности буквы ‘Ч’ – 0,017, буквы ‘Җ’– 0,004, буквы ‘Ц’ – 0,0004. То есть буква ‘Ч’ встречается более чем в 4 раза чаще буквы ‘Җ’ и более чем в 40 раз чаще чем буква ‘Ц’. Этот показатель для буквы ‘Ц’ относительно ‘Ч’ может варьировать от 10-12 раз в технических текстах до более чем 400 раз в татарской прозе. Например, в произведении Н.Гиматдиновой «Пәри утарында» [5] буква ‘Ч’ встречается более 1200 раз, тогда как буква ‘Ц’ - всего в трех случаях.

Приведенные частотные показатели достаточно убедительно демонстрируют, что татарские тексты будут существенно короче, и также потребуется меньше нажатий на клавишу, когда именно фонема ‘Ч’, а не фонемы ‘Җ’ и ‘Ц’, будет отображаться одной буквой. В нашем случае – латинской буквой ‘C’ (как это было в Yanalif 1). Исходя из этого, наиболее предпочтительными во всех пяти вариантах нам представляются алфавиты, приведенные в таблицах C).

Вариант 1 (на основе английского алфавита)

A) Ц - C, Җ – Jh, Ч - Ch

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
A	Ä	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Ö	P	R	S	T	U	Ü	W	Y	Z
а	ә	б	ц	д	э	ф	г	h	И	й	к	л	м	н	о	ө	п	р	с	т	у	ү	в	ы	з

В) Ж - C, Ч - Ch, Ц – Ts

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
A	Ä	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Ö	P	R	S	T	U	Ü	W	Y	Z
а	ә	б	ж	д	э	ф	г	h	и	й	к	л	м	н	о	ө	п	р	с	т	у	ү	в	ы	з

С) Ч - C, Ц – Ts, Ж – Jh

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**264**
A	**Ä**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**Ö**	**P**	**R**	**S**	**T**	**U**	**Ü**	**W**	**Y**	**Z**
а	**ә**	**б**	**ч**	**д**	**э**	**ф**	**г**	**h**	**и**	**й**	**к**	**л**	**м**	**н**	**о**	**ө**	**п**	**р**	**с**	**т**	**у**	**ү**	**в**	**ы**	**з**

Вариант 2 (вариант 1, в котором графемы сгруппированы справа–налево по признаку фонемной близости)

А) Ц - C, Ж – Jh, Ч - Ch

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	B	P	C	Z	S	E	Y	I	J	F	W	U	Ü	K	G	H	L	R	M	N	O	Ö	T	D
а	ә	б	п	ц	з	с	э	ы	и	й	ф	в	у	ү	к	г	h	л	р	м	н	о	ө	т	д

В) Ж - C, Ч - Ch, Ц – Ts

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	B	P	C	G	K	H	D	T	E	Y	F	W	I	J	L	R	M	N	O	Ö	S	Z	U	Ü
а	ә	б	п	ж	г	к	h	д	т	э	ы	ф	в	и	й	л	р	м	н	о	ө	с	з	у	ү з

С) Ч - C, Ц – Ts, Ж – Jh

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**264**
A	**Ä**	**B**	**P**	**C**	**K**	**G**	**H**	**D**	**T**	**E**	**Y**	**F**	**W**	**I**	**J**	**L**	**R**	**M**	**N**	**O**	**Ö**	**S**	**Z**	**U**	**Ü**
а	**ә**	**б**	**п**	**ч**	**к**	**г**	**h**	**д**	**т**	**э**	**ы**	**ф**	**в**	**и**	**й**	**л**	**р**	**м**	**н**	**о**	**ө**	**с**	**з**	**у**	**ү з**

Вариант 3 (на основе русского алфавита по фонемному принципу)

А) Ц - C, Ж – Jh, Ч - Ch

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	B	W	G	D	Z	I	J	K	L	M	N	O	Ö	P	R	S	T	U	Ü	F	H	C	Y	E
а	ә	б	в	г	д	з	и	й	К	л	м	н	о	ө	п	р	с	т	у	ү	ф	h	ц	ы	э

В) Ж - C, Ч - Ch, Ц – Ts

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	B	W	G	D	C	Z	I	J	K	L	M	N	O	Ö	P	R	S	T	U	Ü	F	H	Y	E
а	ә	б	ж	д	з	ф	г	h	И	й	к	л	м	н	о	ө	п	р	с	т	у	ү	в	ы	э

С) Ч - C, Ц – TS, Ж – JH

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**264**
A	**Ä**	**B**	**W**	**G**	**D**	**Z**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**Ö**	**P**	**R**	**S**	**T**	**U**	**Ü**	**F**	**H**	**C**	**Y**	**E**

а	**ә**	**б**	**в**	**г**	**д**	**з**	**и**	**й**	**к**	**л**	**м**	**н**	**о**	**ө**	**п**	**р**	**с**	**т**	**у**	**ү**	**ф**	**h**	**ч**	**ы**	**э**

Вариант 4 (на основе частотности фонем)

А) Ц - С, Ж – Jh, Ч - Ch

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	N	Y	E	R	L	K	T	I	J	G	M	B	D	S	U	Z	P	Ü	Ö	O	W	F	H	C
а	ә	н	ы	э	р	л	к	т	и	й	г	м	б	д	с	у	з	п	ү	ө	о	в	ф	h	ц

В) Ж - С, Ч - Ch, Ц – Ts

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	N	Y	E	R	L	K	T	I	J	G	M	B	D	S	U	Z	P	Ü	Ö	O	W	C	F	H
а	ә	н	ы	э	р	л	к	т	и	й	г	м	б	д	с	у	з	п	ү	ө	О	в	ж	ф	h

С) Ч - С, Ц – Ts, Ж – Jh

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**264**
A	**Ä**	**N**	**Y**	**E**	**R**	**L**	**K**	**T**	**I**	**J**	**G**	**M**	**B**	**D**	**S**	**U**	**Z**	**C**	**P**	**Ü**	**Ö**	**O**	**W**	**F**	**H**
а	**ә**	**н**	**ы**	**э**	**р**	**л**	**к**	**т**	**и**	**й**	**г**	**м**	**б**	**д**	**с**	**у**	**з**	**ч**	**п**	**ү**	**ө**	**о**	**в**	**ф**	**h**

Вариант 5 (по созвучию фонем на основе алфавита по частотности)

А) Ц - С, Ж – Jh, Ч - Ch

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	N	M	Y	E	R	L	K	G	H	T	D	I	J	B	P	S	Z	C	U	Ü	O	Ö	W	F
а	ә	н	м	ы	э	р	л	к	г	h	т	д	и	й	б	п	с	з	ц	у	ү	о	ө	в	ф

В) Ж - С, Ч - Ch, Ц – Ts

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	264
A	Ä	N	M	Y	E	R	L	K	C	G	H	T	D	I	J	B	P	S	Z	U	Ü	O	Ö	W	F
а	ә	н	м	ы	э	р	л	к	ж	г	h	т	д	и	й	б	п	С	з	у	ү	о	ө	в	ф

С) Ч - С, Ц – Ts, Ж – Jh

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**264**
A	**Ä**	**N**	**M**	**Y**	**E**	**R**	**L**	**K**	**C**	**G**	**H**	**T**	**D**	**I**	**J**	**B**	**P**	**S**	**Z**	**U**	**Ü**	**O**	**Ö**	**W**	**F**
а	**ә**	**н**	**м**	**ы**	**э**	**р**	**л**	**к**	**ч**	**г**	**h**	**т**	**д**	**и**	**й**	**б**	**п**	**С**	**з**	**у**	**ү**	**о**	**ө**	**в**	**ф**

Ниже приведена таблица, отображающая два алфавита: Yanalif 2, принятый по Указу Президента Республики Татарстан и признанный в настоящее время экспериментальным, и Yanalif-SLT (26 букв), предлагаемый в качестве варианта.

	Yanalif-2	Произно-шение	Кирил-лица	Yanalif-SLT	Произно-шение	Кирил-лица
1.	A, a	a	А, а	A, a	A	А, а
2.	Ә, ә	ә	Ә, ә	Ä, ä; Ə, ə	ә	Ә, ә

3.	B, b	be	Б, б	B, b	Be	Б, б
4.	C, c	ce	Ҷ, ҷ	C, c	Che	Ч, ч
5.	ζ, ς	che	Ч, ч	-	-	-
6.	D, d	de	Д, д	D, d	De	Д, д
7.	E, e	e	Е, е (э)	E, e	E	Е, е (э)
8.	F, f	ef	Ф, ф	F, f	Ef	Ф, ф
9.	G, g	ge	Г, г	G, g	Ge	Г, г
10.	Ğ, ğ	ğı	Гъ, гъ	-	-	-
11.	H, h	he	Ҳ, ҳ	H, h	He	Ҳ, ҳ
12.	I, ı	ı	Ы, ы	-	-	-
13.	İ, i	i	И, и	I, i	I	И, и
14.	J, j	je	Ж, ж	J, j	Jod	Й, й
15.	K, к	ke	К, к	K, к	Ke	К, к
16.	Q, q	q	Къ, къ	-	-	-
17.	L, l	el	Л, л	L, l	El	Л, л
18.	M, m	em	М, м	M, m	Em	М, м
19.	N, n	en	Н, н	N, n	En	Н, н
20.	Ŋ, ŋ	eng	Ӈ, ӈ	-	-	-
21.	O, o	o	О, о	O, o	O	О, о
22.	Ɵ, ɵ	ɵ	Ɵ, ɵ	Ö, ö; Ɵ, ɵ	ɵ	Ɵ, ɵ
23.	P, p	pe	П, п	P, p	Pe	П, п
24.	R, r	er	Р, р	R, r	Er	Р, р
25.	S, s	es	С, с	S, s	Es	С, с
26.	Ş, ş	şa	Ш, ш	-	-	-
27.	T, t	te	Т, т	T, t	Te	Т, т
28.	U, u	u	У, у	U, u	U	У, у
29.	Ü, ü	ü	Ү, ү	Ü, ü; ɣ, ɣ	Ü	Ү, ү
30.	V, v	ve	В, в	-	-	-
31.	W, w	we	В, в (у)	W, w	We	В, в (у)
32.	X, x	xa	Х, х	-	-	-
33.	Y, y	y	Й, й	Y, y	Ы	Ы, ы
34.	Z, z	ze	З, з	Z, z	ze	З, з

Ц – Ts, Ҷ – Jh, Ш – Sh, Щ – Tc/ Tsh, Ж – Zh, Ӈ - Nh

Известно, что порядок расположения букв и их численное значение имеют огромную важность в эзотерических толкованиях [6, 7], что все чаще находит свое научное обоснование. Поэтому, при окончательном утверждении алфавита представляется необходимым (целесообразным) проведение анализа не только состава алфавита (количества графем и их начертаний), но и иследование порядка следования графем и расположение их исходя из метафизического смыла их численных значений, проверяя их соответствие на ключевых словах мироздания. Тем более, что разница между числовыми значениями по частотности в большинстве своем настолько несущественны, что они вполне могут быть смещены в определенных пределах, возможно даже, что полученные данные могут изменяться (уточняться) в ту или другую сторону при увеличении тестового объема текстов. (Кроме фонемы *‘A’*, имеющей наибольший коэффициент встречаемости: 0,113, и занимающей безусловно первую позицию. Другие фонемы по частотности расположены в следующих интервалах: от *‘Ä’*(0,072) до *‘T’*(0,050)*;* от *‘I’*(0,042) до *‘U’*(0,025)*;* от *‘Z’*(0,016) до*‘O’*(0,010)*;* от *‘W’*(0,007) до *‘H’*(0,003). Расположение графем по варианту 4 С).

Очевидно, что такая работа сама по себе очень емкая, ответственная и требует проведения глубоких исследований на основе огромного текстового материала с разработкой лингвистических моделей и применением математического аппарата.

Заключение

На семинарах Лаборатории проблем искусственного интеллекта АНТ и КГУ обсуждались разные варианты латиницы, и даже был выпущен специальный сборник [4], в котором, в частности, анализируется ряд вариантов алфавита на основе латиницы и предлагается некий собственный алфавит, предложенный специалистами лаборатории. Однако, поддерживая переход на латиницу в принципе, и беспокоясь, что в мелких спорах (и как оказалось - обоснованно) можно потерять саму возможность перехода на латинскую графику, мы практически отказались от продвижения своего варианта.

Пользуясь тем, что эксперимент по поводу перехода на латиницу продолжается, в данной статье проанализированы технические и лингвистические аспекты перехода на латиницу, и предложен новый вариант татарского алфавита на основе латиницы, состоящий из 26 букв.

Литература

1. Сулейманов Д.Ш. Формальная элегантность и естественная сложность морфологии татарского языка // Электронная конф.: Информационные технологии в гуманитарных науках (Казань, 5-31 мая, 1998г). -Казань, 1998. - HTTP://www.kcn.ru/_tat_ru/universitet/gum_konf/ot7.htm.

2. Сулейманов Д.Ш. (Кыпчак И.И.). Татарский язык - язык компьютерных технологий // Материалы научн.-практ. конференции “Языковая ситуация в республике Татарстан: состояние и перспективы” (17 ноября 1998 года). В 2 частях. Ч.1. -Казань: Изд-во “Мастер Лайн”, 1999. -196-198.

3. Suleymanov D.S. Natural possibilities of the Tatar morphology as a formal base of the NLP // In Proceedings of the First International Workshop “Computerisation of Natural Languages” (Varna, Sept. 3-7, 1999). –Sofia (Bulgaria): Information Services Plc, 1999. -P.113-117.

4. Татарский язык и информационные технологии. Вып.2. Изд-во Каз.гос. ун-т. –1995. – 122 с.

5. Гыйматдинова Н. Пәри утарында // «Идел» журналы, № 10, 1996. – 4-17, 32-39 б.

6. Папюсъ. Каббала или наука о боге, вселенной и человеке /Пер. С фр. А.В.Трояновского. Издание В.Л.Богушевского. –СПБ, 1910. – 359 с.

7. Плешанов А.Д. Руский алфавит как инструмент научного познания вселенной. –М.: Новый Центр, 2000. - 68 с.

Ә. ЖҮНІСБЕК

А. Байтұрсынұлы атындағыТіл білімі институты, Мемлекеттік тілді дамыту институты, Алматы, Қазақстан

КІРМЕ (БҮЛДІРГІ) КИРИЛЛ ТАҢБАЛАРЫНЫҢ ДЫБЫС ТАЛДАНЫМЫ

1.Алдымен қазақ тілінің төл дыбыс құрамы ғылыми тұрғыдан анықталып, ***төл дыбыстар*** өз алдына, қазақ тіліндегі ***кірме (бүлдіргі) дыбыстар*** мен ***әріптер*** өз алдына топтастырылу керек. Өйткені кірме дыбыстар мен таңбалар қазақ сөзінің айтылым әуезін, морфем құрамын, буын тұрқы мен тасымал ретін, тіптен, сөйлеу (сөйлем) ырғағын бұзып бітті. Сонда қазақ тілінің фонетика деңгейінен (дыбыс құрамы) бастап, лексика, морфологияны қамтып, синтаксиске дейінгі табиғи қалпын сетінетіп отыр. Ендеше, бірінші кезекте, төл

дыбыстардың құрамы мен әріп таңбалары анықталып, қазақ тілінің төл «Әліпбиі» түзіледі. «Әліпби» төл дыбыстар құрамын түгел қамтып, олар үйлесімді орналастырылуға тиіс.

Қазақ тілінің төл дыбыс құрамы:

А	**Ә**	**П**	**Б**	**М**
Ы	**І**	**Т**	**Д**	**Н**
	Е	**Қ(-К)**	**Ғ (-Г)**	**Ң**
Ұ	**Ү**	**С**	**З**	**Р**
О	**Ө**	**Ш**	**Ж**	**Л**
				Й
				У(w)

Мұндағы ***у*** таңбасының дыбыс мәнін латын әліпбиінің құрамындағы ***w*** әрпімен белгілеп беріп отырмыз. Өйткені қазақ тілінде ***ерін-еріндік, жуысыңқы, үнді*** дауыссыз бар. Ол дауыссыз халықаралық практикада осы таңбамен беріледі. Ондай дыбыс ағылышын тілінде де бар. Ағылшын тілінде сауатты өсіп келе жатқан жас ұрпақ оны тез аңғарып, жеңіл қабылдайды. Ал орыс тілінде ондай дауыссыз дыбыс жоқ. Олардың ***Көкшетау*** дегенді ***Көкчетав*** деп өз тіліндегі ***тіс-еріндік, жуысыңқы, ұяң в*** дауыссызымен алмастыратындығы сондықтан. Қазақ тіліндегі дауыссыз ***у[w]*** мен орыс тіліндегі дауысты ***у[u:]*** дыбыстарын естілім ұқсастығына қарап өзара шатастырып жүргеніміз сондықтан. Тіптен ***и*** мен ***у*** әріптерін (дыбыстарын емес) өз алдына дауысты «фонема» деуге де жеттік. Оны қазақ тілінің дамуы деп түсіндіруге тырысып жатырмыз.

2.Қазақтың төл сөздерінің жазылымындағы кірме әріптер түгел шығарылып тасталады. Олар:

И У Я Ю Щ Х Һ

Өкінішке орай, орыс тілінің емле-ережесі арқылы қазақ тілінің төл сөздерінің жазылымына енген ***и, у, я, ю*** таңбаларының (дыбыстарының емес) әрқайсысын өзалдына дербес дауысты дыбыстар деп қабылдау бар. Көпшілік (тіптен, тіл мамандарының өздері, қазақ тілі пәнінің мектеп мұғалімдері мен жоғары оқу орындарының оқытушылары) осыған әбден ден қойып, психологиялық тұрғыдан мойындап алған. Олай етпеске амалы да жоқ, өйткені мектептен бастап, жоғары оқу орнын бітіргенше оларды дауысты дыбыс деп санасына құя берсе, кімде болса сеніп қалады. Оның үстіне оқу бағдарламалары мен оқулықтар түгелдей осы ережені құптап отыр. Қазақ тілі дыбыстарының құрамына зорлықпен енгізіліп отырған «дауысты дыбыстардың» жөнін табу үшін олар «дауыстыдан кейін дауыссыз болады, дауыссыздан кейін дауысты болады» деген (мұндайда болады екен!?) әлеми жұртқа күлкі болатын ереже де ойлап таптық.

Оның зардабын әлі көріп келе жатырмыз. Тіл мамандарының өзі өздері ұсынып отырған әліпби жабаларында осы жалған қағиданы ұстанып отыр. Мысалы, орыс жазуының ережесімен жазып келген ***ми, ти*** т.б. сөздерді латын негізінде де ***mi, ti*** деп жазып, ***ми, ти*** деп оқуды ұсынып отыр. Сонда қазіргі орыстақы емлеміз бойынша да ***[ми, ти]***, жаңа латынша да ***[mi, ti]*** болып шыққалы тұр. Жоба авторлары осы жолмен жазу желісін үнемдеуге болады, яғни екі таңбаның орнына бір таңба ғана жазамыз деген болжамды алға тартады.

3.Жоба авторларының дәйектемесіне көз жеткізу үшін «Абай жолы» романы, Әнұран т.б. мәтіндерге есептеу жүргізіп көрдік. Сондағы үнем 0,5-0,7% ғана болып шықты. Сонда түкке тұрғысыз үнем үшін, егемендіктің арқасында ғана құтылғалы отырған, тілбұзар емле-ережелерді тағы несіне қайталаймыз деген ой туады. Оның үстіне өздері ұрынған шым-шытырықтан шығу үшін құрау-құрау таңбалар ұсынады. Мысалы, ***ы[ı], і[i], ӱ[ï], ұ[u], ү[ü], у[û].*** Мұндай шешім әдістемелік тұрғыдан өте тиімсіз. Өйткені ұқсас таңбалардың үстіне үстеме белгілер қоя беруге болмайды. Себебі сауат ашуды

қиындатып жібереді: ұқсас таңбалар екіден аспау керек; үстеме таңба біріне қойылып, екіншісіне қойылмау керек. Ал, жазу реформасының басты мақсаты орыс тілінің тіл бұзар ережелерінен арылып, қазақтың ***мый*** деген сөзін айтылуына лайық ***мый*** деп жазып, жуан айтқызу, ***тій*** деген сөзін айтылуына лайық ***тій*** деп жазып, жіңішке айтқызу үшін жасалып отырғанын ескерсек, әріптестеріміздің бұл әрекетіне қалайша «жол болсын» демекпіз.

4.Кез-келген жазу тілдің ішкі өзіне тән құрылымы мен табиғи айтылымына зиян тигізбеу керек. Керісінші, тілдің әдеби айтылымының қорғанышы болу керек. Ол үшін:

1. Жазу сөздің дыбыс құрамын дәл бейнелегенде ғана дұрыс жазу болады. Ендеше қазақ жазуы қазақ сөзінің үндесім әуезін дұрыс көрсетіп отыру керек, басқаша айтқанда, жуан айтылатын сөздің таңбасы мен жіңішке айтылатын сөздің таңбасының арасында көрнекі айырма болсын.

2. Жазу қазақ сөзінің буын құрамын бұзбау керек. Мысалы, ***ми-ы, су-ы [мый-ы], [сұу-ы]***болмау керек, ***мы-йы, сұ-уы*** болу керек. Қазақ тілінде сөз ішінде буын дауыстыдан басталмайды. Оқулықтардағы осындай кереғарлықтардың қиындықтарын айтып, балаларға түсіндіре алмай, олардың алдында «бірде олай болады, бірде бұлай болады» деп екі сөйлеп, әлек болып жатқан ұстаз мұғалімдердің пікіріне де құлақ асу керек қой.

3. Жазу қазақ сөзінің тасымал ережесіне сай болу керек. Тағы да ***ми-ы, су-ы*** болып тасымалданбау керек, ***мы-йы, сұ-уы*** болып тасымалдану керек.

4. Жазу қазақ тілінің морфем (түбір мен қосымшалар) желісін үзбеу керек. Мысалы, қазақ тілінің «егер түбір дауыссызға бітсе, қосымша дауыстыдан басталады» деген іргелі заңдылығы бар. Сонда ***мың – мы-ңы, сұр – сұ-ры*** тәрізді ***мый – мы-йы, сұу – сұ-уы*** болып, тіл жүйесі бұзылмайды.

5.Қазақ тіліндегі дифтонгтардың фонологиялық мәртебесін (статусын) анықтау бүгінгі қазақ фонетика ғылымының өзекті мәселесіне айналып отыр. Өйткені бұл мәселенің теориялық және практикалық маңызы қатар өзекті (актуал) болып отыр. Мәселенің теориялық шешімі қазақ тілі дыбыстарының құрамын дәл анықтап, «қазақ тілінде қанша дауысты дыбыс бар?» деген сұраққа нақты жауап тауып беретін болады. Ал оның практикалық шешімі қазіргі қазақ жазуындағы қайшылықтардан арылуға мүмкіндік береді.

Жазуда ***и, у*** әріптерімен таңбаланып жүрген дыбыстарға, бір жағынан, **«дауыстыдан кейін дауыссыз, дауыссыздан кейін дауысты болады»** деген мектеп және жоғары оқу орындарының оқулықтарындағы жаңсақ анықтамамен қатар, екінші жағынан, олар **«дауысты мен дауыссыздың тіркесін белгілейтін таңба»** деген академиялық грамматикадағы ақиқат анықтама беріледі. Біз бұл жерде қазақ тілі оқулықтары мен ғылыми грамматикаларында кездесіп жүрген бір-біріне қарама-қарсы пікірлердің негізгілерін ғана келтіріп отырмыз.

Өзара артикуляциялық үйлесімі күшті, естілімі ұқсас болып келетін қысаң дауыстылар мен үнді дауыстылар тіркесінің құрамды бөліктерін ажыратып алу оңай болмайды. Оның үстіне орыс тілінің ***ы, и, у*** дауыстыларының естілімі мен қазақ тілінің ***ый, ій, ұу, үу*** дыбыс тіркестерінің естілімі өзара ұқсас болуы да өз әсерін тигізді. Кезінде қазақтың төл сөздерінің құрамында ***и, у*** таңбаларын жазу туралы тоталитарлық ереже қабылдауға да, сөз жоқ, осы жай себеп болған. Олардың құрамды бөліктерінің жігін ажыратудың бірнеше тәсілі бар.

6.Солардың бірі – морфем талдау: егер өзге морфем жігі анық тіркестермен ұқсас келсе, яғни тіркесім сыңарларының арасынан морфем жігі өтетін болса, онда олар дербес-дербес дыбыстардың тіркесі болып саналады.

Ендеше, бірінші, морфем жігі анық тіркестермен салыстырып шығамыз. Төмендегі әртүрлі түбір морфем дыбыс құрамының әртүрлілігіне қарамай, бірыңғай тілдік бірліктер болып табылады (көпшілікке түсінікті болу үшін тіркес транскрипцияларын да кирилше беріп отырмыз):

ай [ай], әй [әй], ей [ей], ой [ой], өй [өй], ұй [ұй], үй [үй], **и [ый], и [ій]**

Алғашқы жеті сөздің (тіркестің) дыбыс құрамына ешкім күманданбайды, ал соңғы ***и*** арқылы жазылып отырған сөзге (тіркеске) келгенде күмән басталады. Шындығында, барлығы ды әртүрлі дауысты дыбыстар мен үнді, жуысыңқы, тіл ортасы (тілшік) ***й*** дауыссызының тіркесі болып келеді. Егер ***ай*** немесе ***үй*** тіркестерінің дыбыс құрамына күман туғызбайтын болсақ, онда солармен теңдес ***ый, ій*** тіркестерінің де дыбыс құрамына күмәнданбауымыз керек.

Тіптен, басқаны былай қойғанда, ***и*** таңбасы бір дауысты дыбыс болса, онда ***ми*** деген сөздің құрамында неге жуан айтылып, ***ти*** деген сөздің құрамында неге жіңішке айтылып, екі түрлі дыбысталады деп ойлану керек қой. Өйткені ***и*** таңбасы бірде құрамында жуан ***ы*** дауыстысы бар ***ый*** тіркесін, бірде құрамында жіңішке ***і*** дауыстысы бар ***ій*** тіркесін белгілейді. ***Тілі қазақы шыққан, естілімі қазақы қалыптасқан адам үшін оны аңғарудың еш қиыншылығы жоқ.***

Енді ***у*** таңбасымен келетін бірыңғай сөздер тобын көрейік (***у*** таңбасының мәні түсінікті болу үшін оны арнайы ***w*** таңбасымен белгілеп отырмыз, өйткені қазақ тілінде ерін-еріндік, жуысыңқы, үнді ***w*** дыбысы бар. Ескерте кететін нәрсе, орыс тілінде ондай дауыссыз дыбыс жоқ, сондықтан да олар қазақтың **Көкшетау** тәрізді сөзінің соңына, ***w*** дауыссызын айта алмайтын болғандықтан, өздерінің үйреншікті ***в*** дыбысымен алмастырып отырады):

ау [aw], әу [әw], оу [ow], өу [өw], **у [ұw], у [үw]**

Алғашқы төрт сөздің (қазақ тілінде ***ы, і, е*** дауыстылары **у[w]** дауыссызымен тіркеспейді) дыбыс құрамына ешкім күманданбайды, ал соңғы ***у*** арқылы жазылып отырған екі сөзге (тіркеске) келгенде тағы да күмән басталады. Шындығында, барлығы ды әртүрлі дауысты дыбыстар мен үнді, жуысыңқы, еріндік ***у*** дауыссызының тіркесі болып келеді. Егер ***ау*** немесе ***әу*** сөздерінің дыбыс құрамына күмән туғызбайтын болсақ, онда солармен теңдес ***ұу, үу*** сөздерінің (тіркестерінің) де дыбыс құрамына күмәнданбауымыз керек.

Тағы да, басқаны былай қойғанда, ***у*** таңбасы бір дауысты дыбыс болса, онда ***алу*** деген сөздің құрамында неге жуан айтылып, ***елу*** деген сөздің құрамында неге жіңішке айтылып, екі түрлі дыбысталады деп ойлану керек болар. Өйткені ***у*** таңбасы бірде құрамында жуан ***ұ*** дауыстысы бар ***ұу*** тіркесін, бірде құрамында жіңішке ***ү*** дауыстысы бар ***үу*** тіркесін белгілейді. Мұның да ***тілі қазақы шыққан, естілімі қазақы қалыптасқан адам үшін аңғарудың еш қиыншылығы болмайды.***

Екінші, түбір морфема мен қосымша морфеманың жігіне қарайық.

май [май], ма-йы [ма-йы]
ми [мый], ми-ы [мы-йы]

тау [тау], та-уы [та-уы]
ту [тұу], ту-ы [тұ-уы]

7.Қазақ тіліндегі қосымшалар талғампаз келеді. Соның бірі түбір морфеманың дауысты не дауыссызға бітуіне байланысты. Егер ***май, тау*** сөздерінің дауыссыз дыбысқа бітуіне байланысты қосымша морфема дауысты дыбыстан басталып тұрса, онда дауысты дыбысқа бітіп тұр (?) деген ***ми, ту*** сөздеріне неге дауысты морфема жалғанып тұр? Ендеше бұл сөздер де дауыссыз дыбысқа бітіп тұр деген сөз. Олай болса, транскрипциядан көрініп тұрғандай, дыбыс тіркесінің екінші сыңары ***й, у*** дауыссыз дыбыстары болғаны. Бұған қоса, мысалға алынып отырған сөздердің бәрінің буын және морфем құрамы біркелкі екенін де ескеру керек.

Ендеше ***и, у*** таңбаларының дыбыс құрамын ***ый, ій, ұу, үу*** дыбыстарының тіркесі деп қарау керек. Өйткені қазақ жазуының құрамына енген кірме әріптердің дыбыстық мәнін анықтау аса күрделі фонетикалық мәселеге айналып отыр. Себебі қазақ тілінің оқулықтары

мен әдістемелік құралдарында кірме әріптер бірде дауысты, бірде дауыссыз, ал енді бірде дыбыс тіркесі деп түсіндіріліп келеді. Соның салдарынан, өкінішке орай, кейбір авторлар оларды қазақ тілінің жаңа латын әліпби құрамына да енгізуге тырысып жатыр. Жаңа фонема болу үшін ол қазақ сөзінің **үндесім әуезін** бұзбайтын (қазақтың жуан үндесім әуезді **мый** сөзін жіңішке әузбен **ми** деп айтқызбайтын), **буын құрамын** өзгертпейтін (**ми-ы** деп тасымалдатпайтын), **морфем жігін** үзбейтін (**мый+ы** деп түбір мен қосымша тіркесіміне нұқсан келтірмейтін) болу керек.

Ал, қазақтың төл сөздерінің жазылымына еніп кетіп, дауысты дыбыс болып жүрген ***я, ю*** таңбаларының тек дыбыс құрамын ғана ашып кетеміз. Дауысты ***я*** деп жүргеніміз ***йа, йә*** тіркестерінің таңбасы болса, дауысты ***ю*** деп жүргеніміз ***йұу, йүу*** тіркестерінің таңбасы болып табылады. Мысалы, **жая [жайа], сая [сайа], әлия [әлійә], дүрия [дүрүйә], аю [айұу], ою [ойұу], үю [үйүу], түю [түйүу].** Бұлардың да ақиқат дыбыс құрамына жоғарыдағы әдістермен көз жеткізуге болады.

Жаңа латын әліпби мен оның емле-ережелері қазақ сөзінің айтылым әуезіне, буын құрамына, морфем жігіне, тасымал ретіне нұқсан келтірмейтіндей болу керек. Мұндай келелі шара мектеп «Әліппесі» мен оқулықтарынан бастау алмаса, «әліпби мен емле-ереже сырқаты» бүгінгіден де асқынып кетері сөзсіз.

КАРИМОВ Б.Р.

Международный институт языка ортатюрк, Ташкент, Узбекистан

ПЕРСПЕКТИВЫ ТЮРКСКОЙ ЦИВИЛИЗАЦИИ И ПУТЬ ОБРЕТЕНИЯ ТЮРКСКИМИ ЯЗЫКАМИ, В ЧАСТНОСТИ КАЗАХСКИМ ЯЗЫКОМ, МИРОВОГО СТАТУСА ПОСРЕДСТВОМ ИСПОЛЬЗОВАНИЯ МЕТОДОВ КОМПЬЮТЕРНОЙ И МАТЕМАТИЧЕСКОЙ ЛИНГВИСТИКИ

Тюркская общность народов внесла и вносит оригинальный и значимый вклад в мировую историю и мировую цивилизацию, наследие Тюркской цивилизации имеет мировое общецивилизационное значение. Тюркская цивилизация способна самоорганизоваться в форме локальной цивилизации. Тюркская цивилизация является одной из древнейших цивилизаций мира. Истоки её уходят к 5-6 тысячелетию до нашей эры, когда Древний Шумер как первое развитое государство, образовавшееся на территории расселения протоалтайских и прототюркских народов (территории нынешних государств Турция, Иран, Азербайджан, Афганистан, Туркменистан, Узбекистан, Таджикистан, Казахстан, Кыргызстан, СУАР в КНР и др.) в числе первых в мире создал великие достижения человеческой цивилизации – развитые формы государства, этнической общности, языка, письменности, мифологии, религии (Тенгрианство), градостроительства и материальной культуры. После завоевания Древнего Шумера семитскими племенами и включения его территории в Древний Аккад, и затем в Древний Вавилон, новые центры развития Тюркской цивилизации возникают на оставшихся территориях расселения протоалтайских, прототюркских и древнетюркских народов. После завоевательных нашествий на эти территории скотоводческо-земледельческих полукочевых племен древних арийцев в XII-YII веках до нашей эры на значительной части этих территорий образуются государства с арийской государственной элитой, но значительную часть населения в них продолжают составлять потомки прототюрков, древние тюрки, а также смешанное арийско-древнетюркское население [1].

В древнекитайских письменных источниках этноним «тюрк» встречается уже 4700 лет до наших дней. Эти же источники свидетельствуют также и о том, что уже в те времена на

территориях от Каспийского моря до территории Маньчжурии проживали древнетюркские племена и народности [2]. Древние археологические памятники культуры на этих территориях древнего Турана свидетельствуют об этом. Тюркская цивилизация, находясь на пути между ведущими государствами, цивилизациями, культурами Востока и Запада, Севера и Юга, выступала как цивилизация, оказывавшая значительное влияние в системе цивилизаций мира. Важную роль она играла на Великом Шелковом пути. Тюркский язык был одним из ведущих международных, мировых языков на этом пути. Тюркская цивилизация создала целый ряд государств, имевших большое влияние на развитие мировой истории, культуры, науки, искусства и мировоззрения. В современную эпоху нового возрождения, Ренессанса Тюркской цивилизации она имеет возможности интенсивно развить свою культуру, науку и язык, посредством использования достижений современной мировой цивилизации. Гениальные и выдающиеся деятели науки, искусства и культуры Тюркской цивилизации внесли огромный вклад в развитие мировой науки, искусства и культуры. Наука, культура и язык Тюркской цивилизации внесли огромный вклад в развитие мировой науки, духовной культуры и языковой системы мировой цивилизации. Тюркская цивилизация в течение тысячелетий создала цепь последовательно возникавших развитых тюркских языков, фольклора, литературы и письменностей. Тюркский язык по длительности периода пребывания в статусе международного, мирового языка занимает первое место среди всех языков, когда-либо имевших такой статус. Тюркская цивилизация оказала и оказывает большое влияние на развитие письменностей цивилизаций Востока и Запада.

В процессе глобализации идет формирование многополюсного мира, что связано с возникновением новых полюсов, центров геополитической силы. Этими центрами выступают чаще всего локальные цивилизации. Тюркской общности народов целесообразно сформировать у своих членов тюркское локально-цивилизационное мировоззрение.

Фундаментальной идеей локального цивилизационного мировоззрения является идея цивилизационизма, идея цивилизационной солидарности, единства исторических корней и исторической судьбы народов, относящихся к данной локальной цивилизации, необходимости их сотрудничества для решения стоящих перед ними общих проблем.

Для перспектив развития Тюркской цивилизации в современной мировой информационной цивилизации большое значение имеет идеология тюркизма, являющаяся идеологией имеющей целью формирование Тюркской локальной цивилизации. Кратко резюмируя ряд исследований в данном направлении можно утверждать, что следующие концепции и положения соответствуют идеологии тюркизма, интересам Тюркской цивилизации в целом и каждого из тюркских народов в отдельности:

По цивилизационным вопросам [3; 4; 5; 6]: 1) Поддержка концепции этнолингвопанизма [3], концепции тюркизма и возрождения Тюркской цивилизации, разъяснение и поддержка концепций демократического тюркизма и ортатюркизма; 2) Позиция против концепции массовой культуры, против ликвидации национальных культур, за равноправный диалог, полилог локальных цивилизаций в рамках мировой цивилизации; 3) Поддержка концепции формирования среднеалтайского языка, общеалтайского мировоззрения и единого общеалтайского информационного пространства [7].

По политическим проблемам: 1) Создание Тюркской Конфедерации государств (ТКГ). Конфедерация должна обеспечивать сохранение суверенитета отдельных тюркских государств и их членство в ООН. В Тюркской Конфедерации допускается на основе консенсуса всех её членов участие отдельных стран ещё и в других межгосударственных объединениях; 2) Соблюдение устава ООН, поддержка её международной деятельности; 3)Поддержка концепции международно-правовой защиты прав человека; 4)Поддержка геополитической концепции многополярного мира. Одним из полюсов при этом может стать Тюркская Конфедерация; 5) Расширение Шанхайской Организации Сотрудничества (ШОС), и превращение её, в Организацию по безопасности и сотрудничеству в Азии (ОБСА), в аналог Организации по безопасности и сотрудничеству в Европе (ОБСЕ). При этом надо пользоваться тем, что тюркских государств в ШОС много – Узбекистан, Казахстан,

Кыргызстан, и добиться расширения за счет добровольного включения всех других азиатских стран. Организация «Совещание по взаимопониманию и мерам доверия в Азии», созданная по инициативе Казахстана, может слиться с ОБСА.

По военным вопросам: 1) Создание в рамках Тюркской конфедерации государств военного оборонительного союза на базе имеющейся системы межгосударственных договоров с сохранением суверенитета каждого тюркского государства; 2) Поддержка договора о нераспространении ядерного оружия и договора о безъядерной зоне в Центральной Азии с адекватным совместным своевременным реагированием ТКГ на происходящие в мире изменения в сфере ядерных угроз и ядерной безопасности.

По национальному вопросу [3; 4; 5; 6]**:** 1) Поддержка ойкуменической концепции нации, этносистского гуманизма и гуманистического этносизма и соответствующей концепции межнациональных, международных отношений. Опровержение крайних форм конструктивизма, инструментализма и примордиализма; 2) Поддержка концепции равноправия наций, борьба против гегемонизма, шовинизма одних наций против других; 3) Поддержка концепции межнациональной толерантности; 4) Поддержка концепции трехуровневого равноправия наций и их языков; 5) Поддержка принципа права наций на самоопределение, концепции адекватного понимания права наций на сепарацию, отделение и образование своего независимого государства признаваемого ООН. При этом использовать принцип компактификации и стягивания этнических территорий наций внутрь образующихся своих отдельных национальных государств [8; 9].

По лингвистическим вопросам [10; 11; 12; 13; 14; 15]**:** 1) Создание среднетюркского языка ортатюрк (анатюрк [14]) методом усреднения норм тюркских языков, признание его в качестве языка межтюркского межнационального общения, языка информации имеющей общетюркское и мировое значение и достижение признания его в качестве одного из языков ООН. Это обеспечивает равноправие всех тюркских языков между собой, каждый из тюркских народов имеет право использовать свой тюркский язык как государственный язык в своем национальном государстве и развивать его в меру своих возможностей. Механизм усреднения при создании норм языка «ортатюрк» дает близость к языку предков современных тюркских народов и внесение в язык ортатюрк основной части своего языкового наследия каждой из тюркских наций; 2) Создание системы усредненных языков для родственных языков; 3) Создание всемирного общечеловеческого языка на основе всех языков мира посредством системы усредненных языков; 3)Поддержка концепции государственного статуса национальных языков; 4) Поддержка концепции «понятных» языков. Близкородственные языки могут официально быть признанными в качестве «понятных» языков на территориях других государств; 5) Создание координированной системы алфавитов национальных тюркских языков и языка ортатюрк; 6) Создание координированной системы терминов тюркских языков [15]; 7) Поддержка концепции межъязыковой толерантности; 8) Поддержка концепции культурного разнообразия и защиты национальных языков; 9) Целесообразно каждый год 21 февраля отмечать в Тюркском мире «День родного языка», солидаризуясь в этом вопросе с решениями ЮНЕСКО.

По экономическим проблемам: 1) Создание Тюркского экономического пространства, экономического союза в рамках Тюркской конфедерации государств; 2) Поддержка концепции Всемирных денег в системе ООН. Национальная валюта отдельной страны и даже группы стран не способна обеспечить устойчивую реализацию функции Всемирных денег; 3) Поддержка Концепции человеческого развития и Концепции Целей развития тысячелетия.

По проблемам культуры: 1) Поддержка Концепции культурного разнообразия; 2) Создание общетюркской культуры путем усреднения и на базе языка ортатюрк для развития культуры, как отдельной тюркской нации, так и Тюркской цивилизации в целом; 3) Поддержка деятельности ЮНЕСКО по защите языков от гибели; 4) Поддержка концепции межкультурного диалога, полилога и толерантности [17]; 5) Поддержка концепции глобальной экокультуры, планетарной этики [18].

По проблемам религии: 1) Поддержка концепции свободы совести и вероисповедания как личной свободы отдельного человека. Отказ от отождествления национальной и религиозной идентичности, признание их взаимной независимости [19]; 2) Поддержка концепции светского государства в мире и в каждом из тюркских государств; 3) Поддержка концепции межконфессиональной толерантности; 4) Поддержка права тенгрианства как древнейшей религии протоалтайцев, прототюрков и древних тюрков [1] на своё добровольное возрождение, без принуждения индивидов.

По проблемам истории [20]**:** 1) Выработка координированной концепции всемирной истории, истории Тюркского мира, Тюркской цивилизации, истории отношений между тюркскими и другими народами; 2) Выработка согласованной концепции этногенеза и культурно-исторического наследия тюркских народов; 3) Развитие исследований о роли Тюркской цивилизации в древнем глобальном взаимодействии цивилизаций через Великий шелковый путь и в процессе формирования великих евразийских империй; 4) Углубление исследований в сфере шумерологии в рамках концепции исторической взаимосвязи Тюркской цивилизации с Шумерской, тюркского языка с шумерским, тюркского тенгрианства с шумерским.

По проблемам науки [21]: 1) Создание Фонда научных исследований ТКГ; 2) Создание Ассоциации Академий наук тюркских государств; 2)Создание в каждом тюркском государстве Тюркской Академии (ТА) с отделами: тюркологии; языка и литературы; истории; этнологии; философии; политики; права; экономики; демографии; экологии; социологии; информационного развития; культурологии; религиоведения; психологии; педагогики и образования; искусствоведения; спорта; здравоохранения и медицины; энергетики и природных ресурсов; водных, сельскохозяйственных и продовольственных проблем; востоковедения; глобалистики и футурологии. Целесообразно, чтобы эти ТА входили в систему международной организации «Международная Тюркская Академия» (МТА). Тюркская Академия, созданная по инициативе Президента Республики Казахстан Н.А.Назарбаева в столице Казахстана Астане, могла бы выступить началом этого великого объединения интеллектуальных сил Тюркской цивилизации. Для равноправия тюркских государств целесообразно, чтобы каждый год председательство в МТА переходило от ТА одного тюркского государства к ТА другого тюркского государства.

В целом система этих концепций даёт основы тюркского мировоззрения, при сохранении постмодернистского мировоззренческого плюрализма во многих других аспектах мировоззрения (онтология, гносеология, методология, этика, эстетика и др.) [22].

В процессе глобализации растет роль языковых реформ нацеленных на самосохранение наций, локальных цивилизаций и Человечества, на рост их интеллектуального и духовного потенциала. Языковый консерватизм при этом выступает как фактор тормозящий ход языковых реформ. Укоренился стереотип о том, что языки должны развиваться «естественным», «естественно-историческим», стихийным образом, что искусственные языки не должны широко использоваться в социальной жизни. Этот стереотип своим философским основанием имеет постулат о том, что языки являются объективной естественной данностью, также как и природная (неживая и живая) данность, объективная реальность. Поэтому, мол, субъективное воздействие на эту реальность путем создания искусственных языков, функционирующих на равных основаниях с естественными языками, неуместно в принципе. На самом деле язык создан обществом, а не непосредственно природой. Поэтому развитие языка должно идти и идет путем воздействия человека, в ходе развития деятельности социальных групп именно как средство реализации такой совместной деятельности. Индивидуальное субъективное воздействие на развитие языка не столь броско вследствие того что язык есть средство деятельности социальной общности, а не только индивида. Поэтому язык выступает для индивида как определенная данность, реальность в значительной мере не зависящая от его сознания и воли. Но социальная общность как субъект истории, создавший данный язык в ходе своей жизнедеятельности, реально преобразует и имеет право преобразовывать в дальнейшем, развивать свой язык.

Естественная природа, неживая и живая природа развивается по естественным, природным законам. В социуме происходит переход к социальным законам и закономерностям и «закон джунглей», утверждающий победу сильного над слабым, в социуме заменяется на закон социального партнерства, социального сотрудничества и толерантного сосуществования по принципу кантовского категорического императива всех субъектов социальной жизнедеятельности вне зависимости от их силы или слабости в рамках социальной общности. От царящей в природе смертельно опасной борьбы за жизнь, за существование в социуме должен произойти переход к ноосфере, к сфере подчиненной разуму, к рациональной регуляции социальных процессов на основе принципов гуманизма, этносизма и этнолингвопанизма [4; 5; 6]. В этой связи в языковой сфере необходимо отказаться от стереотипа языкового консерватизма о неотвратимости естественно-исторического процесса развития языков, так как этот процесс противоречит гуманизму и этносизму, приводит к интенсивной гибели языков народов мира. Но ведь эти языки являются бесценным культурным наследием Человечества, которое необходимо беречь. Поэтому необходимо перейти к сознательной регуляции процессов языкового развития всех социальных общностей. Лингвистические инновационные концепции формируются в форме гипотез и постепенно превращаются в теории, которые могут быть реализованы в практике языковой жизни, если они ей соответствуют. Этим концепциям приходится преодолевать языковый консерватизм и его стереотипы, укоренившиеся в сознании личностей и социальных групп.

В самосохранение и развитие Тюркской цивилизации вносят вклад все тюркские народы. В данной статье будет рассмотрен вклад казахского народа и перспективы развития казахского языка в системе тюркских языков, Тюркской и мировой цивилизаций.

Казахский народ внёс и продолжает вносить огромный вклад в восстановление исторической памяти Тюркской цивилизации, в развитие интеграционных процессов в её системе, в развитие взаимосвязей Тюркской цивилизации с другими цивилизациями в системе евразийских народов и мировой цивилизации. Казахский народ имел широкую систему контактов и взаимодействий со многими урало-алтайскими, индоевропейскими, китайскими, тюркскими, славянскими народами. Одновременно казахский народ достаточно устойчиво сохранял свои внутренние духовные ценности, свою культуру. Этому способствовало доминирование номадической компоненты в структуре культуры казахского народа, что позволяло подвергаться меньшему изменяющему влиянию внешних инокультурных социальных воздействий. Оседлое население чаще всего оказывается не способным уйти с территории завоеванной агрессором и подвергается сильному всестороннему социальному воздействию с его стороны, в ходе которого происходят большие изменения в его языке, культуре и мировоззрении. В отличие от оседлого населения кочевое население имеет больше возможностей уйти от удара агрессора, не стать завоеванным народом и в большей мере сохранить свой язык, свою культуру и мировоззрение. Казахский народ в большей мере, чем многие другие тюркские народы, сохранил тюркское ядро в системе своего языка. Это связано с тем, что казахский народ является крупнейшим тюркским народом с историческим доминированием номадической компоненты в своем составе. Этому способствовало и то, что казахский народ территориально находится в центре Тюркской цивилизации, в основном окружен тюркскими народами и тем самым в определенной мере защищен от нетюркских внешних воздействий. Казахский язык испытал воздействие многих вариантов тюркского языка, и кипчакских, и огузских, и карлукских, и алтайско-сибирских. Впитав в себя богатства этих вариантов тюркских языков казахский язык в лексическом аспекте исторически оказался близким к большинству тюркских языков, и в этом смысле среднетюркским. Это проявилось и в том, что язык великого Абая (1845-1904), являвшийся региональным вариантом литературного тюркского языка «тюрки», оказался наиболее близким к среднетюркскому языку ортатюрк, нормативная система которого выявляется посредством методов математической и компьютерной лингвистики в ряде современных исследований [10; 11; 12; 13; 14; 15].

Идеи Президента Республики Казахстан Н.А.Назарбаева о единстве тюркских народов дали мощный духовно-интеллектуальный импульс тюркологическим исследованиям современных ученых и будут способствовать дальнейшим исследованиям последующих поколений ученых в сфере истории, этнологии, лингвистики, графемики, культурологии, политологии, социологии, экологии, геополитики и философии Тюркской цивилизации. Республика Казахстан внесла большой вклад в формирование и развитие Тюркского совета, Межпарламентской Ассамблеи тюркских государств, Тюркской Академии и других интеграционных структур Тюркского мира. В Казахстане создан Евразийский национальный университет им. Л.Н.Гумилева, Международный Центр алтаистики и тюркологии, в которых ведется много исследований, посвященных истории древних алтайцев, древних тюрков, современному состоянию и перспективам развития Тюркской цивилизации.

После достижения независимости Республикой Казахстан и казахской нацией делаются большие усилия для подъема внутригосударственного и международного социального статуса казахского языка. Принята и проводится в жизнь государственная программа по поддержке развития казахского языка как государственного языка страны. Созданы центры по обучению казахскому языку и традициям. Однако в этом направлении есть еще много трудностей и нерешенных проблем. Значительная часть делопроизводства в стране не ведется на государственном языке. Не достаточным для современного уровня развития мировой цивилизации является уровень развития терминологической системы казахского языка и объем информации, имеющейся на казахском языке. В межнациональной и в международной системе коммуникации казахский язык используется недостаточно активно. В этих функциях в основном используются русский и английский языки. В комплексе эти факторы понижают социальный статус и имидж казахского языка, создавая о нем ошибочное мнение как о сугубо национальном языке неконкурентоспособном с «мировыми языками». Для прорывного ускоренного повышения социального статуса казахского языка требуются оригинальные нестандартные решения, так как стандартные решения не дают ожидаемого эффекта в требуемые сроки, несмотря на очень большие выделяемые средства.

Цель данного проекта: Обеспечить научную основу для прорывного инновационного модернизационного повышения внутригосударственного и международного социального статуса казахского языка посредством создания близкородственного среднетюркского языка ортатюрк, имеющего потенциальные возможности для превращения в международный язык признаваемый ООН [10; 11; 12; 13; 14; 15]. Для достижения этой цели требуется решение следующих задач: 1) создать среднетюркский язык ортатюрк; 2) создать координированную и унифицированную систему алфавитов национальных тюркских языков и языка ортатюрк; 3) создать координированную терминологическую систему тюркских языков и языка ортатюрк [15]; 4) развить информационные ресурсы на языке ортатюрк и национальных тюркских языках, в том числе на казахском языке [23; 24].

Среднетюркский язык ортатюрк в лексическом аспекте будет близок к языку Абая (1845-1904), одного из последних классиков Тюркской цивилизации, писавших на региональном варианте литературного языка «тюрки», который был среднетюркским и общетюркским языком той эпохи. Этот великий язык Тюркской цивилизации в 1924 году был юридически превращен в мертвый язык не имеющий социального функционирования в значимых сферах жизни общества и государства. На всех активных сторонников сохранения единства тюркского языка, тюркского народа и Тюркской цивилизации большевики навесили ярлык «пантюркист», объявили их «врагами народа» и жесточайшим бесчеловечным образом преследовали и уничтожали [25]. Эта шовинистическая, отрицающая право наций на самоопределение акция была осуществлена в ходе так называемого «национально-государственного размежевания» произведенного посредством массового террора под руководством Сталина без выявления свободного волеизъявления народа Туркестана и тюркских народов. Но те, которые желали уничтожить и думали, что уничтожили языковое единство Тюркской цивилизации, ошиблись. Создание языка ортатюрк выступает как оживление великого тюркского языка «тюрки», выведение его из состояния «клинической

смерти» с помощью современных методов математической и компьютерной лингвистики, современных средств информационной цивилизации. Это фундаментальное лингвистическое единство Тюркской цивилизации независимые тюркские народы смогут восстановить.

Реализация предлагаемого нами проекта создания языка ортатюрк и совершенствования системы функционирования тюркских языков и письменностей выступила бы как фактор прорывного, инновационного, модернизационного социального развития государств и наций, входящих в Тюркскую, Центрально-азиатскую цивилизацию и мировой цивилизации.

Для социально-экономического развития Казахстана выполнение данного проекта имело бы большое значение, так как позволило бы оперативно, без очень больших затрат собственных ресурсов, коллективными усилиями всех тюркоязычных народов и структур ООН, получать основную мировую информацию и передавать миру свою информацию на близкородственном мировом языке ортатюрк. Это подняло бы социальный статус и имидж казахского языка почти до уровня международных, мировых языков. В свою очередь, это привело бы к повышению и его внутригосударственного социального статуса.

Возможны следующие негативные последствия в случае отказа от данной программы. Будет упущена стратегически важная историческая и геополитическая возможность самосохранения и успешного саморазвития Казахстана, казахской нации и казахского языка на основе опоры на сотрудничество и взаимопомощь родственных тюркских народов. Казахский язык, может быть, будет оттеснен из основных сфер социальной жизни такими международными, мировыми языками как русский, английский и китайский языки или подвергнется их очень сильному деформирующему воздействию. В результате этого языковая, национальная и государственная идентичность населения Казахстана может испытать большие колебания и стать нестабильной в процессе глобализации и мощного воздействия нетюркских геополитических акторов. Это противоречит национальным интересам и национальной безопасности Казахстана и казахской нации.

Существуют следующие основные варианты решения проблемы языка межтюркского межнационального общения: 1) Признание английского языка как лидирующего мирового языка, который знают многие представители тюркских народов в качестве языка межтюркского межнационального общения; 2) Признание русского языка как одного из мировых языков, который знают многие тюркские народы в качестве языка межтюркского межнационального общения; 3) Признание одного из национальных тюркских языков в качестве языка межтюркского межнационального общения; 4) Возрождение языка «тюрки» в качестве языка межтюркского межнационального общения; 5) Создание среднетюркского языка «ортатюрк» методом усреднения тюркских языков и признание его в качестве языка межтюркского межнационального общения; 6) Синтетический вариант по отношению к вариантам 4 и 5, в этом варианте можно, приняв за основу фонетику и грамматику «тюрки», в аспекте лексики использовать усреднение лексики современных тюркских языков.

В целом представляется целесообразным реализовать пятый вариант, то есть создание среднетюркского языка «ортатюрк» методом усреднения тюркских языков, признание его в качестве языка межтюркского межнационального общения, языка информации имеющей общетюркское и мировое значение и достижение признания его в качестве одного из языков ООН [10; 11; 12; 13; 14; 15]. Этот вариант обеспечивает равноправие и равенство всех тюркских языков между собой, не предоставляет привилегий и не создает ущемления прав и достоинства каждого из тюркских народов. Каждый из тюркских народов имеет право использовать свой тюркский язык как государственный язык в своем национальном государстве и развивать его в меру своих возможностей. А при решении очень крупных задач, которые не под силу отдельному тюркскому народу, таких как освоение общемировой информации и внесение на этой основе большого вклада в мировую цивилизацию, каждая тюркская нация будет иметь возможность добровольно и в меру своих интересов и желаний дополнительно использовать язык «ортатюрк» как наиболее близкий родственный язык из признанных ООН мировых языков. Механизм усреднения при создании норм языка «ортатюрк» дает также и близость к языку предков современных тюркских народов и

внесение в язык ортатюрк основной части своего языкового наследия каждой из тюркских наций. Создание усредненного языка ортатюрк, выступающего в функциях добровольно признанного нейтрального языка межнационального общения, накопления информации общетюркской и общечеловеческой значимости способствовало бы развитию взаимопонимания и равноправного сотрудничества между тюркскими народами, социально-экономическому, политико-правовому, информационно-коммуникативному, профессионально-квалификационному, духовно-интеллектуальному развитию личностей и социальных групп, как тюркского мира, так и всего человечества.

Ортатюрк способствовал бы сотрудничеству между тюркскими народами, информационно-коммуникативному, духовно-интеллектуальному развитию личностей и социальных групп, как Казахстана и Тюркского мира, так и всей Мировой цивилизации.

Тюркская группа языков входит в состав более крупного генетического объединения языков, названных алтайскими. Возникает возможность варианта языкового развития с созданием среднеалтайского языка. Меру целесообразности вариантов языкового развития следует исследовать отдельно. Расчет меры близости различных генеалогически родственных языков в рамках групп и подгрупп языков может быть произведен на основе разработанного в математической лингвистике метода определения количественной меры близости и удаленности родственных языков и диалектов [26]. Это позволит найти оптимальные пути усреднения. На этой основе можно выяснить, какой из вариантов языкового развития лучше: 1) произвести усреднение по всей группе генеалогически родственных языков; 2) произвести усреднение по отдельным подгруппам генеалогически родственной группы языков; 3) не производить усреднения.

Процессы развития языков, особенно проблемы изменения языковой идентичности, целесообразно оценить с точки зрения соотношения прав человека и прав наций и языковых групп личностей. В индивидуалистической концепции личности личность отрывается от сформировавшего её социума и рассматривается как независимое априорно самодостаточное исходное начало для оценки всех процессов происходящих в мире. Права социума, общности, коллектива, который сформировал данного человека, при этом учитываются в недостаточной мере. Целесообразно достичь большей гармонизированности в соотношении прав личности и прав социума. Для этого необходимо учесть социальную сущность человека и нации. В этой связи целесообразно использовать ойкуменическую теорию нации, концепции этносизма, этнолингвопанизма, усредненных языков и среднемирового языка [10; 11; 12; 13; 14; 15], которые в целом составляют лингвогеополитическую концепцию, ведущую к синтезу языков Востока и Запада.

Концепции этносизма и этнолингвопанизма основываются на ойкуменической концепции нации [4; 5; 6]. Они дают основу для конвергентного развития цивилизаций, наций и культур Востока и Запада, так как эти концепции показывают, что такие характеристики нации, как общность территории, экономики, языка, культуры, социально-психологических черт являются акциденциями, но не атрибутами. В контексте ойкуменической теории нации рассмотрим проблемы глобального развития и пути сохранения национальных культур, языков, развития языковой и культурной конвергенции Востока и Запада. В этих вопросах целесообразно достичь оптимального сочетания общечеловеческих и национальных интересов. Для обеспечения единства Человечества и сохранения его многообразия целесообразно создание системы усредненных языков для групп генеалогически родственных языков [10; 11; 12; 13], а в дальнейшем создание среднемирового языка посредством усреднения в многообразии усредненных языков и изолированных языков на основе ностратической (борейской) концепции, концепции языковых универсалий и статистических методов усреднения языковых феноменов [10; 11; 12; 13]. Создаваемый таким путем всемирный вспомогательный язык межкультурного, межнационального общения, накопления мировой информации и глобального обучения способствовал бы решению многих глобальных проблем мировой цивилизации и духовному взаимообогащению всех локальных цивилизаций и народов. Создание среднемирового языка

могло бы выступить как путь языковой и культурной конвергенции Востока и Запада в рамках системы единого Человечества [27].

В современную эпоху глобализации и формирования мировой информационной цивилизации важно решение проблем формирования единого информационного пространства для каждой из групп родственных по языку народов. Рассмотрим это на примере алтайских народов. Языковые барьеры, как обусловленные различием языков, так и различием их письменностей, являются препятствиями развитию данного единого информационного пространства. Иероглифы, используемые в японском языке, относящемся к алтайской семье языков, создают «иероглифический барьер», который также препятствует развитию единого информационного пространства алтайских народов. Для алтайской семьи в целом, включающей в себя тюркские, монгольские, тунгусо-маньчжурские, корейский и японский языки, проблему языкового барьера между алтайскими языками предлагается преодолеть посредством использования метода создания усредненных языков для соответствующих групп родственных языков, то есть путем создания среднеалтайского языка на основе создания среднетюркского, среднемонгольского языков и усредненного тунгусо-маньчжурского языка. Для создания среднеалтайского языка целесообразно усреднить следующие пять языков: среднетюркский, среднемонгольский, усредненный тунгусо-маньчжурский, японский и корейский языки. При этом предлагается использовать метод усреднения в меру его применимости, используя также достижения современной алтаистики и борейской, ностратической теории. При таком построении среднеалтайский язык не будет достаточно целостным. Поэтому для дополнения недостающих компонентов целесообразно использовать теорию языковых универсалий, статистические методы переработки баз данных. При создании усредненного языка для других семей и групп языков (романской, германской, индийской, дравидийской, индонезийско-малайзийской, славянской, семитской, иранской, уральской и др.) целесообразно применять метод аналогичный методу, примененному в отношении алтайской семьи языков [28].

Целесообразно также создать глобальную единую всемирную систему письменности [29], охватывающую как письменности на основе алфавитов, так и иероглифические и силлабарийные системы письменности [30]. Ныне иероглифический барьер, сильнее чем «Великая китайская стена» в древности, разделяет мировую цивилизацию и мировое информационное пространство. Для реализации западно-восточного синтеза культур необходимо решение этой глобальной проблемы развития письменности. Богатства человеческой культуры мирового значения, закодированные в сложнейших системах письменности связанных с иероглифическим принципом письма, целесообразно перекодировать, перейдя к оптимальной по затратам человеческих сил новой кодировке построенной на основе алфавитного принципа. Это было бы благом для всего Человечества, в том числе для самих народов Китая и Японии, являющихся неотъемлемой частью мировой цивилизации. Это позволит объединить интеллектуальные и материальные ресурсы Человечества для инновационного решения стоящих перед всеми нами сложнейших и опаснейших глобальных проблем. Самореализация, наполнение смыслом жизнедеятельности человека в этих сообществах приобрело бы характер, соответствующий требованиям современной информационной и инновационной эпохи.

В процессе формирования мировой информационной цивилизации для каждого тюркского языка целесообразно создание компьютерных программ, которые преобразуют тексты на одном алфавите в тексты на другом алфавите. Целесообразно создание компьютерных программ для перевода с одного тюркского языка на другой. При этом перевод на язык ортатюрк мог бы служить основным этапом для последующего перевода на другие тюркские языки [23]. Необходимо увеличить информационные ресурсы в Интернет на национальных тюркских языках и на языке ортатюрк [23; 24]. Осуществление этих предложений способствовало бы развитию Тюркской цивилизации в системе мировой информационной цивилизации.

В решении проблем межкультурной коммуникации близкородственных по языку наций целесообразно использование концепции понятных языков [16]. Понимать неродной родственный язык значительно легче, чем свободно говорить на этом языке, поэтому концепция понятных языков существенно облегчает общение наций на территориях, включающих в свой состав население с родственными языками. Юридически признанный статус понятного языка позволил бы этим этноязыковым группам активно участвовать в государственной, общественной и производственной жизни страны. Это помогает государствообразующей нации обеспечить государственный статус своего национального языка, так как расширяет сферу его применения на близкородственные по языку этнические группы и предотвращает их переход на неродственный официальный, региональный или международный языки. Ответ на обращение на понятном языке можно будет давать на государственном языке, или на другом официальном языке данного государства, или на языке, официально признанном понятным.

В течение столетий в значительной части территории Евразии было широко распространено знание двух языков: тюркского («тюрки») и иранского («фарси»). В XXI веке можно было бы содействовать добровольному изучению желающими двух усредненных языков – среднетюркского («ортатюрк») и среднеиранского («ирани»). Это способствовало бы мирному сосуществованию, взаимопониманию, сотрудничеству и межкультурному диалогу между представителями Тюркской и Иранской цивилизаций [31; 32; 33; 34].

Реализация этих проектов выступила как фактор инновационного модернизационного развития мировой цивилизации. Развитие системы информации, трансфер технологий, информационное обеспечение инновационной деятельности, защита права интеллектуальной собственности, международное сотрудничество в этих сферах обеспечивается в большей мере при преодолении языковых и иероглифических барьеров и формировании единого мирового информационного пространства, использующего алфавитный принцип и единую координированную и унифицированную систему алфавитов [29; 30; 35].

Предлагаемые преобразования соответствуют тенденциям развития и расширяют горизонты развития межкультурной коммуникации в процессе формирования мировой информационной цивилизации в XXI веке. Они направлены на решение проблем в данной сфере на основе общепризнанных в системе норм международного права принципов равноправия, суверенитета государств, обеспечения прав и свобод человека и коллективных прав социальных групп (национальных, языковых, этнических, расовых, конфессиональных, локально-цивилизационных и др.).

Создание среднетюркского языка ортатюрк, который в лексическом аспекте будет близок к языку Абая, приведет к возрождению каждого из тюркских языков и всей системы тюркских языков в качестве международных, мировых языков. Это послужит основанием для прорывного инновационного модернизационного повышения внутригосударственного и международного социального статуса каждого тюркского языка, в том числе и казахского языка, и всестороннему развитию каждой из тюркских национальных культур [36].

Литература

1. Каримов Б.Р. Тенгрианство и тюрки: история и современность // «Шаманизм как религия: генезис, реконструкция, традиции». Якутск: ЯГУ, 1992.

2. Ходжаев А. Из истории древних тюрков (сведения древнекитайских источников). Ташкент: Tafakkur, 2010.

3. Каримов Б.Р. Этнолингвопанизм как одна из форм идеологического обоснования взаимообогащения культур // Проблемы обоснования в контексте развития культуры. Уфа. 1991.

4. Каримов Б.Р. Миллат, инсон ва тил: тараққиёт муаммолари. Қарши: Насаф, 2003.

5. Karimov B.R. The oikumenic concept of the nation and development of languages. Ойкуменическая концепция нации и развитие языков. Qarshi, 2003.

6. Каримов Б.Р. Ойкуменическая концепция нации и развитие языков. Якутск, 2004.

7. Каримов Б.Р. Проблемы формирования единого информационного пространства алтайских народов // Актуальные проблемы комплексного исследования алтаистики и тюркологии. Материалы Международного конгресса. Кокшетау, 2009.

8. Каримов Б.Р. Формирование нации и проблема соотношения этнической территории нации и территории ее национального государства // Актуальные проблемы социально-гуманитарных наук. Ташкент, 2005.

9. Каримов Б.Р. Ойкуменическая концепция нации и проблема разумного урегулирования территориальных проблем // VIII Конгресс этнографов и антропологов России «Границы и культуры». Оренбург, 2009. С. 285-286.

10. Каримов Б.Р., Муталов Ш.Ш. Ўртатурк тили. Тошкент, 1992.

11. Karimov B.R., Mutalov Sh.Sh. Averaged languages: an attempt to solve the world language problem. Tashkent: Fan, 1993. (второе издание в 2008 г.).

12. Каримов Б.Р., Муталов Ш.Ш. Усредненные языки: попытка решения мировой языковой проблемы. Т.: Фан, 2008.

13. Karimov B., Mutalov Sh. Ortak Turkce // «Bilig» Dergisi. Sayi-3/Guz'96, S.190-199.

14. Каримов Б.Р. Лингвистические основы единства и сотрудничества тюркских народов // Султанмурат Е., Мухаметдинов Р., Каримов Б. Тюркский пояс стабильности. Алматы, 2008. С.38-52. (Второе издание: Казань, 2009. С. 35-48.)

15. Каримов Б.Р. Координирующая терминологическая система для родственных языков // Компьютерный фонд терминов тюркских языков. Туркистан-Шымкент, 1995.

16. Karimov B.R. Interrelation of concepts of understandable and averaged languages in the decision of national-language problems // Замонавий тилшунослик ва таржимашуносликнинг долзарб масалалари. Тошкент. 2012. С. 168-170.

17. Karimov B.R. Oikumenic theory of nation and problem of tolerance in the conception of ethnolinguopanism and ethnosism. Ойкуменическая теория нации и проблема толерантности в концепциях этносизма и этнолингвопанизма // Толерантность: идея и традиции. Материалы Международной научной конференции «Через толерантность к взаимопониманию и миру» (Якутск,12-15 июля 1994 г.). Якутск, 1995.

18. Аюпов Н.Г., Нысанбаев А.Н. О тюркской фальсафе // Тюркская философия: десять вопросов и ответов. Алматы, 2006. С.112-127.

19. Каримов Б.Р. Проблемы соотношения национальной и конфессиональной самоидентификации личности и гуманизм // Логос. Культура. Цивилизация. Якутск, 2003.

20. Каримов Б.Р. Пути укрепления международного сотрудничества в социогуманитарных исследованиях // Актуальные проблемы филологии и социально-гуманитарных наук. Шымкент, 2006. С.4-9.

21. Каримов Б.Р. Тюркская академия и Центр исследования Тюркского мира // Гуманитарная наука сегодня. II Международная конференция. Т.1. Караганда, 2010. С.49-51.

22. Каримов Б.Р. Национальная философия отдельного тюркского народа в контексте тюркской и мировой философии // Туркология, № 3(53), 2011. С.134-139.

23. Karimov B.R. Ortatürk dili ve kitle iletişim araçlarında Türkçenin kullanımını yoğunlaştırmak // I Uluslararası “Kitle iletişim araçlarında Türkçenin kullanımı” bilgi şöleni bildirileri. 16-17 Nisan 2009, Kırıkkale, 2011. S. 665-667.

24. Karimov B.R. Türk kültürünün cihan dil kültürü gelişmesindeki katkısı ve şimdiki zaman dil problemleri // 2 Uluslararası Türk kültürü kurultayı. Fethiye, 03-04 Aralık 2009, Fethiye, 2010.

25. Аншин Ф.Д., Алпатов В.М., Насилов Д.М. Репрессированная тюркология. Москва: «Восточная литература», 2002. – 296 с.

26. Каримов Б.Р., Муталов Ш.Ш. О количественной оценке синхронической близости родственных языков и диалектов // Тюркское языкознание. Материалы III-ей Международной конференции по тюркологии (10-12 сентября 1980 г.). Ташкент, Фан, 1985. С. 126-129. (Тезисы конференции были опубликованы в Ташкенте в изд-ве «Фан» 1980 году).

27. Каримов Б.Р. Глобальное обучение и пути языковой и культурной конвергенции Востока и Запада // VII Международная конференция «Образование личности и развитие межкультурной компетентности в новом тысячелетии». Хабаровск, 2010. С.34-35.

28. Каримов Б.Р. Проблемы формирования единого информационного пространства алтайских и уральских народов // Наука Удмуртии, № 5(43), июнь 2010. С.63-66.

29. Каримов Б.Р. Проблема создания единого унифицированного алфавита как глобальная проблема // Актуальные вопросы в области гуманитарных и социально-экономических наук. Вып.2. Ташкент, 2005, с.22-23.

30. Каримов Б.Р., Каримова У.Б. Проблемы развития письменностей языков в процессе глобализации. Ташкент: IFEAC, 2006. – 28 с.

31. Каримов Б.Р. Среднеиранский язык «ирани» и пути развития языка межиранского межнационального общения // Эроний тилларнинг лексикологияси. Тошкент, 2010. 80-88-б.

32. Каримов Б.Р., Муталов Ш.Ш. Метод создания усредненного языка «ирани» (на примере его лексического уровня) // Эроний тилларнинг лексикологияси. Т., 2010. 89-96-б.

33. Каримов Б.Р. Проблемы развития этноязыковых контактов в регионе "Великого шелкового пути" // Цивилизация. Мустакиллик. Инсон. Тошкент, 1996. С.128-130.

34. Karimov B.R. Bukhara in the system of ethnolinguistic communications of the great silk route: past, present and future // Scientific and cultural heritage of mankind – to the third millennium. Theses of reports of the international symposium dedicated to the 2500 anniversary of Bukhara and Khiva. Tashkent, 1997. P.87-89.

35. Каримов Б.Р. Стереотипы языкового консерватизма, концепции языка ортатюрк и единой системы письменности // Материалы Международной научной конференции «Актуальные вопросы фонетических наук: истоки и перспективы». Алматы, 2012. С.278-284.

36. Каримов Б.Р. Как превратить язык Абая в мировой язык? (Проблемы и пути развития тюркских языков и их письменности) // Наука. Философия. Религия. Алматы: КазНПУ им.Абая, 2008. С.22-26.

Ж.А. ЕСЕНБАЕВ [1, 2], М.Х. КАРАБАЛАЕВА [2], Ф.К. ШАМАЕВА[3]

[1] Nazarbayev University Research and Innovation System, Астана Казахстан
[2] Евразийский национальный университет им. Л.Н. Гумилева
[3] Кызылординский государственный университет им. Коркыт Ата, Кызылорда, Казахстан

К ВОПРОСУ ОБ УТОЧНЕНИИ ФОНЕТИЧЕСКОГО СТРОЯ КАЗАХСКОГО ЯЗЫКА ПОСРЕДСТВОМ АКУСТИЧЕСКОГО АНАЛИЗА ЗВУКОВ

В своем Послании народу Казахстана в декабре 2012 года Президент РК Нурсултан Назарбаев поручил профильным ведомствам разработать латинский алфавит для казахского языка и проработать поэтапный переход к нему [1]. Таким образом, на политическом уровне решение принято. Настало время сосредоточиться на главном вопросе: каким будет новый алфавит?

Система графем нового алфавита не должна быть трудоемкой и времяемкой по чтению и письму, загруженной правилами правописания. Также нежелательно, чтобы превышались нормы среднестатистической длины слова в тексте и допускались кардинальные различия в произношении и написании. При этом важно, чтобы буквы алфавита правильно отражали фонологические звуковые законы. Таким образом, при разработке нового алфавита для казахского языка следует ориентироваться на число основных фонем казахского языка с тем, чтобы избежать включения избыточных, нефункциональных символов и знаков в новый алфавит [2].

В данной работе на основе экспериментальных данных был проведен акустический анализ звуков казахского языка с целью уточнения его фонетического строя.

Традиционная классификация звуков казахского языка

Согласные звуки

К согласным звукам относят звуки, получаемые в результате образования препятствий на пути воздушного потока в процессе произношения речи. Обычно препятствиями могут быть следующие события:

- блокирование речевых органов;
- сильное или слабое ограничение речевых органов;
- перенаправление воздушного потока через носовые органы.

В казахском языке существуют 26 согласных звуков [3]: Б, В, Г, Ғ, Д, Ж, З, Й, К, Қ, Л, М, Н, Ң, П, Р, С, Т, У, Ф, Х, Һ, Ц, Ч, Ш, Щ. Здесь У является согласным, если используется после гласного звука («ауа», «әуен»). Данные звуки можно классифицировать по трем различным признакам:

1) по участию голоса: глухие, звонкие, сонорные.
2) по способу образования: смычные, щелевые, сонорные.
3) по месту образования: губные, губно-зубные, зубные, переднеязычные, среднеязычные, заднеязычные, увулярные.

В таблице 1 приведена классификация согласных звуков по указанным признакам.

Таблица 1. Классификация согласных звуков казахского языка

		Место образования						
Способ образования	Участие голоса	губные	губно-зубные	зубные	передне-язычные	средне-язычные	задне-язычные	увулярные
Смычные	глухие	П		Т, Ц	Ч	К	Қ	
	звонкие	Б		Д		Г		
Щелевые	глухие		Ф	С	Ш, Щ		Х	Һ
	звонкие		В	З	Ж		Ғ	
Сонорные		М, У			Н, Р, Л, Й		Ң	

Гласные звуки

К гласным звукам относят звуки, произносимые свободно без каких-либо препятствий со стороны речевых органов. В казахском языке выделяют 12 гласных звуков: А, Ә, Е, И, І, О, Ө, У, Ү, Ұ, Ы, Э. Звуки И и У являются дифтонгами: И = Ы+Й или І+Й; У = Ұ+У или Ү+У. Звук Э употребляется только в словах, заимствованных из русского языка.

Гласные звуки казахского языка классифицируют по трем признакам:

1) по положению языка: мягкие и твердые;
2) по положению челюсти: открытые и закрытые;
3) по положению губ: огубленные и неогубленные.

В таблице 2 приведена классификация гласных звуков по указанным признакам.

Таблица 2. Классификация гласных звуков казахского языка

	По положению губ			
	неогубленные		огубленные	
По положению челюсти	открытые	закрытые	открытые	закрытые
твердые	А	Ы (И)	О	Ұ (У)

мягкие	Ә, Е	І (И)	Ө	Ү (У)

Акустические характеристики согласных звуков казахского языка

В данной работе был проведен акустический анализ согласных звуков казахского языка во временно́й и частотной областях на основе аудио данных двух мужских и двух женских голосов носителей языка. В процессе анализа были использованы программные продукты Audacity и Praat. Согласные звуки казахского языка сгруппированы в соответствии с классификацией международного фонетического алфавита (IPA).

Смычные (взрывные) согласные

Во время произношения взрывных согласных воздушный поток полностью блокируется на мгновенье речевыми органами, формируя небольшое давление за образовавшимся препятствием. Уменьшение этого препятствия и дальнейшее движение артикуляторов к следующему звуку определяют два основных признака, сигнализирующих о месте артикуляции взрывных согласных: частота взрыва и переходы формант [4]. Кроме того, длительность самого препятствия также является важной характеристикой взрывных согласных, которая на спектрограмме отображается в виде интервала либо не содержащего существенной энергии (для глухих звуков), либо содержащего энергию в низкочастотном диапазоне (для звонких звуков). Таким образом, основными критериями при анализе данного класса звуков являются длительность паузы и взрыва, диапазон частот концентрации энергии взрыва, а также переходы второй (F2) и третьей (F3) формант смычных звуков.

Пауза, образующаяся при полном блокировании речевых органов во время произношении взрывных согласных, различает между собой глухие [П, К, Қ, Т] и звонкие [Б, Г, Д] смычные согласные. Глухие звуки не имеют энергии в данном интервале, средняя длина паузы составляет 40-56 мс. Звонкие звуки имеют квазипериодичный сигнал во временной области, образованный вибрацией голосовых связок, что отражается в виде низкочастотной полосы на спектрограмме. Длина паузы колеблется в диапазоне 25-50 мс. Длина взрыва в глухих звуках также больше, чем у звонких звуках и составляет, соответственно, 18-40 мс и 5-16 мс. Отдельно можно выделить заднеязычный звук [Қ], у которого интервал паузы достигает 100 мс, а взрыва – 60 мс. Более того, зачастую паузообразный интервал у данного звука может отсутствовать, если перед ним употребляется гласный звук. В таких случаях звук [Қ] звучит как щелевой [Х].

Губные звуки [П, Б] имеют низкое расположение частоты взрыва – около 500-1400 Гц, зубные звуки [Т, Д] – 2000-4000 Гц, а среднеязычные [К, Г] – 2000-3500 Гц. Наиболее высокую частоту взрыва имеет заднеязычный звук [Қ] – 6000-7500 Гц.

Переход формант от предыдущего гласного звука к взрывному звуку и от взрывного к последующему гласному также определяет место артикуляции согласного. Так, для губных звуков энергия взрыва расположена в низкочастотной области, а, следовательно, F2 и F3 опускаются от предстоящего гласного звука и поднимаются к последующему гласному. В зубных и заднеязычном звуках F2 и F3 поднимаются от предстоящего гласного звука и опускаются к последующему гласному, в силу высокого расположения частоты взрыва. В случае со среднеязычными звуками, F2 поднимается, а F3 опускается от предыдущего гласного, и наоборот, F2 опускается, а F3 поднимается к последующему гласному.

Смычные (аффрикаты)

Аффрикаты [Ц, Ч] представляют собой совмещенные пары звуков [ТС] и [ТЩ], соответственно. Данные звуки перешли из русского языка и в основном используются в заимствованных словах. Акустические характеристики аффрикат аналогичны смычным и щелевым. Отличием является лишь короткая длительность фрикативной части – около 47-90 мс.

Щелевые (фрикативные) согласные

Щелевые согласные образуются в результате большого ограничения речевыми органами звукового потока. В казахском языке щелевые согласные разделяют по участию голоса: глухие – [Ф, С, Ш, Щ, Х, Һ], звонкие – [В, З, Ж, Ғ]. Данное свойство выражается в наличии

квазипериодичного колебания во временно́й области и, соответственно, образованием энергии в низкочастотной области на спектрограмме. Другими основными отличительными характеристиками щелевых звуков являются диапазон частот концентрации энергии и длительность самих звуков.

Наиболее высоко располагается энергия для звуков [С, З] – от 4500 до 15000 Гц, среднее значение показывают звуки [Ш, Щ, Ж, Х, Һ, Ғ] – около 2000-8000 Гц, и наименьшее значение у звуков [Ф, В] – 700-2000 Гц. Стоит отметить, что звук [Щ], заимствованный из русского языка, в казахском языке является аллофоном звука [Ш]. В частности, [Ш] звучит как [Щ] в мягких словах. Звуки [Ж] и [Ш] зачастую произносятся как аффрикаты [ДЖ] и [ТЩ] в некоторых южных регионах страны. Звуки [Ф] и [В] являются заимствованными звуками, в силу чего они зачастую произносятся сельскими жителями или менее образованными людьми как [П] и [Б], соответственно.

Длительность звуков также способна различать классы звуков между собой. Так, глухие звуки длиннее, чем звонкие, и имеют в среднем продолжительность 73-116 мс и 44-58 мс, соответственно. Кроме того, длительность звуков [Ф, В] гораздо меньше, чем у остальных звуков – около 30-40 мс.

Сонорные (носовые, назальные) согласные

Назальные согласные звуки [М, Н, Ң] по артикуляции совпадают со смычными [Б, Т, Қ] за тем лишь исключением, что при произношении первых открывается носовая полость, и поток воздуха выходит оттуда беспрепятственно, не образуя взрыва, как в смычных звуках. В связи с этим, назальные звуки характеризуются выраженной низкочастотной энергией и слабыми формантами. Переходы второй форманты способны сигнализировать о месте артикуляции назальных звуков. Так, для звука [М] вторая форманта F2 располагается низко – около 1300 Гц, а, соответственно, в контексте VCV («гласный–согласный–гласный») F2 опускается от предстоящей гласной и поднимается к последующей гласной. В звуке [Н] форманта F2 находится около 2000 Гц, поэтому переходы F2 от предстоящей и последующей гласных происходят плавно на одном уровне. В звуке [Ң] форманта F2 находится на уровне 2500 Гц, что приводит к поднятию F2 от предстоящей гласной и опусканию к последующей гласной. Отметим, что переходы второй форманты явно видны на спектрограмме ввиду того, что форманты назального звука очень слабы. Длина назальных звуков может колебаться от 45 до 73 мс.

Сонорные (аппроксиманты)

При произношении аппроксимант речевые органы не сильно блокируют воздушный поток, в результате чего их акустические характеристики аналогичны гласным звукам. Аппроксиманты обладают явной формантной структурой, но менее выраженной, чем у гласных.

Аппроксимант [У] по артикуляционным признакам совпадает с гласным [У], но у первого звука артикуляторы более напряжены и сжаты. Форманты F1 и F2 расположены очень близко около 1800 Гц, а F3 проходит равномерно на уровне 2700 Гц.

Аппроксимант [Й] также похож на гласный звук [И]. Он характеризуется низкой первой формантой (700 Гц) и высокими вторыми и третьими формантами (2300 Гц, 3000 Гц). Оба звука отличаются явно выраженными переходами формант в контексте гласных.

Аппроксимант [Р] произносится небольшим дребезжанием языка, в результате чего на спектрограмме внутри самого звука образуется паузообразный интервал (длительностью около 7-8 мс), как у смычных звуков.

Аппроксимант [Л] отличается от других сонорных тем, что воздушный поток проходит по краям ротовой полости, а кончик языка прижат к альвеоле. Его первые три форманты располагаются в частотах 450 Гц, 1300 Гц и 2500 Гц, соответственно.

Акустические характеристики гласных звуков казахского языка

Акустический анализ гласных звуков казахского языка был проведен во временно́й и частотной областях на основе аудио данных восьми мужских и восьми женских голосов носителей языка.

В казахском языке различают следующие основные гласные [А, Ә, Е, И, О, Ө, У, Ұ, Ү, Ы, І, Э]. Остальные гласные [Ё, Ю, Я] являются дифтонгами, то есть, фактически каждый из них представляет собой сочетание двух голосовых фонем [3].

При образовании гласных воздушный поток, сформированный легкими, свободно проходит через весь речевой тракт (для нормальной нешепотной речи). Благодаря определенной конфигурации речевого тракта обеспечивается специфическая форма спектра, типичная для данного звука. Гласные могут искусственно продолжительно «тянуться», и в речевом потоке при нормальном темпе ударные гласные обычно имеют участок, где их характеристики оказываются относительно стационарными. Однако зачастую гласные (особенно в безударных позициях) проявляются в своей характерной форме всего лишь на очень короткий период, а иногда артикуляционный аппарат вообще не успевает принять правильное положение для произнесения гласного. Окружающие звуки (контекст) и темп речи значительно влияют на качество гласных и на их акустические характеристики при конкретной реализации [5].

Гласные фонемы различают при этом: по степени подъема языка; по степени его продвинутости вперед – назад; по участию губ [8]. Однако эти показатели являются трудноизмеримыми и относительными.

Наиболее полезную информацию при акустическом анализе гласных мы можем извлечь из спектрограмм, изучая формантную структуру звуков. Гласные звуки имеют ярко выраженную формантную структуру, что отражается в виде чётких полос на спектрограмме. С целью минимизации влияния соседних звуков на гласные, мы рассматривали каждый гласный между двумя смычными (взрывными) согласными (контекст CVC). Основными критериями при анализе гласных звуков являются значения первой (F1) и второй (F2) формант. В дополнение к этим критериям мы рассматривали третью (F3) форманту, а также длительность звука, начиная с установления чёткой квазипериодичности сигнала после начала озвончения (VOT, voice onset time) вплоть до затухания второй форманты сигнала. Усредненные акустические параметры исследованных гласных звуков приведены в таблице 3.

Таблица 3. Акустические параметры гласных звуков

Гласный звук	Среднее значение F1, Гц	Среднее значение F2, Гц	Среднее значение F3, Гц	Диапазон длительности, мс	Среднее значение длительности, мс
А	653	1438	2841	40-130	74
Ә	647	1871	2839	48-195	80
Е	380	2112	2870	35-123	66
И	320	2206	2873	17-131	54
О	470	1113	2987	33-145	70
Ө	411	1276	2660	47-142	74
У	386	1185	2733	24-233	68
Ұ	464	1223	3031	10-99	40
Ү	386	1445	2801	17-87	42
Ы	499	1445	2922	21-66	37
І	447	1722	2745	18-85	38
Э	393	2112	2772	31-214	71

Как известно, артикуляционные признаки, характеризующие позицию языка при произнесении гласных, коррелируют с формантами F1 и F2 следующим образом [9]:

1) чем выше подъем языка, тем ниже F1;

2) чем больше язык продвинут вперед, тем выше F2.

Таким образом, казахские гласные изменяют подъем языка от нижней позиции языка к верхней примерно в следующем порядке: [А – Ә – Ы – О – Ұ – І – Ө – Э – У – Ү– Е – И]. Гласные [А, Ә] являются самыми «нижними», поскольку среднее значение их первой форманты F1 превышает 600 Гц. Гласные [О, Ө, Ұ, Ы, І] занимают среднее положение, поскольку среднее значение их форманты F1 находится в диапазоне 400-500 Гц. Гласные [Е, И, У, Ү, Э] являются «верхними», поскольку среднее значение их форманты F1 не превышает 400 Гц. При этом самым «верхним» является звук [И], его первая форманта заметно ниже, чем у других гласных.

Далее, казахские гласные изменяют позицию языка от задней до передней примерно в следующем порядке: [О – У – Ұ – Ө – А – Ы – Ү – І – Ә – Э – Е– И]. Гласные [О, У, Ұ, Ө] являются самыми «заднеязычными», поскольку среднее значение их второй форманты F2 не превышает 1300 Гц. Гласные [А, Ы, Ү, І, Ә] занимают среднее положение, поскольку среднее значение их форманты F2 находится в диапазоне 1400-1900 Гц. Гласные [Э, Е, И] являются «переднеязычными», поскольку среднее значение их форманты F2 превышает 2100 Гц. При этом самой «переднеязычным» снова является звук [И] с его максимальным значением второй форманты.

Длительность гласных значительно варьируется, поскольку гласные в свободной речи можно и «тянуть» (удлинять, особенно в ударной позиции), и «съедать» (укорачивать, редуцировать). Средняя длительность гласных звуков колеблется в диапазоне 37-80 мс, хотя крайние значения параметра длительности могут значительно отличаться от среднего и колебаться в диапазоне 17-233 мс. Можно видеть, что самыми короткими по длительности произнесения гласными являются звуки [Ы, І, Ұ, Ү]. Максимальная длительность этих звуков гораздо меньше, чем у других гласных, и не превышает 100 мс. Эти четыре звука в слитной речи нередко редуцируются, вплоть до полного исчезновения голосовой составляющей.

На рисунке 1 приведены спектрограммы исследованных гласных звуков.

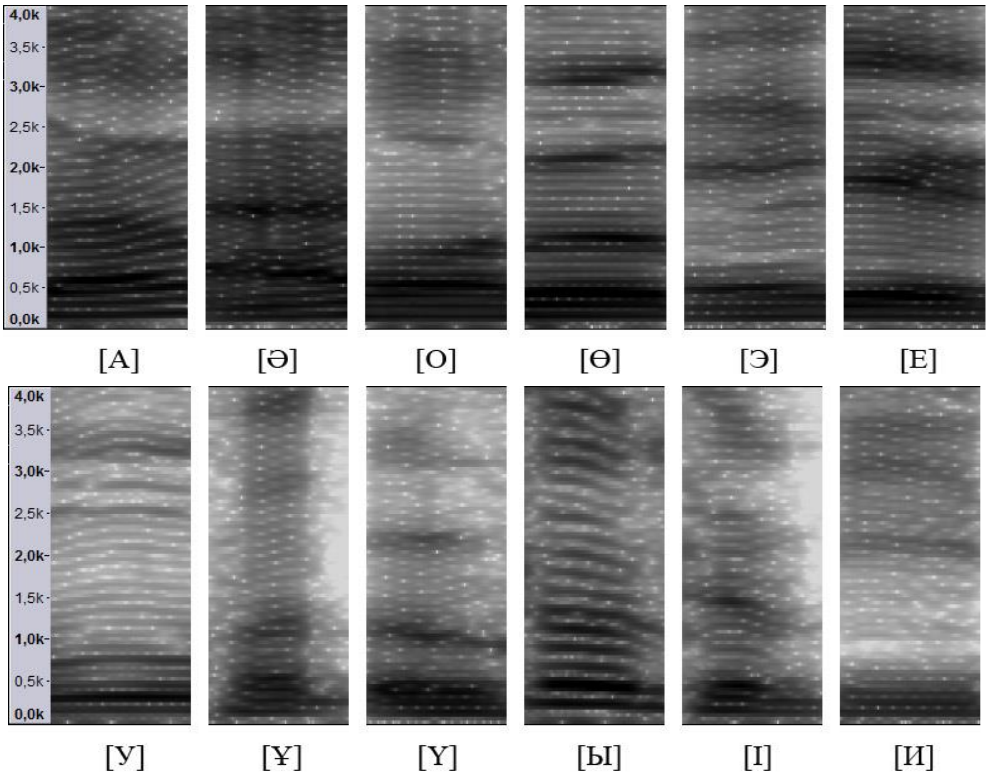


Рис. 1 Спектрограммы гласных звуков для мужского голоса в ударной позиции, отрезки длительностью 80 мс.

Следует отметить, что звук [Э], употребляющийся только в словах, заимствованных из русского языка, по акустическим параметрам очень близок к звуку [Е] в речи носителей казахского языка. Возможно, следует рассматривать эти звуки как аллофоны одной и той же фонемы, аналогично тому, как эти звуки рассматриваются в русском языке.

Заключение

В текущей работе на основе экспериментальных данных был проведен акустический анализ звуков казахского языка с целью уточнения его фонетического строя. Результаты работы дают качественную и количественную оценку акустических характеристик звуков казахского языка во временно́й и частотной областях. Классификация звуков представлена в соответствии с международной фонетической ассоциацией, что дает возможность проведения дальнейших работ по сравнению полученных результатов с аналогичными результатами для других языков.

Литература

1. Послание Президента Республики Казахстан Н. Назарбаева народу Казахстана. 14 декабря 2012 г.
2. Исабаева С. Иммунитет для Qazaq tili. Каким будет латинский алфавит для казахского языка? URL: http://camonitor.com/archives/6768. Дата обращения: 15.09.2013.
3. Исаев С. М. Қазақ тілі. Оқу құралы. – Алматы: Өнер баспасы (ISBN 9965-768-05-6), 2007. – 208 б.
4. Reetz H., Jongman A. Phonetics - Transcription, Production, Acoustics and Perception. – Oxford: Wiley-Blackwell (ISBN 9-78063123-226-1), 2011. – p. 317.
5. Olive J. P., Greenwood A., Coleman J. Acoustics of American English speech. A dynamic approach. – Springer (ISBN 0-387-97984-0), 1993. – p. 396.
6. Баданбекқызы З. Ағылшын және қазақ тілдерінің салыстырмалы фонетикасы. – Алматы: Бастау, 2010. – 227 б.
7. Jones D., Ward D. The Phonetics of Russian. – Cambridge University Press (ISBN: 9780521153003), 2011. – p. 324.
8. Ронжин А.Л., Карпов А.А., Ли И.В. Речевой и многомодальный интерфейсы. – М.: Наука, 2006. – 173 с.
9. Rezaei N., Salehi A. An Introduction to Speech Sciences (Acoustic Analysis of Speech). – IRJ. 2006; 4 (4):5-14.

Ж.А.ЖАҚЫПОВ

Л.Н.Гумилев атындығы Еуразия ұлттық университеті, Астана, Қазақстан

ҚАЗАҚ ЖАЗУЫН ЛАТЫНДАНДЫРУДЫҢ ФЕНОМЕНОЛОГИЯЛЫҚ МӘСЕЛЕСІ

Латын таңбалы қазақ әліпбиіне байланысты нұсқаларды көктей шолғанда оларға ортақ жалпы сипаттарды былайша топтауға болатын сияқты:

біріншісі - бұрынғы 41 кирилдік әріпке латынша таңба қою (бүйткенше, бұрынғы орыс қарпінде қала берген артық);

екіншісі – өткен ғасырдың 30-жылдарында қазақ тілінде қолданылған латын әліпбиін жаңғырту (ол әліпби көпшіліктің келісімімен енбегенін айтсақ та жетер);

үшіншісі - орыс сөздерімен келетін дыбыстардың таңбасын алып тастап, кирилл сипатын қолдану (бұл механикалық әрекет, ұлт рухы ескерілмеген);

төртіншісі – жат тілге тән дыбыстарды сақтау (бұл – қазақ тілін жат тілдің қысымына салу);

бесіншісі – қазақ дыбысталымын негізге ала отырып жат тіл дыбыстарына басқаша дыбыстық тіркесімдер, таңбаларды қосымшалау (мұндайда жат тілдік сөздер сол қалпымен өзгермей енеді де, қазақ тілінің дыбысталымдық мәні, үндесімі бұзылады. Жат тілден келген сөзге әлдебір аса мәртебелі, киелі сөздей қарайды жалпыш қазағымыз. Қазір университетті «үніберсійтет» деп айтып жазып көр, сенен асқан надан адам болмай шығады, вагонды «бәгөн» деп көр, ешкімнің мінбей қоюы мүмкін!).

Көріп отырғанымыздай, бұларда кемелдік болмай тұр. Қалай десек те, отарлық сананың санамызды таптап тастағаны сондай, ғалымдар біле тұра қазақ тілінің асыл нұсқасын қалпына келтіруге дәрменсіз болып тұр. Ал, әсіресе, совет дәуірінде қазақ тілінің дыбысталым, айтылым, жазылым заңдылықтары әбден бұзылып «қыр құлақ» болып тұрғаны мынау! Осы тұста ана тіліміз қазақ тілінің әзелден Алла Тағала тағайынаған дүниетанымдық, ұлттың рухи әлемінің ұстыны екендігі туралы ғылыми, тарихи шегініс жасап кеткен артық болмас.

Қазақ тілінде сөздің дыбысталуында терең этногенезистік, рухани танымдық негіз бар. Бізде сөздің дыбыстық құрамы «ерлік (немесе аталық)» және «әйелдік (немес аналық)» негізден тұрады. «Ерлік» негіз дауысты дыбыстар болса, «әйелдік» негіз - дауыссыз дыбыстар. Екеуі қосылғанда әріптік дыбыстардың таңбасы жасалып қана қоймайды, «рух, жан» бітеді. Қазақ тілінің үндестік заңында осындай Ғаламдық Гармония (Жарасым, Үйлестік) бар. Оқығандар біледі, батыстық, орыстық философия бойынша қарама-қарсылықтар қайшылықта болып, солардың күресі арқылы әлем дамиды, ал қазақта қарама-қарсылықтар үйлесім жасайды. Біздің Әлеміміз басқа ғой, басқа! Бұл Үйлесім сонау көне түркі бітігінде орнаған еді. Айрылып қалдық! Қазақ соны қайта қалпына келтірсе, түркі әлемінің нағыз рухани көшбасшысы болар еді, рухани әлеміміз де ретке келер еді. Амал жоқ! Мұны айтсаң, қазіргі қоғам динозаврды қайта тірілтейік деген сөзбен бірдей көреді.

Бұл мәселені тереңдей зерттесе, талай сыр ашылатын сияқты. Оның үстіне, батыстық сананы зерттеушілер әріпте Құдайдан түскен сыр бар екеніне әбден сендіреді. Олардың айтуынша, латын әрпін Құдай католиктерге жолдаған, кирилл әрпін православтарға тиесілі еткен.

Осы сияқты терең рухани-танымдық негізді жоғалтпау үшін, сақтау үшін біз **тіліміздіәліпбиге емес, әліпбиді тілімізге бейімдеуді** ұмытпауымыз қажет. Ал қазір жұртшылық таңба ауыстыру ретінде ғана қарап жүрген сыңайы бар. Латын қарпіне көшуде таза технократтық бағыттар басымдық алып кететін сияқты. Сондықтан бұл жерде тіл фонологиясының, семантикасының мамандары шешуші тұрғыға ие болғанын жөн санаймын. Біз басқа тілден енген сөздерді таңбалауда да, айтуда да осыны ұмытпағанымыз жөн.

Бұл істің саяси, экономикалық, мәдени-рухани жақтары туралы аз сөз. Латын әліпбиіне көшсек, орыс досымызбен қалай боламыз? Ұлттың рухани мәселесін шешерде оған жалтақтасақ, тәуелсіздігімізде не құн қалады! Ақпаратты әлемдік дереккөздерден тікелей алғанымызға, жаңа технологияға қазақ тілін оңтайлы енгізуге шын дос неге қарсы болады? («Шын дос неге қарыс болады» дегеннен шығады, бар сенген Ағамыз Мұқтар Шаханов латынға қарсы шалқалағанын көріп, шалқамнан түсе жаздадым). Орыс бауырларға айтарымыз бір ғана жауап – латын әрпіне мемлекеттік тіл көшеді.

Экономикалық жағы соңыра бұқара халықты сауаттандыру, баспаға енгізу, бұрынғы жазба мұраны латын әрпіне көшірумен байланысты. Қазақ жазуын латын әліпбиіне Ахмет Байтұрсынұлы шынайы да нақты көрсеткен жолмен көшірсек, 28 төл дыбысты белгілейтін 28 әріп болады, 42 әріпті бір айға жетпей үйреніп алғанда, 30 шақты әріпті онан да тез үйреніп алмаймыз ба! Аз уақыт оқытуға мемлекетте қаражат жетеді деп ойлаймын.

Баспаның көшуі онан да оңай, латын әріптері қазіргі компьютерлерде, баспа жабдықтарында баршылық. Бұл көп шығын талап ете қоймас.

Ал мұраларымызды жаппай көшірудің қажеті де шамалы, керегін ұрпақ алады, оның үстіне кирилл әріптері санамыздан өше де қоймас, ол әріптер әрі-беріден соң археологияны қажет етпейді ғой.

Ресми тараптың көздеп отырғаны – латын әліпбиіне 2025 жылы көшу. Бір күнде ауыса салу мүмкін емес екені белгілі. Биыл күзде Үкімет жанынан арнайы комиссия құрылмақшы екен. Демек, 2025 жылы толықтай көшуге дайындық жүреді деген сөз. Ол дайындықтар кезең-кезеңімен іске асатын болады. Соған орай мынадай істерді қазірден қолға алса жақсы болар еді.

Астанада салтанат құратын «Экспо-2017» дүниежүзілік көрмесіне дейін латын қарпіне көшіп алсақ құба-құп болар еді, ең болмаса сол уақытқа дейін көрнекі ақпаратты осы қаріпке түсірсек тиімді болатынын айтқымыз келеді.

«Қазақ» гәзетінің 20-ғасыр басында қазақтарды тамаша сауаттандырған тәжірибесін ескеріп, мерзімдік басылымдар да қазірден ептеп кірісе берсе, теріс болмас еді (айталық, кейбір маңызды да қысқа ақпарларды латын әрпімен беріп тұрса, т.с.с.).

Тіл мамандары қазақ фонологиясы тұрғысынан емле сөздігін жасауға кірісе бергені жөн.

Қалай дегенмен, осы бір мүмкіндікті жіберіп алмай, қазақ тілінің Гармониясын қалпына келтіріп алу қажет-ақ. Осы мүмкіндіктен айрылып қаламыз ба деп қорқамын. Біздіңше, көне түркіден бері біздің рухымызды ұстап тұрған – Гармония.

Ғалым Қ.Сартқожаұлы түркі жазба жадыгерліктеріндегі сөздің дүниетанымдық негізі туралы былай деп жазады: «Байырғы түркі графикасы түркілік дүниетанымды тұғыр еткен түркілік мәдениет қазанында пісіп жетілген түрік тілінің фонетикалық жүйесіне негізделген. ... «Екі негіз» ұғымы – байырғы түркілердің әлемді түсіну философиясы. Көне түркілер пайымында әлем аталық пен аналықтан тұрады. ... ...қытай философиясындағы «инь» мен «янь»ұғымына ұқсас. Европа түсінігінде дуализмге келеді. Европа философиясы бойынша болмыс немесе құбылыс қарама-қайшы, бір-біріне бағынбайтын тең құқықты екі нәрседен тұрады. Олар өзара бәсекелестікте, бірін-бірі жоққа шығару арқылы дамуды алға жылжытады. Ал Шығыс халықтарында, соның ішінде байырғы түркілер дүниетанымында жоғарыдағы екі ұғым бір-біріне сүйене, демеу болып, бірін-бірі толықтыра отырып дамиды. ... Түркілер «екі негіздік» дүниетанымды ұстана отырып, түркі әлемінің барлық болмыс-бітімін осы жұптық жүйеге бейімдеді. Түркілердің ...ауызекі сөйлеу және графика жүйесіндегі дыбыстың сингармонизмі мен консонантизмі, т.с.с. өмір мен өлім арасындағы болмыс пен құбылысты «екі негізге» ... негіздегенін көрсетеді» /2/. Бұл екі негізді ғалым амал және білік деп атайды. Осы ойдың ағымымен көне түркі тіліндегі дауысты дыбыстарды - Жанға, дауыссыз дыбыстарды Тәнге балайды. Көне түркі мәтіндері арқылы әлем бейнесін бақылап отырған тіл тарихшысының бұл танымы болмыс пен сана диалектикасына қайшы келмейді.

Қазақ тілінде сөздің дыбысталуында терең этногенезистік, рухани танымдық негіз бар. Бізде сөздің дыбыстық құрамы «ерлік (немесе аталық)» және «әйелдік (немес аналық)» негізден тұрады. «Ерлік» негіз дауысты дыбыстар болса, «әйелдік» негіз дауыссыз дыбыстар. Екеуі қосылғанда әріптік дыбыстардың таңбасы жасалып қана қоймайды, «рух, жан» бітеді. Қазақ тілінің үндестік заңында осындай Ғаламдық Гармония бар.

Гармония, әрине, негізінен музыкаға тән ұғым. Алайда музыканың тілдің фонологиясына сүйенетінін ескерсек, Гармония құбылысының тілге де байланысты айтуға әбден болады деп санаймыз.

Араб философиясының халифат дәуірінде Гармонияға зор мән берілген. Оның алғашқы ірі өкілі әл-Кинди түрлі ғылымдарды жіктеу барысында жарасымды жеке ғылым ретінде бөледі. Оның бұл ойлары «Гармония туралы үлкен кітап» атты еңбегінде баяндалады. Әл-Кинди: «Гармония (мұнан былай – Жарасым деп аламыз) ғылымы бір санның келесі санмен жалғануы мен қатынасуын анықтаудан, өлшемдестік пен өлшемдес еместікті ажыратудан тұрады. Гармония барлығында бар, ал оның анық табылатын жері дыбыстарда, ғалам мен адам жанында» (курсив авт.), - деп жазады [3]. Бізге бұл тұжырымдамадағы маңыздысы сол, мұның негізі шығыстық орта ғасырда болған әлем суретін барынша жақын бейнелейді. Ол

суретте Дыбыс, Адам мен Ғарыш біріге келе жарасымды бүтін түзеді. Біз зерттеп отырған көне замандағы арғы қазақтар мен ортағасырлық дәуірдегі қазақтардың әлем суретінде де, олар үшін Дыбыс пен Ғарыштың жарасымды байланысы шешуші орынға ие болған.

Жарасымды түсіндірудің ұтымды тұжырымдамасы әл-Фарабидің (шамамен 870 – 950 жылдары) философиялық көзқарастарынан көрініс тапты. Ол - Пифагордың жарасымның аспан сфераларының қозғалысына тәуелді болатындығы туралы тұжырымдамасынан бас тартты. «ғаламшарлар мен жұлдыздар қозғалғанда жарасымды түрде сабақталатын дыбыс туғызады дейтін пифагорлықтардың пікірі жаңсақ, - деп жазады ол. – Олардың болжамы іске аспайтындығы, аспан шырақтары мен жұлдыздардың қандай да бір дыбыс туғыза алмайтындығы физикада дәлелденген» [4]. Осылайша жарасымды космогония тұрғысынан түсіндірудің орнына антропологиялық түсіндірме келді. Мұның қағидаттары ағзаның әрі физикалық, әрі рухани дамуы үшін бірдей қолданылады. Әл-Фараби музыкасында Жарасымның шешуші белгісі адами түйсіктер болып саналады, соған сәйкес тыңдарман рақат пен эмоционалдық ләззат шақыратын табиғи (орындалған) түйсікті және қажу мен шамырқану шақыратын бейтабиғи түйсікті ажыратады [5].

Жарасым санаты әл-Фараби еңбектерінде тек музыкада ғана айқындалып қоймайды, музыкалық ғылымға тұтастай қолданылады. Өйткені музыка мен оның психофизикалық ықпалы адамның рухани күйіне де, физикалық күйіне де жарастырушылық қызмет жасайды. «Бұл ғылым мынадай мәнде алғанда пайдалы: тұрақты күйден айырылғандардың мінез-қылықтарын жұмсартады, кемелдікке әлі жетпегендерді кемелдендіреді және тұрақты күйде тұрғандардың тұрақтылығын бекітеді. Бұл ғылым тән саулығы үшін де пайдалы, өйткені тән ауырғанда, жан да жүдейді, тән ауру кешкенде, оны жан да бастан кешеді. Сондықтан жан сауыққанда, жанның қуаты жұмсарып, осыған ықпал ететін дыбыстардың арқасында жанның субстанциясына бейімделгенде, тән сауығады»[6].

«Жарасым» ұғымы Таяу Шығыста «Тазалық бауырлары» (X ғасыр) дейтін атпен белгілі діни қозғалыстың «Жолдауларында» натурфилософиялық тұрғыдан айқындалды. Олар табиғат пен өнерде жарасым идеясы бұларға ұқсастық, сәйкестік пен тектестік тән болғандықтан және табиғи заттардың әсемдігі олардың құрылымының пропорционалдылығы және оны құрастыратын бөліктердің жарасушылығына тәуелді болғандықтан салтанат құрады.

Батыста антикалық эстетикаға тән ғарыштық жарасым идеясы ортағасырлық дәуірде жер мен көк, адам мен құдайдың иерархиясымен таразыланды. Қайта ояну дәуірінде жарасымның мағынасы кеңейіп (әсемдік ұғымы енгізілді) әсемдіктің ішкі мазмұны ретінде ұғынылды, бұл санатты эстетикалық тұрғыдан негіздеуге алып келді.

Әлеуметтік психология мен тәрбие теориясы салаларына жарасым туралы ілім Ағарту дәуірінде енді, бұл дәуірде дамыған адамды тәрбиелеп шығаруға, жеке мүдде мен қоғамдық мүддені үйлесімді сабақтастыру проблемасына қатты зейін қойылған болатын.

Романтизм дәуірінде Жарасымды ретсіздік пен дисгармонияны еңсеру арқылы жарасымдылықты көрсетуге ұмтылатын көркем санат ретінде ұғыну ұсынылды.

Қазіргі ғылымда эстетикалық жарасым әмбебап санатқа айналды, оның қағидаттары табиғатта, адамда, өнерде де, жалпы эстетикалық қарекетте көрініс табады. Дәл осы себептен бұл ұғым ырғақ, пропорция, симметрия, кемелдік сияқты шектес ұғымдармен, сондай-ақ ақыл, қарекет пен эмоциялылық сияқты адами қабілеттермен қатынасқа түсетін болды. Мәселен, XX ғасырдағы батыс эстетикасында француз «реалдық эстетикасының» өкілі Шарль Лоло жарасым санатын әр түрлі күйде: парасат пен талғамға негізделген әсемдік ретінде; қарсыласқан нысанды жеңетін салтанаттылық ретінде; ұнатудан (симпатиядан) туындап маңызсыз нәрсеге дейін жететін сымбаттылық ретінде қарастырады. Бұған қоса, мұны зерттеудің философиялық, санаттамалық, өнертанулық сияқты түрлі аспектілері өмірге келді.

Сөйтіп, нақты әлеуметтік шарттардың ықпалымен анықталатын «жарасым (гармония)» ұғымы бір дәуірден келесі дәуірге өзгере жетіп, әр тарихи кезеңдегі Әлем-үй (Мироздание) туралы түсініктерді бейнелеп отырғанын анық көреміз. Демек, жарасым жайында теориялық

және тарихи санат ретінде сөз қозғауға болады. Екі жолды сабақтастыру бізге уақыт жағынан алшақ дәстүрлі мәдениетті зерттеуде оңтайлы жол болып көрінеді. Өйткені жарасым туралы бүгінгі түсініктің біздің ата-бабаларымыз мұндай түсінігінен айырмашылығы мол, уақыттың өзі мәдениетке де, «жарасым» санаты ұғымына да түзетулер енгізіп отырған.

Жарасым туралы мұндай теориялық түсінік қазақтардың Универсум туралы түсінігіне анағұрлым жақын келеді. Қазақтар Универсумды барлық элементтері теңгерілген қайшылықсыз бүтін деп ұғады, мұның өзі қазақ мәдениетінің жарасымын белгілейді.

«Жарасымда» әлемді игерудің ұлттық ерекшелігінің заңдылықтары анағұрлым толық бейнеленеді, себебі бұл санат (категория) қазақтар үшін дәстүрлі мәдениеттің түрлі элементтерінің арасындағы логикалық және тұрақты байланыстардың өзі болып саналады. «Жарасым» санатының тағы бір ерекшелігі – болмыс пен таным санаттарының ұқсастығын жай білдіріп қана қоймайды, олардың арасындағы логикалық, себеп-салдарлық байланыстарды да орнатады. Сондай-ақ қазақтардың әлемді игеруі болмыс пен ойлаудың негізгі санаттарынан да, қандай да бір құндылық жасау ісінен көрінеді.

Қазақ тілінің гармониясы жат сөздер ықпалымен, нақты айтқанда, оларды сол қалпынша енгізудің кесірінен бұзылып жатыр. Енді осы бір рухани Жаңғыру, Қайта ояну сияқты мүмкіндік туып тұрғанда, соны пайдаланып қалу Ғаламдық міндет деп санаймын. Жат сөз келген тіліне бейімделсін, бұл кез келген дамыған тілдің қағидасы, ал әріп тілге бейімделуге тиіс.

Әріптің мәні тереңде дейтініміз де осы айтылған тылсым жайттармен байланысты.

Әдебиеттер

1. Шәріпбай А. Қазақ жазуын латын әліпбиіне көшіру жобасы. – Астана, 2013.

2. Сартқожаұлы Қ. – Орхон мұралары. – Алматы, 2012. – 30-45-беттер.

3. Избранные произведения мыслителей стран Ближнего и Среднего Востока IX–XIV вв. / Сост. С.Н. Григорян и А.В. Сагадеев. – М.: Соцэкгиз, 1961. – С. 49–50.

4. Аль-Фараби. Естественно-научные трактаты / Пер. с арабского. – Алма-Ата: Наука, 1987. – С. 208.

5. Сонда. Б. 221-222.

6. Аль-Фараби. Трактаты о музыке и поэзии / Пер. с арабского. – Алматы: Ғылым, 1993. – С. 338–339.

А.А. ШАРИПБАЙ, А.С. ОМАРБЕКОВА

Евразийский национальный универсистет имени Л.Н.Гумилева, Астана, Казахстан

КОНВЕРТАЦИЯ ТЕКСТА НА КАЗАХСКОМ ЯЗЫКЕ С КИРИЛЛИЦЫ НА ЛАТИНИЦУ

В научно-исследовательском институте «Искусственный интеллект» ведутся исследования по переходу казахского языка с кириллицы на латиницу.

1 Предложен проект алфавита казахского языка на основе латиницы и разработан алгоритм перевода казахской письменности с кириллицы на латиницу.

Сначала предлагаются следующие критерий определения нового алфавита:

1) новый алфавит должен создаваться только на основе звуковой системы казахского языка.

2) новый алфавит должен создаваться на основе научного анализа частоты встречаемости букв действующего алфавита в казахском тексте.

3) новый алфавит должен использовать только буквы из классического латинского алфавита, имеющиеся в стандартной клавиатуре.

4) адаптация латинского алфавита казахскому языку производится путем изменения значений некоторых букв казахскими звуками.

5) в новом алфавите порядок следования букв должен совпадать с порядком их следования в классическом латинском алфавите

Затем предлагается новый казахский алфавит, который основывается на классическом латинском алфавите. Для его определения с помощью компьютера исследованы звуковая система казахского языка и частота букв и буквосочетаний в текстовом корпусе, состоящего из 100 миллионов букв используемого в настоящее время кириллического алфавита.

Для автоматизации перевода казахской письменности с кириллицы на латиницу построен алгоритм конвертации в казахских текстах кириллических букв на латинские буквы.

Конвертация будет проходить в два этапа: на первом этапе исходный текст на кириллице преобразуется в промежуточный текст тоже на кириллице, где освобождаются от всех букв (*ё, э, и, ю, я, ц, ч, щ, ь, ъ*), которые обозначают не исконно казахские звуки; на втором этапе промежуточный текст на кириллице преобразуется в результирующий текст на латинице в соответствии с алфавитом.

2 Разработан конвертер казахского языка с кириллицы на латиницу.

Конвертер размещен на сайте www.alphabet.kz (рисунок 1).

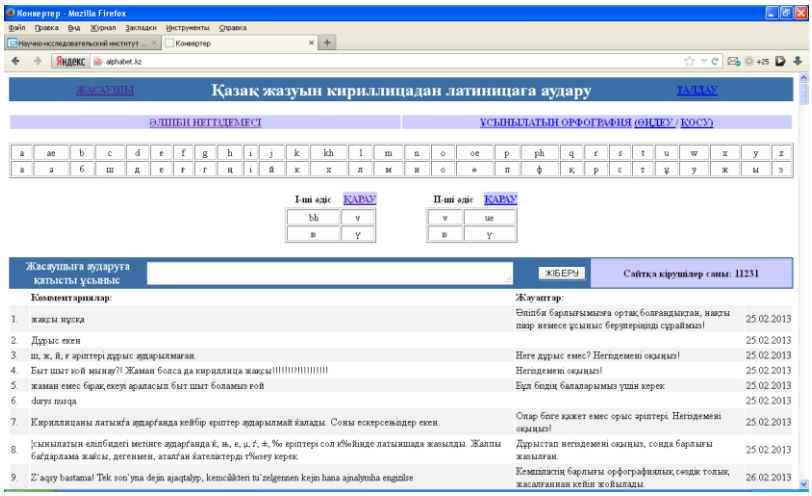


Рисунок 1. Конвертер казахского языка с кириллицы на латиницу

Данный сайт предоставляет возможность пользователю просмотреть обоснование (ӘЛІПБИ НЕГІЗДЕМЕСІ), утвержденную орфографию (ҰСЫНЫЛАТЫН ОРФОГРАФИЯ), оставить комментарий/предложение. Реализовано два варианта конвертера. Для просмотра конвертера необходимо нажать соответствующую ссылку «ҚАРАУ» (рисунок 2).

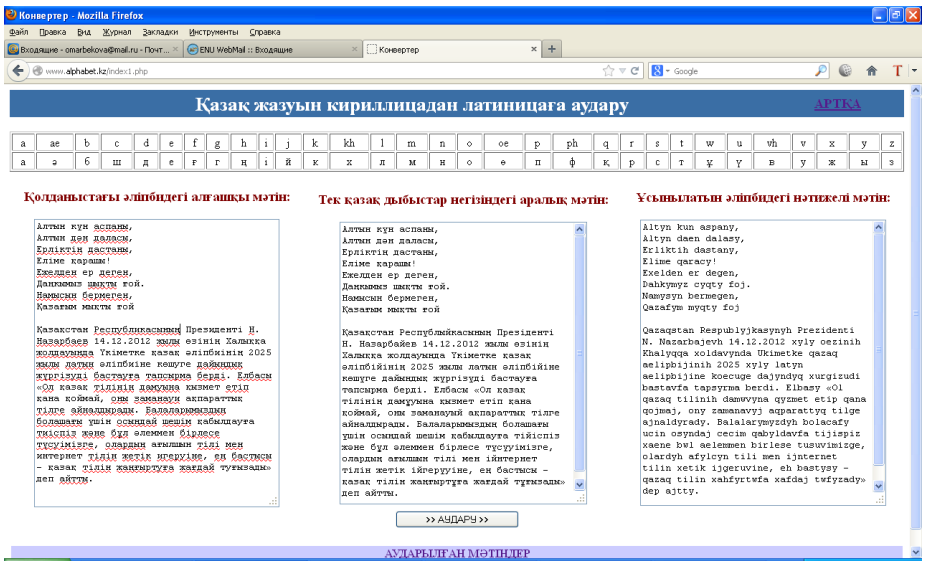


Рисунок 2. Форма конвертирования текста

Для конвертации в первое окно необходимо ввести или скопировать исходный казахский текст в действующем кириллическом алфавите. После нажатия на кнопку «АУДАРУ» в среднем окне появится казахский текст в промежуточном кириллическом алфавите, содержащего буквы только для казахских звуков (без букв *ё, э, и, ю, я, ц, ч*, щ ь, ъ), а в третьем окне казахский текст отражается в предлагаемом казахском алфавите, состоящего только из 26 латинских букв.

При нажатии на ссылку «АУДАРЫЛҒАН МӘТІНДЕР» откроется база сконвертированных текстов на казахском языке на кириллице и латинице (рисунок 3).

Рисунок 3. Сконвертированные тексты

3 Разработан генератор электронных учебных изданий на латинице

Авторская подсистема генератора ЭУИ

С помощью данной системы преподаватель-непрограммист может создать свое ЭУИ в сети Интернет (на кириллице, на латинице), которое полностью соответствует государственному стандарту СТ РК 34.017-2005 «Информационная технология. Электронное издание. Электронное учебное издание».

Запустите браузер Internet Explorer или Mozilla Firefox (нужно установить дополнения для работы с файлами mht), введите в адресной строке адрес **www.e-zerde.kz/t_oas/**. Начальная страница генератора ЭУИ показана на рисунке 4.

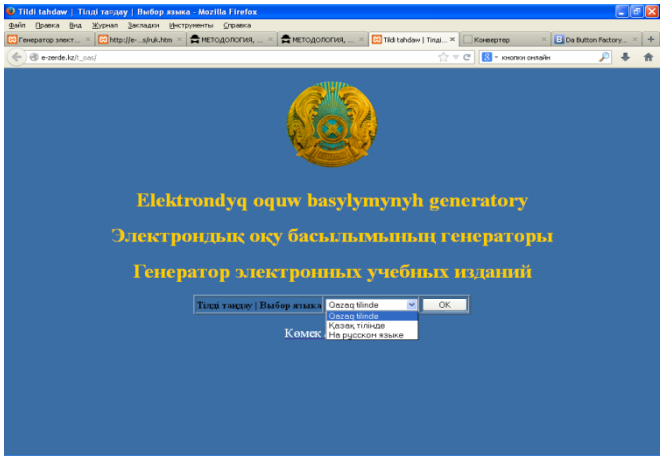


Рисунок 4. Титул генератора ЭУИ

После выбора языка (казахский язык на латинице, кириллице, русский язык) пользователя необходимо пройти авторизацию.

При первом входе в систему необходимо пройти регистрацию, для этого нажмите ссылку «Tirkew» («Регистрация»).

Сначала необходимо из ниспадающего списка выбрать тип пользователя: vjrenvci (обучающийся), tvjtor (тьютор). Если пользователь намерен самостоятельно создавать ЭУИ, необходимо выбрать пункт «тьютор». Если пользователь намерен обучаться по ЭУИ, созданным с помощью генератора другими тьюторами, необходимо выбрать пункт «обучающийся». Также необходимо ввести логин, пароль, повтор пароля, фамилию, имя, отчество, e-mail, факультет, кафедру, телефон, должность.

Если Вы зарегистрировались как тьютор, откроется окно формирования ЭУИ.

Авторская система позволяет создавать ЭУИ, формировать трехуровневую структуру ЭУИ, вводить теоретический материал, примеры, задания, вопросы, тесты, справочник, мультимедиа к каждому уроку (рисунок 5).

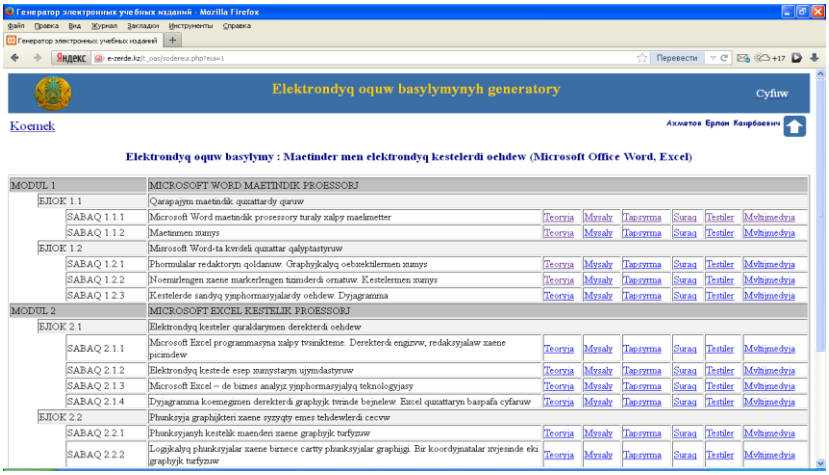


Рисунок 5. Ввод содержимого уроков

После завершения формирования ЭУИ, тьютор сохраняет ЭУИ. Теперь данное ЭУИ будет доступно администратору системы, который должен дать разрешение на опубликование

ЭУИ. После чего оно будет доступно другим пользователям генератора электронного учебного издания.

Пользовательская подсистема генератора ЭУИ

Для просмотра готовых ЭУИ, разработанных тьюторами на сайте генератора ЭУИ, необходимо зарегистрироваться как «обучающийся». Появиться список ЭУИ.

При нажатии на наименовании ЭУИ откроется титульная страница ЭУИ, на которой отражается информация об авторах, аннотация, оглавление, помощь.

Для начала использования ЭУИ необходимо нажать «Mazmuny» (Содержание). Откроется форма позволяющее обучаемому выбрать режим работы.

Рисунок 6. Выбор режима работы

Первый режим просмотра (QARAW REXIMI). В этом режиме обучающая программа обеспечивает просмотр только учебного материала. При этом нет доступа к заданиям, вопросам, тестам, анимациям.

Второй режим тестирования (TESTILEW REXIMI). В этом режиме обучающая программа обеспечивает тестирование по всему объему учебного материала. При этом после тестирования можно получить информацию о результате тестирования.

Третий режим начала обучения (OQJTUDJ BASTAW REXIMI). Режим начала обучения позволяет сформировать траекторию обучения тремя способами: ручным, тестовым или полным выбором (рисунок 7).

Четвертый режим продолжения обучения (OQJTVDJ XALFASTJRUW REXIMI). В этом режиме обучающая программа обеспечивает продолжение обучения по выбранной траектории. При этом процесс обучения начинается со следующего урока после прерывания.

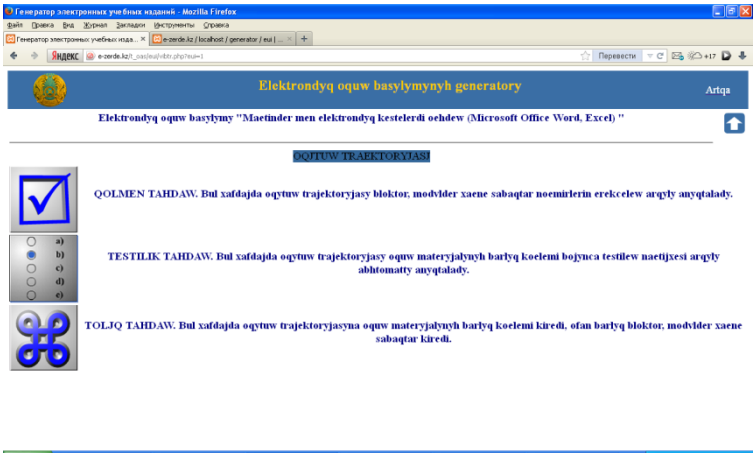


Рисунок 7. Выбор траектории обучения

При ручном выборе (QOLMEN TAHDAW) траектория определяется обучаемым самостоятельно путем отметки номеров модулей, блоков, уроков.

При тестовом выборе (TESTILIK TAHDAW) траектория определяется автоматически по результатам тестирования по всему объему учебного материала. В этом случае в траекторию обучения включаются только те уроки, по вопросам которых были получены недостаточное количество правильных ответов.

При полном выборе (TOLJQ TAHDAW) в траекторию включается весь объем учебного материала данной дисциплины, включая все уроки, модули и блоки.

После определения траектории пользователь переходит непосредственно к обучению текущего урока (рисунок 8).

Рисунок 8. Изучение текущего урока

Для перехода к изучению следующего урока необходимо правильно ответить на тестовые вопросы текущего урока. В случае недостаточного количества правильных ответов на тесты (<75%), обучаемый не сможет перейти к следующему уроку в траектории и будет продолжать изучение текущего урока.

Контролирующая подсистема генератора ЭУИ

При входе в систему под администратором предоставляется список всех ЭУИ.

Если формирование ЭУИ уже завершено, справа от наименования будет находится ссылка «XARYJALAW» («ОПУБЛИКОВАТЬ») (рисунок 9).

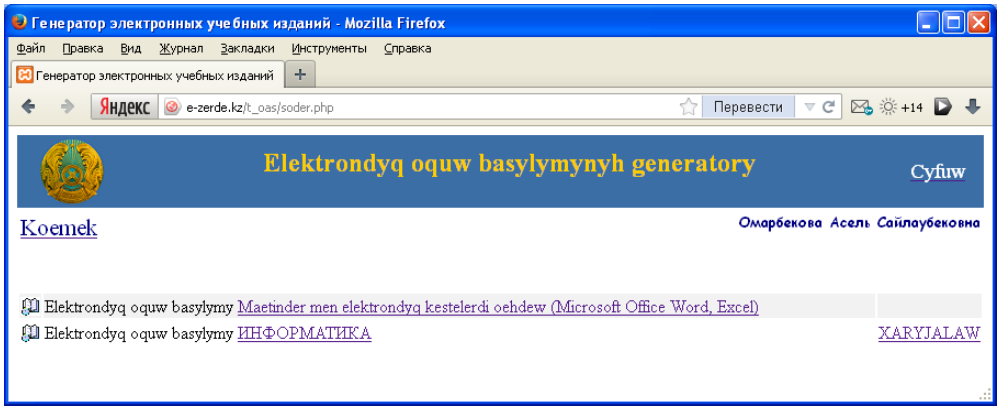


Рисунок 9. Форма опубликования ЭУИ

При нажатии на ссылку «XARYJALAW», ЭУИ станет доступен для просмотра обучающимся. Ссылка «XARYJALAW» будет заменена на строку «XARYJALANFAN» («ОПУБЛИКОВАНО»).

4 Разработана база обучающих анимационных роликов для обучения казахскому языку на латинице по различной тематике (рисунок 10).

Рисунок 10. Обучающие анимационные ролики.

Заключение

Разработанные конвертер, генератор ЭУИ, обучающие анимационные ролики предназначены для облегчения процесса перехода казахской письменности с кириллицы на латиницу.

A.A.SHARIPBAY, K.ANUARBEKOV

L. N. Gumilyov Eurasian National University,
Scientific-Research Institute "Artificial intelligence"

CONVERSION OF KAZAKH WRITING INTO LATIN ALPHABET PROJECT

Annotation

The project proposes conversion of Kazakh language writing into Latin alphabet, which alows working with various types of computers and communication devices more effectively.

Voice system of Kazakh language and frequency of letters and letter combinations in a text body, which consists of 100 million letters of current Cyrillic alphabet, were investigated to support the project.

1. Introduction

President N. A. Nazarbayev in his message to the people of Kazakhstan from 14.12.2012 said: "We need to start **conversion of our alphabet to Latin alphabet** starting from the year of 2025. It is a principled issue that nation must solve. For the sake of the future of our children we must make this decision and it will provide conditions for our integration in the world and for more effective

study of **English language** and **language of the Internet** by our children, but the most important thing is that it will give a push to modernisation of Kazakh language. We must **modernise and develop Kazakh language**. It is necessary to make the language modern, to search for a consensus in the questions of terminology, once and for all to solve a question about translating of established international and foreign words into Kazakh language. This issue must not be solved by the group of the separate personalities. The Government should take care of it".

In many countries there is a discussion that in the epoch of globalization global informative space must have a single alphabet and all informational resources must be created with the use of the classic Latin alphabet. Our proposal directly relates to this issue. Because all computers and telephones, produced all over the world, initially support and will support the classic Latin alphabet, used in English language. If another national language has even one letter which differs from the classic Latin alphabet, then there is a need to create additional fonts, drivers, sorting and searching programs in order to work on these devices with this alphabet. It also requires a lot of labor and financial expenses.

2. Classic Latin alphabet and methods of adaptation

The classic Roman alphabet consists of 26 letters: ***Aa, Bb, Cc, Dd, Ee, Ff, Gg, Hh, Ii, Jj, Kk, Ll, Mm, Nn, Oo, Pp, Qq, Rr, Ss, Tt, Uu, Vv, Ww, Xx, Yy, Zz.*** It is the basis of the writing of Romanic, Germanic and many other languages.

In the process of its adaptation to the phonetic systems of some languages there was a problem of indication of sounds which had no corresponding letters in the Latin alphabet. There are the following methods of adaptation:

1. Addition of the new letters to the alphabet, for example: ***ŋ*, ç, ş**.

2. The use of diacritical marks, for example: ***á, ä, ğ***.

3. The use of negative diacritics, such as the Turkish ***I*** («***ы***» sound in Kazakh language) – without a point ***i***.

4. The use of a number of letters to write one sound (examples: "*sh" in english*, "*sch" in german* to indicate the Kazakh sound «*ш*», "*ch" in English* and "*tsch" in German to indicate the Kazakh sound «ч»*, etc.).

5. The use of a number of letters and diacritical mark to indicate one sound, for example, "*c'h"* in modern Breton is «*х*» in Kazakh.

6. Change of the value of one or more letters, for example letter "*x"* indicates:

1) Kazakh sound «*ш*» *in* Portugal;
2) Kazakh sound «*ы*» In Polish.

3. The experience of the Turkic-speaking countries

Before discussing application of these methods to the conversion of Kazakh writing to the Latin alphabet, we will consider the experience of Turkic-speaking countries (Turkey, Turkmenistan, Uzbekistan, Azerbaijan) which already use Latin alphabet.

The Turkish alphabet, which was accepted in 1928, consists of **29** letters and its **6** letters (**ğ**, **ü**, **ş**, **ı**, **ö**, **ç**) are not in the classic Latin alphabet. The Turkmenian alphabet, which has changed several times in 1990s, had **30** letters, and **8** of them (**ç**, **ä**, **ž**, **ň**, **ö**, **ş**, **ü**, **ý**) were not in the classic Latin alphabet. The Uzbek alphabet in 1993 had **6** letters out of the classic Latin alphabet. They were **ç**, **ş**, **ğ**, **ö**, **ñ** and **j**. Later, in 1995, they were removed from the alphabet and orthographic rules were rewritten: the previous letters were substituted to **ch, sh, g', o'**, **ng** and **j** respectively. Aksant aigues (apostrophe) ' before the sign is presented in this way **‘**, and after a sign it is written as **’**, for example, **‘**alphabet - алфавит - әліпби**’**. The Azerbaijan alphabet of the year of 2004 consists of **32** letters, **8** letters in it (**ç**, **ə**, **ğ**, **ı**, **İ**, **ö**, **ş**, **ü**) are not in the classic Latin alphabet.

Among these alphabets only the alphabet of the Uzbek language does not require additional fonts, drivers, sorting and searching program utilities to work with computers and telephones.

4. Stages and problems of Kazakh language writing

The writings of Kazakh language come from ancient Turkic rune inscriptions. Arabic alphabet in Turkic languages began being used since the second half of the VIII century. It came together with the Islam.

In the XX century Kazakh language alphabet had changed three times. In 1912 the great Kazakh enlightener Akhmet Baytursynov worked with the voice system of Kazakh language and on the basis of the Arabic graphics he created a new alphabet of 28 letters and defined the rules of direct writing. This alphabet was used in our country till 1929. Kazakhs, who live in other countries, (for example, Afghanistan, Iran, China, Pakistan) still use this alphabet.

In 1929 the Kazakh language writing passed to 29 letter alphabet, based on the Latin graphic. The new sound "**хы**" was added and it was indicated by the Latin letter **h**.

In 1940 the Kazakh writing passed to 42 letter alphabet, based on the Cyrillic graphic. To indicate specific Kazakh sounds 9 letters (**ә, ғ, қ, ң, ө, ұ, ү, һ**), which are out of Cyrillic alphabet, were added to the 33 letters of Russian language alphabet.

The last reform was carried out without taking into account the phonetic features of Kazakh language: the sounds of Russian language were unnecessary added. Some extra letters for their indication were included in Kazakh alphabet (**в, ё, и, ц, щ, ф, э, ю, я, ь, ъ**). The new requirements in the Kazakh orthography (spelling rules), and a pronunciation (pronunciation rules) appeared. In Kazakh language spelling and pronunciation of borrowed words from Russian must be in accordance with the rules of Russian language. Some believe that such innovation develops Kazakh language (imagine that English sounds are added to Russian language and to indicate them the appropriate letters are included in the Russian alphabet and make Russian speaking people write and pronounce borrowed English words in the Russian language in accordance with the norms of the English language. Does it “develop” Russian language?). Many contradictions in orthography and orthoepy appeared as a result of this "development" of the Kazakh language. Language has become deformed. It started losing its internal unity and sounding. The Russian accent appeared in Kazakh speech. Textbooks and scientific works devoted to Kazakh language are contradictory to each other and still are published.

(Appendix A contains only Kazakh language schoolbooks and Kazakh language textbooks where the voice system of Kazakh language is given differently).

Let’s have a look at the use of the Russian vowel sounds embedded into Kazakh language "**и**" and "**у**" in order to show appeared contradictions in Kazakh language. For example, in a modern writing such words as **"би, "ми", "бу", "су"** end with the sounds "**и**" and "**у**". In the third person attractive completion form we write **"би+і", "ми+ы", "бу+ы", "су+ы".** This is contrary with the rule: "In Kazakh language in the third person "**і**" or "**ы**" attractive endings are attached to words which end with the consonants, and "**сі**" or "**сы**" are attached to the words which end with vowels. If in Kazakh language sounds "**и**" and "**у**" will be considered as consonants, then in words "**би**", "**ми**", "**бу**", "**су**", "**ту**" there will not be any syllable. And we have 2 questions from that:

1. “Are the rules about endings correct in Kazakh language?”
2. “Are there any words without syllables in Kazakh?”

To make these rules correct these words must be rewritten with the use of Kazakh sounds: "бій", "мый", "бұу", "сұу", "тұу".

In such a state Kazakh language does not develop as the official language of Republic of Kazakhstan. This state destroys Kazakh language. So there is no doubt that new reform in Kazakh language is needed. If the process of reform will not start now, Kazakh language will transform in some kind of hybrid language and lose its natural features.

Necessity of the realization of reform on the basis of the voice system of Kazakh language by transformation to Latin alphabet, which makes it easy to use and create information technologies effectively, is obvious. During its realization it is necessary to renew orthographic, orthoepic, morphological, syntactic rules and to work out technologies for processing them in a computer and prepare schoolbooks and textbooks for all education levels.

5. Sound system of Kazakh language

There are **28** native sounds in Kazakh language: 9 of them are vowels and 19 are consonants. In **1929** during reform the new vowel "**хы**" was added and we had 29 sounds. In the alphabet used nowadays they are determined as follows: **а, ә, е, о, ө, ұ, ү, ы, і** - vowels; **б, ғ, г, д, ж, з, й, қ, к, л, м, н, ң, п, р, с, т, у, ш, h** −consonants (**а, о, ұ, ы, е** are phonemes, and **ә, ө, ү, і** – their allophones; **ғ-г** and **қ-к** are types of consonance of one phoneme). And in 1940 during reform 11 sounds of Russian language were added to Kazakh language. In the applied alphabet they are indicated by **в, ё, и, ц, ч, щ, ф, х, э, ю, я**. Let's take a closer look at the sounds **в, х** and **ф**, which do not violate phonetic patterns of Kazakh language. First of them is widely used in writing of the last name of a person. In spite of the fact that at writing of the last name of all Kazakhs in Kazakh language (without using the sound **в**) a problem of the correct writing of the last names of representatives of other nationalities in official documents in official language still takes place. Besides, if we take into account importance of numerous terms, such as «*валюта, вакуум*, *вакцина*, *вариант*, *ватт, вектор*, *вексель, вето, викторина, вирус, виртуал, вице, вокал, вольт*, хадис, *хаки, халат, хаос, химия, хлор, хор, хром, хроника, хрусталь, хунта*, *факт, факультет, фаза, файл, фауна, федерация, фельетон, физика, филармония, фильм, фонетика, формула, форфор, фосфор, фотон, фракция, функция*», it is better to leave embedded sounds **в, х, ф** in the structure of the voice system of Kazakh language. Kazakh language will not suffer from this; on the contrary, these sounds will help to pronounce the international terms correctly. Thus, we can assume that there will be 31 sounds in the sound system of the Kazakh language. These are: ***а, ә, б, в, ғ, г, д, ж, з, е, й, к, қ, л, м, н, ң, о, ө, п, р, с, т, у, ұ, ү, ф, х, ш, ы, і.***

6. Criterion of determination of Kazakh alphabet

In accordance with the status of official language the electronic resources, created in the country, must be in Kazakh language. Creation of informative resources is a process that requires a lot of finances and work, so their creation with the use of 26 letter classic Latin alphabet, which all the keyboards of all the computers have, will be more effective. Because in order to create electronic resources in other national alphabets we need to create additional fonts, drivers, sorting and searching programs, which do not develop without additional financial expenses.

Considering all these arguments it is possible to offer the following criterion of convertion of Kazakh alphabet to Latin alphabet:

1. New alphabet should be developed only on the basis of a sound system of Kazakh language. This requires Russian language sounds which conflict with the phonetics of Kazakh language to be removed.

2. New alphabet must use only the classical Latin alphabet letters, which are available in a standard computer keyboard. It will make written communication in Kazakh language by computer and telephone much easier without setting fonts, drivers, software, sorting and searching programs.

3. New alphabet will change the values of some Latin letters to identify some sounds of Kazakh language. This requires some Latin letters to be voiced in Kazakh and in this way they will be treated as the letters of Kazakh alphabet.

4. To indicate some Kazakh language sounds Latin letter combinations will be used. This is required because of lack of some Latin letters to indicate some of Kazakh Latin sounds. The combinations of letters, which do not appear together in the simple Kazakh word, must be used and the appropriate spelling rules must be developed.

5. New alphabet will have the sequence of the letters, which is fully identical to the sequence of letters in a classic alphabet. Sorting and searching programs, which are available in all system programs, can be used in the creation or use of information resources in Kazakh language without any damage and delay.

7. Tasks at the construction of new alphabet

Let's have a look at the indication of Kazakh sounds with the letters of classical alphabet in

accordance with the above criteria. We have **26** letters in the classical Latin alphabet and **31** sounds in Kazakh language (**28** natural sounds, **3** embedded sounds). Therefore to determine the alphabet, which does not extend the length of Kazakh words, we should solve following three tasks:

1) Identify the sounds indicated only by a single letter;
2) determine the sounds, indicated by the letter combnations;
3) Identify the specific pair combinations.

In order to solve these problems, we have investigated Kazakh sound system using a computer and have analyzed the types and properties of sounds. By constructing mathematical models of vowel sounds we proved that phonemes and their allophones have the following relationship: **а+е=ә, о+е=ө, ұ+е=ү, ы+е=і** (Mathematical model and oscillograms of vowel sounds of the Kazakh language are given in the Appendixes b.a and b.b). These relationships show that two letter combinations can be used to indicate allophones. Such a method is also used in other languages. For example, in English the sound **ш** is indicated by a combination of letters **s** and **h**. In addition, we have built a body of Kazakh language, which contains 100 million letters of our current alphabet and done a research about the frequency of letters and letter combinations with the help of computer programs (The frequencies of letters in the text body of Kazakh language are given in Appendix c).

The results of these studies were needed to determine which letter and which letter combination can efficiently indicate sound (not extending the length of the words). The results also, before the conversion of Kazakh writing to the new alphabet, will help to create a new spelling dictionary without the letters **ё, и, ц, ч, щ, ь, ъ, э, ю, я**, which will be excluded from the current alphabet (Appendix d shows examples of terms written in only Kazakh letters of the current alphabet).

8. Construction of the new Kazakh alphabet

Indication of Kazakh vowels **а, о, е, ы** by Latin letters **a, o, e, y**, and of consonants **б, г, д, з, л, м, н, п, р, с, т** by Latin letters **b, g , d, z, l, m, n, p, r, s, t** does not cause any problem.

Among the unmarked allophones **ә, ө, і** sound frequency of sound **і** is the highest and this sound, while participating in suffixes and endings, may occur in a speech several times (for example, in the word ‘біліктіліктің' there are 5 **і** sounds). Therefore, despite the fact that this sound is an allophone, it will be indicated by a Latin letter **i**. And allophones **ә** and **ө** do not meet often and they do not participate in any suffixes and endings, i.e. they are used only in root words and do not participate in the composition of the derivative words. If we will identify them by letter combinations, writings of the words will not be too long. As a first combination pair let’s take a Latin letters **a** and **o**, which indicate phonemes **а** and **о**, which are similar with **ә** and **ө** in soundings. As the second pair combination let’s take Latin letter **e**, which indicates the intermediate phoneme from the relationship **а + е = ә** and **о + е = ө**, i.e. **ә** and **ө** allophones will be marked by the **ae** and **oe** Latin letter combinations respectively.

Remarks:

1. In our investigation of Kazakh text body there are no simple words, which have **ae** and **oe** Cyrillic letter combinations. Besides, our articulation abilities do not allow pronouncing these sound combinations. Therefore in the new alphabet Latin letter combinations **ae** and **oe** do not generate any orthographic problem.
2. If such letter combination is found in compound words in accordance with the existing orthography, it is only between the constituent words and therefore they have to be written separately. For example, the components of the word "қараемел" will be written separately as "қара емел".
3. If such letter combination is found in the record of someone’s last name in accordance with the currently existing orthography, then we need to rewrite it by placing letter **й** between the letters of such letter combinations, because we hear the sound **й** when we pronounce last names with such letter combinations in it. (e. g. Абаев=Абайев; Оразаев=Оразайев, Ризоев=Ризойев, Сусоев=Сусойев)

In the new alphabet high-frequency consonant sounds **ғ** and **г** will be indicated by nearly standing letters **f** and **g**, and consonant sounds **қ** and **к** will be indicated by the letters **q** and **k,** because it seems like the word ‘қазақ’ is correctly spelled as 'qazaq' rather than a ' kazak''.

Now let’s have a very close look at the sonar consonant sound **ң** (**ың**). There is no word in the Kazakh language that starts with this sound, i.e. the use of the capital letter that will indicate this sound will be very rare. Therefore, paying more attention to the lowercase letter that indicates the sound **ң**, we will choose Latin letter **h**, which is similar in shape to the Latin letter **n**, which indicates sonar sound **н** that is in tune with **ң** sound (In some languages sound **ң** is indicated by a combination of two Latin letters. For example, in English and Uzbek languages sound **ң** is indicated by the combination of letters **n** and **g**. This method cannot be used in Kazakh language, because some different words in the new alphabet will be written identically and pronounced differently: different words **" күнгі, күңі"** are written as **"kungi"**) .

Let’s change the values of the remaining Latin letters **c** and **x** to the unmarked Kazakh sounds **ш(шы)** and **ж(жы)**respectively (here we took into account that the letter **C** was used in the same value in the Kazakh alphabet in 1929).

Another unmarked borrowed low frequency sounds **ф(фе)** and **х(хы)** will be indicated by the combinations of two Latin letters. We will take Latin letters **p** and **k** as a first pair combination respectively, which indicate sounds **п** and **к** that are close to **ф** and **х** sounds in soundings. As a second pair of these combinations we will take Latin letter **h**, which indicates sonar sound **ң** (**ың**). Because in the investigated body text we did not find any simple word, where there are **пң** and **кң** letter combinations. This means that the Latin letter combinations **ph** and **kh** will not generate any spelling problems in the new alphabet.

Further let’s consider 3 Kazakh similar-sounding sounds: vowels **Y** and **ұ** and consonants **y**. Note that in the body text frequencies of letters **Y, ұ, y** were closely to be alike, but frequencies of letters **ү** and **ұ** will increase during the creation of new orthography of Kazakh language. It is because if the sound occurs in the beginning or in the middle of a word before consonants, then according to the softness or hardness of the nearest syllable it is replaced by the sound of **Y** or **ұ** (e.g. улеу=үлеу, умаждау=ұмаждау, удмурт=ұдмұрт, сурет=сүрет, формула=формұла) . Also, if the sound occurs in the end of a word after the consonants, it is influenced by the softness and hardness of the nearest syllable and is replaced by a combination of үу or ұу (e.g. келу=келүу, бару=барұу). In spite of the fact that ү is the allophone, we must indicate it by the special Latin letter. To indicate sounds **y, ү and ұ** we have three appropriate letters **u, v and w**. In Latin alphabet they stand in sequence. Therefore we will indicate Kazakh sounds **y, ү, ұ** by the Latin letters **u, v, and w** respectively. We will indicate the low frequency sound **в (вы)** by a combination of two Latin letters. As the first pair of combination let’s take the letter **b** that indicates close sounding sound **б (бы)** and as the second pair we will take Latin letter **h.** There is no letter combination **бң** found in the ivestigated body text.

So we have built the first draft of the proposed alphabet and in the alternative to him second draft the sound **в (вы)** is indicated by the one letter **v** and the sound **ү** is indicated by the **ue** letter combination. The rest of Latin letters that indicate Kazakh sounds do not change (in Appendix i presentation of notation of Kazakh sounds in Latin alphabet , Appendix j.a shows the Latin alphabet of the Kazakh language , and in Appendix j.b orthographic rules of the submission of Kazakh sounds are given) .

In the new Latin alphabet of Kazakh language built above you can type it on the keyboard of any computer and telephone without any additional costs. There is no need to switch to the other tabs and you can process Kazakh text efficiently without using additional fonts, drivers, software, sorting and searching programs. (in Appendix k examples of the application of the new alphabet are shown) .

9. Future challenges

Before transforming the Cyrillic alphabet Kazakh writing into Latin alphabet we should get rid of all letters that indicate sounds, which do not belong to Kazakh language. There are no letters and

orthographic rules for them in the offered new alphabet.This is due to the fact that no language takes sounds of another language without legitimizing them and words that are formed with the use of the sounds of another language, which subordinate these words according to their own phonetic rules, create appropriate orthographic and orthoepic rules. For example, in Russian , some people's names and the names of the places and water bodies are not written in Kazakh. They were converted in Russian style. As the result Әділ became Адиль, Іңкәр became Инкар,Ұзынағаш became Узунагач, Іле became Или. This method is used in any languages with different sound systems and alphabets. Therefore, we have developed some rules how to remove Cyrillic alphabetic Russian letters ё, и, э, ю, я , ц, ч, щ, ь, ъ from Kazakh language words. The automatic transformation of Kazakh texts from Cyrillic into Latin alphabet computer converter program was created. Its internet address: www.alphabet.kz.

After the approval of the new Kazakh alphabet we need to work in the following 7 directions, each of them consists of 7 issues:

I. Standardization of Kazakh language:

I.1. Creating a standard phonetics of Kazakh language.
I.2. Creating a standard orthography of Kazakh language.
I.3. Creating a standard morphology of Kazakh language.
I.4. Creating a standard syntax of Kazakh language.
I.5. Creating industrial terms of Kazakh language.
I.6. Creating an onomastics standard and toponymy of Kazakh language.
I.7. Creating a standard of measurement of Kazakh language knowledge.

II. Electronic audio dictionaries of Kazakh language:

II.1.Orthographic dictionary of Kazakh language.
II.2.Orthoepic dictionary of Kazakh language.
II.3.Phraseological dictionary of Kazakh language.
II.4.Synonymic audio dictionary of Kazakh language.
II.5.Russian-Kazakh common language audio dictionary.
II.6.Diversified Russian-Kazakh audio glossary.
II.7.Diversified explanatory audio glossary of terms in Kazakh language.

III. Formalization of Kazakh language

III.1. Formalization of the phonetic rules of Kazakh language.
III.2. Formalization of the morphological rules of Kazakh language.
III.3. Formalization of the syntax rules of Kazakh language.
III.4. Formation of semantic models of word forms of Kazakh language.
III.5. Formation of semantic models of the word combinations of Kazakh language.
III.6. Formation of semantic models of the sentences of Kazakh language.
III.7. Formation of semantic models of the texts of Kazakh language.

IV. Computer processing of Kazakh language writing

IV.1. Creation of the converter programs for transforming of Kazakh language texts from one encoding system to another and from one drawing to another.
IV.2. Creation of a thematic electronic Kazakh language text fund on Arabic, Cyrillic and Latin.
IV.3. Creation of a computer program of analysis and synthesis of word forms of Kazakh language.
IV.3. Creation of a computer program of morphological analysis and synthesis of Kazakh language.
IV.4. Creation of a computer program of syntactic analysis and word combinations synthesis of Kazakh language.
IV.5. Creation of a computer program of syntactic analysis and synthesis of simple sentences of Kazakh language.
IV.6. Creation of a computer program of syntactic analysis and synthesis of complex sentences of Kazakh language.
IV.7. Creation of a system of semantic information search in Kazakh language.

V. Speech technologies of Kazakh language

V.1. Creation of phonetically complete acoustic enclosure based on the new orthography.

V.2. Creation of a program of generation of phonetically complete Kazakh word forms and their transcriptions.

V.3. Creation of program of synthesis of individual oral Kazakh words.

V.4. Creation of a program of synthesis of continuous speech of Kazakh language.

V.5. Creation of program of recognition of individual oral Kazakh words.

V.6. Creation of a program of recognition of continuous speech of Kazakh language.

V.7. Creation of a program of recognition of noises of Kazakh language.

VI. Electronic learning (e-learning) of Kazakh language:

VI.1. Creation of multimedia programs that train listening and writing sounds, syllables and words of Kazakh language.

VI.2. Creation of a multimedia program to train listening and writing sentences of Kazakh language.

VI.3.Creation of an intellectual web-system that will train using of Kazakh language dialogues in different situations (at home, at work, on way).

VI.4. Creation of the multi-media office training programs in Kazakh language.

VI.5. Creation of the multi-media Civil Code training programs in Kazakh language.

VI.6. Creation of an intellectual civil servants training web-system in Kazakh language.

VI.7. Creation of an educational Web-portal in Kazakh language.

VII. Knowledge certification of Kazakh language

VII.1. Creation of beginning level of knowledge of Kazakh language determining intellectual system.

VII.2. Creation of simple level of knowledge of Kazakh language determining intellectual system.

VII.3. Creation of basic level of knowledge of Kazakh language determining intellectual system.

VII.4. Creation of middle level of knowledge of Kazakh language determining intellectual system.

VII.5. Creation of good level of knowledge of Kazakh language determining intellectual system.

VII.6. Creation of high level of knowledge of Kazakh language determining intellectual system.

VII.7. Creation of fluent level of knowledge of Kazakh language determining intellectual system

10. Performing activities

1.To determine the type and number of sounds in Kazakh language and to approve its new Latin alphabet. Activity must be performed before 2013.

2.To develop the 100,000 units spelling dictionary with the use of Cyrillic alphabet. There must be no letters which indicate non-Kazakh sounds. Activity must be performed before 2013.

3.To accept a state program "Scientific, normative, technological and methodical bases of the transformation of the Kazakh writing to the Latin alphabet". Activity must be performed before 2013.

4. The national company, in which 100% of its share capital is held by the country, must be created in order to reach the complete and qualitative transformation of the Kazakh writing into the Latin alphabet and efficient use of allocated funding. Activity must be performed before 2013.

Appendix a. Turkish alphabet, 1928

№	Letters	Transcription	Comparision
1	A a	/a/	'a' as in father
2	B b	/b/	'b' as in book
3	C c	/dʒ/	'j' as in Joke
4	Ç ç	/tʃ/	'ch' as in chimpanzee
5	D d	/d/	'd' as in day
6	E e	/e/, /ɛ/	'e' as in red or 'a' as in cat

7	F f	/f/	'f' as in far
8	G g	/g/, /ɟ/	'g' as in game
9	Ğ ğ	/ɣ/1	No equivalent, watch video below
10	H h	/h/	'h' as in hot
11	I ı	/ɯ/	'e' as in open
12	İ i	/i/	'i' as in machine
13	J j	/ʒ/	's' as in pleasure
14	K k	/k/, /c/	'k' as in kilo
15	L l	/l/, /ɬ/	'l' as in life
16	M m	/m/	'm' as in master
17	N n	/n/	'n' as in nice
18	O o	/o/	'o' as in more
19	Ö ö	/ø/	'u' as in turn
20	P p	/p/	'p' as in spin
21	R r	/ɾ/	the 'r' as in car
22	S s	/s/	's' as in smile
23	Ş ş	/ʃ/	'sh' as in shine
24	T t	/t/	't' as in stop
25	U u	/u/	'u' as in ultimate
26	Ü ü	/y/	'u' as in cube
27	V v	/v/	'v' as in victory
28	Y y	/j/	'y' as in you
29	Z z	/z/	'z' as in zigzag

Appendix b. Turkmenian alphabet, 1991

№	Cyrillic	Latin	Name	Transcription
1	Аа	A a	a	/a/
2	Бб	B b	be	/b/ — /-ß-/
3	Чч	Ç ç	çe	/ç/
4	Дд	D d	de	/d/
5	Ее	E e	e	/e/
6	Әә	Ä ä	ä	/ä/
7	Фф	F f	fe	/f/
8	Гг	G g	ge	/g-/ — /-γ-/
9	Хх	H h	he	/h/
10	Ии	I i	i	/i/
11	Ҷҷ	J j	dƶe	/dʒ/
12	Жж	Ž ž	že	/ʃ/ö
13	Кк	K k	ka	/q/ — /k/
14	Лл	L l	el	/l/
15	Мм	M m	em	/m/
16	Нн	N n	en	/n/
17	Ңң	Ň ň	eň	/ň/
18	Оо	O o	o	/o/
19	Өө	Ö ö	ö	/ö/
20	Пп	P p	pe	/p/
21	Рр	R r	er	/r/
22	Сс	S s	es	/θ/
23	Шш	Ş ş	şe	/ş/
24	Тт	T t	te	/t/

25	Уу	U u	u	/u/
26	Үү	Ü ü	ü	/ü/
27	Вв	W w	we	/w/ — /v/
28	Ыы	Y y	y	/y(ï)/
29	Йй	Ý ý	ýe	/ýe/
30	Зз	Z z	ze	/ ð/

Appendix c. Uzbek Alphabet, 1995

№	Cyrillic	Latin	Name	Transcription
1	A a	A a	a	[a, æ]
2	B b	B b	be	[b]
3	D d	D d	de	[ɖ]
4	E e	E e	e	[e]
5	F f	F f	ef	[ɸ]
6	G g	G g	ge	[g]
7	H h	H h	he	[h]
8	I i	I i	i	[i, ɨ]
9	J j	J j	je	[ʒ]
10	K k	K k	ke	[k]
11	L l	L l	el	[l]
12	M m	M m	em	[m]
13	N n	N n	en	[n]
14	O o	O o	o	[o]
15	P p	P p	pe	[p]
16	Q q	Q q	qe	[q]
17	R r	R r	er	[r]
18	S s	S s	es	[s]
19	T t	T t	te	[ʈ]
20	U u	U u	u	[u, y]
21	V v	V v	ve	[ʋ]
22	X x	X x	xe	[χ]
23	Y y	Y y	ye	[j]
24	Z z	Z z	ze	[z]
25	O’ o’	O’ o’	o’	[o, ø, ʏ]
26	G’ g’	G’ g’	g’e	[ʁ]
27	Sh sh	Sh sh	she	[ʃ]
28	Ch ch	Ch ch	che	[ʧ]
29	’	’	apostrof	[ʔ]

Appendix d. Azerbaijan alphabet

№	Latin	Name	Transcription
1	A a	a	[ɑ:~a]
2	B b	be	[b]
3	C c	ce	[ʤ]
4	Ç ç	çe	[ʧ]
5	D d	de	[d]
6	E e	e	[ɛ~e]
7	Ə ə	ə	[æ]

8	F f	fe	[f]
9	G g	ge	[g]
10	Ğ ğ	ğe	[ɣ]
11	H h	he	[h]
12	X x	xe	[x]
13	I ı	ı	[ɯ~ə]
14	İ i	i	[ɪ~i]
15	J j	je	[ʒ]
16	K k	ke, ka	[kʲ, k]
17	Q q	qe	[g]
18	L l	el	[l]
19	M m	em	[m]
20	N n	en	[n]
21	O o	o	[ɒ~o]
22	Ö ö	ö	[ɜ:~ø]
23	P p	pe	[p]
24	R r	er	[r]
25	S s	se	[s]
26	Ş ş	şe	[ʃ]
27	T t	te	[t]
28	U u	u	[ʊ~u]
29	Ü ü	ü	[ʏ:~ʉ:]
30	V v	ve	[v]
31	Y y	ye	[j]
32	Z z	ze	[z]

Appendix e. Textbooks and schoolbooks of Kazakh language

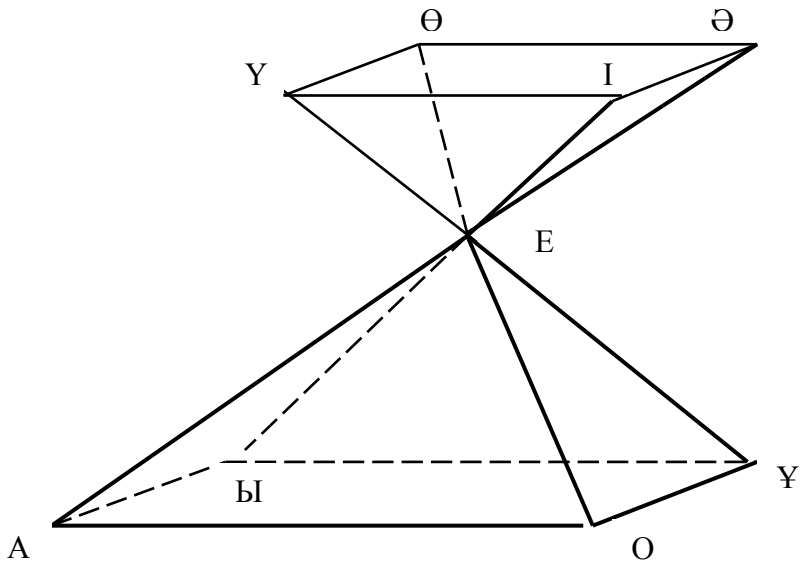
№	Authors	Title of work	Vowels, amount, form	Consonant, amount, form
1.	А. Байтұрсынов	Тіл тағылымы. Алматы: Ана тілі, 1992. –448 б.	5 дыбыс ***а, е, о, у, ы*** 2 жарты дауысты ***у, й***	17 дыбыс ***б, г, ғ, д, ж, з, к, қ, л, м, н, ң, п, р, с, т, ш***
2.	Қ. Жұбанов	Қазақ тілі жөніндегі зерттеулер. Алматы: Ғылым, 1966. -362 б.	7 дыбыс ***а, е, о, у, ы, ұу, ый***	12 дыбыс ***б, г, ғ, д, ж, з, к, қ, п, с, т, ш, л, н, ң, м, у, й, р***
3.	І. Кеңесбаев	Қазақ тіл білімі туралы зерттеулер. Алматы: Ғылым, 1987. -352 б.	11 дыбыс монофтонг ***а, ә, е, о, ө, ұ, ү, ы, і, э,*** 2 дифтонг ***и, у***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, (у), ф, х, һ, ц, ч, ш***
4.	Ж. Аралбаев	Қазақ фонетикасы бойынша этюдтер. Алматы: Ғылым, 1988.-144б	11 дыбыс монофтонг ***а, ә, е, о, ө, ы, і, ұ, ү, э*** 2 дифтонг ***и, у***	25 дыбыс ***б, в, г, ғ, д, ж***(дж), ***з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ, ц, ч, ш***
5.	М.Қараев	Қазақ тілі. Алматы:Ана тілі, 1993. -216 б.	11 дыбыс монофтонг ***а, е, ы, і, ә, о, ө, ұ, ү*** дифтонгоид ***и, у***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, л, м, н, п, р, с, т, (у), һ, ф, х, ц,ч, ш, қ, ң***
6.	Н.Оралбаева, Т.Әбдіғалиева, Б.Шалабаев	Практикалық қазақ тілі. Алматы: Ана тілі, 1993. -272 б.	12 дыбыс ***а, ә, е, о, ө, ұ, ү, ы, і, э***	24 дыбыс ***б, в, г, ғ, д, ж, з***

			дифтонг ***и, у***	***й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ц, ч, щ***
7.	С. Мырзабеков	Қазақ тілі фонетикасы. Алматы: Қазақ университеті, 1993. - 136 б.	11 дыбыс ***а, ә, о, ө, ұ, ү,ы, і, е, у, и***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, л, м, н, қ, ң, п, р, с, т, у, һ, ф, х, ц, ч, ш***
8.	Ғ.Әбуханов	Қазақ тілі. (Пед. училищеге арналған оқулық). Алматы: Мектеп, 1982. -284 б.	12 дыбыс ***а, ә, е, и, о, ө,у, ұ, ү, ы, і, э***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ц, ч, ш, щ***
9.	Н.Оралбаева, Ғ.Мадина, А.Әбілқаев	Қазақ тілі. (Филология мамандығына арналған оқулық). Алматы: Мектеп,1982. -296 б.	12 дыбыс ***а, ә, е, и, о, ө,у, ұ, ү, ы, і, э***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ, ц, ч, ш, щ***
10.	С.Рахметова	Қазақ тілін оқыту методикасы (Педагогика училище арналған оқулық). Алматы: Ана тілі, 1991. -184 б.	13 дыбыс ***а, ә, е, ё, и, о, ө,у, ұ, ү, ы, і, э***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ц, ч, ш, щ***
11.	С.М.Исаев	Қазақ тілі. Алматы: Қайнар,1993. -170 б.	12 дыбыс ***а, ә, е, и, о, ө, у, ұ, ү, ы, і, э***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ,ц,ч, ш, щ***
12.	С.Исаев, К.Назарғалиева	Қазақ тілі. (7-сыныпқа арналған оқулық). Алматы: Рауан, 1997. -192 б.	12 дыбыс ***а, ә, е, о, ө, ұ, ү, ы, і, э, и, у***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ, ц, ч, ш, щ***
13.	Жалпы Редакциялаған А.Сейдімбек	Қазақ тілі. (әмбебап анықтамалық) Алматы: Болашақ балапандары, 1999. –150 б.	13 дыбыс ***а, ә, е, и, о, ө, ұ, ү, ы, і, ю, я***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, ц, ч, ш, щ, һ***
14.	И.Маманов	Қазақ тілі. Алматы: Мектеп, 1989. –176 б.	15 дыбыс ***а, ә, е, ё, и, о, ө,у, ұ, ү, ы, і, э, ю, я***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ц, ч, ш, щ***
15.	Ш.Бектұров, М.Серғалиев	Қазақ тілі. Алматы: Білім, 1994. –224 б.	15 дыбыс монофтонг ***а, ә, е, о, ө, ұ, ү, ы, і*** дифтонг ***и, у,*** орыс дыбыстары ***э, е, ю, я***	25 дыбыс б***, в, г, ғ, д, ж, з, й, к, л, м, н, ң, п, р, с, т, (у), һ, ф, х, ц, ч, ш, щ***
16.	К.Сариева	Қазақ тілі. Оқу жаттығу құралы. Алматы: Білім, 2000. –136 б.	15 дыбыс ***а, ә, е, е, и, о, ө, у, ұ, ү, ы, і, э, ю, я***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р,***

				с, т, ф, х, һ, ц, ч, ш, щ
17.	Р.Әміров, А.Бәкірова	Қазақ тілі. (1-сыныпқа арналған оқулық). Алматы: Мектеп, 1990. -158 б.	15 дыбыс ***а, ә, е, е, и, о, ө,у, ұ, ү, ы, і, э, ю, я***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ц, ч, ш, щ***
18.	Ш.Әуелбаев, Ә.Наурызбаева Р.Ізғұттынова, А.Құлжанова	Әліппе. Алматы: Ата-мұра,1997-160 б.	15 дыбыс ***а, у, о, ы, ұ, е, і, ә, и, ө, ү, я, ю, ё, э***	25 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ш, щ, ц, ч***
19.	Ш.Әуелбаев, Ә. Наурызбаева, Р.Ізғұттынова, А.Құлжанова	Әліппе. Алматы: Ата-мұра, 2002. -144 б.	15 дыбыс ***а, ә, е, ё, и, о, ө, у, ұ, ү, ы, і, э, ю, я***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ, ц, ч, ш, щ***
20.	А.Ысқақов, К.Аханов, Б.Кәтенбаева	Қазақ тілі. (5-сыныпқа арналған оқулық. 18-ші басылым). Алматы: Мектеп, 1989.-192 б.	12 дыбыс ***а, ә, и, е, э, о, ө, ү, ұ, (у), ы, і***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, һ, ш, щ, ц, ч***
21.	Г.Қосымова, Ж.Даулетбекова	Қазақ тілі (5-сыныпқа арналған оқулық). Алматы: Ата-мұра, 2005.-192 б.	12 дыбыс ***а, ә, е, и, о, ө, у, ұ, ү, ы, і, э***	26 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, у, ф, х, ц, ч, ш, щ һ, (у)***
22.	С.Аманжолов, Ә.Хасенов, И.Ұйықбаев	Қазақ тілі. (9-сыныпқа арналған оқулық. 20-басылым). Алматы: Рауан, 1996.-126 б.	***12 дыбыс*** ***а, ә, э, е, и, о, ө, у, ұ, ү, ы, і***	24 дыбыс ***б, в, г, ғ, д, ж, з, й, к, қ, л, м, н, ң, п, р, с, т, ф, х, һ, ш, ц, ч***

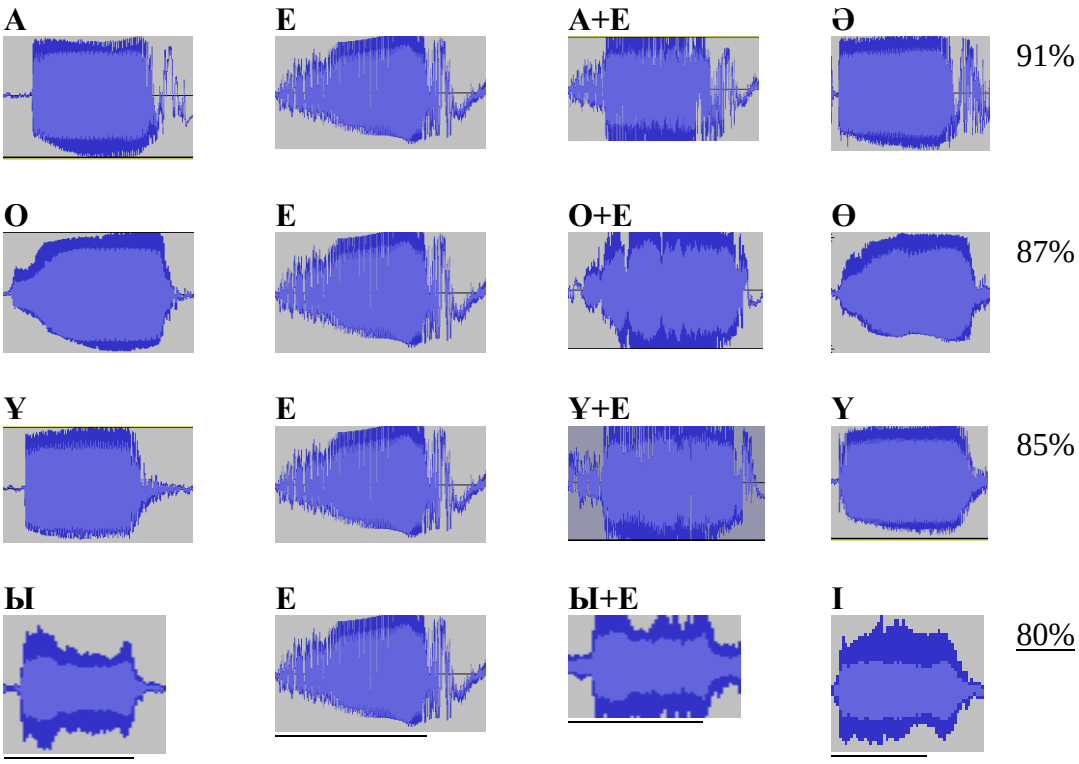
Here you can see that in the last four books during passing to the older grades the number of Kazakh language vowels reduces: in one ABC book for 1th grade pupils there are 15 vowels + 25 consonants = 40 sounds, in the next 15 vowels + 26 consonants = 41 sounds, in the book for 5th grade 12 vowels + 26 consonants = 38 sounds, and in the schoolbook for 9th grade pupils there are 12 vowels and 24 consonants = 36 sounds

Appendix f. a. Models of vowel sounds of Kazakh language



АОҰЫ – жуан (твердые, hard), **ӘӨҮІ** – жіңішке (мягкие, soft), **АОӘӨ** – ашық (открытые, open), **ЫҰІҮ** – қысаң (закрытые, closed), **ОҰӨҮ** – еріндік (губные, labial), **АЫІӘІ** – езулік (негубные, non-labial), **Е** – аралық (промежуточный, transitional).

Appendix f. b. Kazakh vowels oscyllograms



Appendix g. Frequency of letters, %

Letter	Art. Lit.	Publicistics	Official	Science	Informal speech	Total
А	12,85	12,89	12,18	11,74	12,65	12,52
Ә	0,71	0,81	0,98	0,87	0,73	0,88
Б	2,87	2,42	2,14	2,21	2,69	2,31

В	0,15	0,29	0,25	0,33	0,22	0,26
Г	1,13	1,16	1,24	1,29	1,19	1,20
Ғ	1,74	1,64	1,42	1,51	1,59	1,54
Д	4,59	4,45	4,04	4,50	4,60	4,27
Е	8,00	7,45	8,10	7,98	8,11	7,82
Ё	0,00	0,00	0,00	0,00	0,00	0,00
Ж	1,84	1,67	1,94	1,57	1,64	1,81
З	1,58	1,62	1,32	1,42	1,74	1,47
И	0,79	1,55	1,44	1,92	1,25	1,45
Й	2,28	1,41	0,98	1,33	1,96	1,27
К	2,88	2,72	3,23	2,95	2,90	2,98
Қ	3,52	3,33	3,36	3,24	3,13	3,35
Л	4,52	5,16	5,41	5,13	4,62	5,23
М	2,90	3,22	3,69	3,19	3,39	3,42
Н	6,90	6,85	6,57	6,75	6,84	6,72
Ң	1,68	1,58	1,50	1,53	1,56	1,55
О	2,58	2,56	2,30	2,82	2,72	2,44
Ө	1,11	0,89	0,86	0,83	0,96	0,89
П	2,42	1,52	1,12	1,48	2,02	1,40
Р	5,10	5,54	5,78	5,77	5,35	5,62
С	4,03	4,29	4,31	3,99	4,14	4,27
Т	5,19	6,01	6,52	6,31	5,54	6,19
У	1,04	0,9	0,73	0,83	0,86	0,83
Ұ	1,18	1,71	2,78	1,97	1,42	2,18
Ү	1,08	0,73	0,59	0,71	0,89	0,69
Ф	0,04	0,12	0,13	0,20	0,17	0,12
Х	0,18	0,28	0,18	0,25	0,20	0,23
Һ	0,02	0,02	0,00	0,01	0,01	0,01
Ц	0,03	0,12	0,19	0,23	0,08	0,15
Ч	0,04	0,03	0,03	0,05	0,05	0,03
Ш	1,44	1,23	1,24	1,14	1,35	1,25
Щ	0,01	0,00	0,00	0,01	0,01	0,01
Ъ	0,01	0,01	0,03	0,02	0,00	0,02
Ы	7,36	7,82	7,89	7,65	7,23	7,81
І	5,84	5,41	4,90	5,41	5,64	5,20
Ь	0,06	0,07	0,07	0,11	0,11	0,07
Э	0,02	0,08	0,08	0,18	0,06	0,08
Ю	0,03	0,04	0,14	0,06	0,05	0,09
Я	0,26	0,40	0,34	0,51	0,33	0,37

Appendix h. Samples of orthography in current alphabet

Current orthography	**Proposed orthography**
Академия	Әкәдемійә

Артист	Әртіс
Банкет	Бәнкет
Банкир	Бәнкір
Бизнес	Бізнес
Буфер	Бүфер
Бюджет	Бүйджет
Бюллетень	Бүллетен
Вьетнам	Вйетнәм
Габардин	Гәбәрдін
Газет	Гәзет
Газик	Гәзік
Галерея	Гәлереә
Галифе	Гәліфе
Гамбит	Гәмбіт
Гарнитур	Гәрнітүр
Гастрит	Гәстріт
Гастроль	Гәстрөл
Генезис	Генезіс
Гепатит	Гепәтіт
Глюкоза	Глүкөз
Губерния	Гүбернійә
Дискрет	Діскірет
Дисплей	Діспілей
Императивті	Імперәтів
Империя	Імперійә
Импликант	Імпілікәнт
Императивті	Імперәтів
Инженер	Інженер
Индустрия	Індүстрійә
Институт	Інстітүт
Интеграл	Інтегрәл
Инфекция	Інфексіә
Комитет	Көмітет
Компьютер	Көмпүйтер
Максимум	Мәксімүм
Минимум	Мінімүм
Министр	Міністір
Натурал	Натұрал
Ноль	Нөл
Президент	Презідент
Премьер	Премійер
Республика	Респүблік
Ревизия	Ревізійә
Регламент	Регләмент
Синус	Сінүс
Синхрон	Сыйнхрон
Термин	Термін
Терминал	Термінәл
Университет	Үніверсітет
Факультет	Фәкүлтет

Фельетон	Фелійетон
Фетиш	Фетіш
Фетишизм	Фетішізм
Филиал	Філійәл
Фильм	Філім
Флюгер	Флүгер
Фольклер	Фөлкілер
Фортопьяно	Фортопыйано
Франция	Франсыйа
Француздар	Франсұз
Функция	Фұныксыйа
Фюзеляж	Фүзеләж
Чемпион	Шемпійөн
Эволюция	Евөлүтсійә
Экскаватор	Екіскәвәтөр
Экскурсия	Екіскүрсійә
Экспансия	Екіспәнсійә
Эксперимент	Екісперімент
Эксперт	Екісперт
Экспонент	Екіспөнент
Экспресс	Екіспресс
Экстракт	Екістрәкт
Экстремист	Екістреміст
Эмиграция	Емігрәсійә
Энцефалит	Енсефәліт
Эскиз	Ескіз
Этика	Етійкә
Эфир	Ефір

Appendix i. Indication of Kazakh letters in Latin alphabet

№	Alphabet 1940	Alphabet 1929	New alphabet, 1 project	New alphabet, 2 project
1.	А	A a	A a	A a
2.	Ә	Ə ə	Ae ae	Ae ae
3.	Б	B в	B в	B в
4.	В		Bh вh	V v
5.	Г	G g	G g	G g
6.	Ғ	Ġ ġ	F f	F f
7.	Д	D d	D d	D d
8.	Е	E e	E e	E e
9.	Ё			
10.	Ж	Č č	X x	X x
11.	З	Z z	Z z	Z z
12.	И			
13.	Й	J j	J j	J j
14.	К	K k	K k	K k
15.	Қ	Қ қ	Q q	Q q
16.	Л	L l	L l	L l
17.	М	M m	M m	M m

18.	Н	N n	N n	N n
19.	Ң	H ŋ	H h	H h
20.	О	O o	O o	O o
21.	Ө	Ө ө	Oe oe	Oe oe
22.	П	P p	P p	P p
23.	Р	R r	R r	R r
24.	С	S s	S s	S s
25.	Т	T t	T t	T t
26.	У	V v	W w	W w
27.	Ұ	U u	U u	U u
28.	Ү	Y y	V v	Ue ue
29.	Ф	F f	Ph ph	Ph ph
30.	Х	H h	Kh kh	Kh kh
31.	h			
32.	Ц			
33.	Ч			
34.	Ш	C c	C c	C c
35.	Щ			
36.	Ъ			
37.	Ы	Ь ь	Yy	Yy
38.	І	I i	I i	I i
39.	Ь			
40.	Э			
41.	Ю			
42.	Я			

Appendix j. a. Latin alphabet of Kazakh language

№	Letter	Name	Trancription	Clarification
1.	A a	(a)	**[ɑ]**	
2.	B в	(бы)	[**b**]	
3.	C c	(шы)	**[ʃ]**	New value
4.	D d	(ды)	**[d]**	
5.	E e	(е)	**[э]**	
6.	F f	(ғы)	**[ɣ]**	New value
7.	G g	(гі)	**[g]**	
8.	H h	(ың)	**[ŋ]**	New value
9.	I i	(і)	**[ɨ]**	
10.	J j	(ый, ій)	**[j]**	New value
11.	K k	(кі)	**[k]**	
12.	L l	(ыл)	**[l]**	
13.	M m	(мы)	**[m]**	
14.	N n	(ны)	**[n]**	
15.	O o	(о)	**[o]**	
16.	P p	(пы)	**[p]**	
17.	Q q	(қы)	**[q]**	
18.	R r	(ыр)	**[r]**	

19.	S s	(сы)	**[s]**	
20.	T t	(ты)	**[t]**	
21.	U u	(ұ)	**[u]**	
22.	V v	1-(ү), 2-(v)	**1- [ʉ], 2-[v]**	New value
23.	Ww	(ұу, үу)	**[w]**	
24.	X x	(жы)	**[ʒ]**	New value
25.	Y y	(ы)	**[ɯ]**	
26.	Z z	(зы)	**[ƶ]**	

Appendix j. b. Orthographic rules of sound indications

№	Sound	Name	Mark	Transcription
1-project	***ə***	(ə)	Ae	**[æ]**
	ɵ	(ө)	Oe	**[ɵ]**
	в	(вы)	Bh	**[v]**
	ф	(фе)	Ph	**[f]**
	х	(хы)	Kh	**[x]**
2-project	***ə***	(ə)	Ae	**[æ]**
	Ɵ	(ө)	Oe	**[ɵ]**
	Y	(ү)	ue	**[ʉ]**
	Ф	(фе)	ph	**[f]**
	X	(хы)	kh	**[x]**

Appendix k. Examples of the use of new alphabet

Word examples which contain sounds **ə, ɵ:**

әдіскер=aedisker, әзірлеу=aezirlew, әкімшілік=aekimcilik, әңгіме=aehgime, бәйге=baeyge, бәсеке=baeseke, өгейшілік=oegejcilik, өміршең=oemirceh, өңгерілген=oehgerilgen, өшіргіш=oecirgic, бөбек=boebek, әзәзіл=aezaezil, зәмзәм=zaemzaem, нәмәрт=naemaertt, тәбәрік=taebaerik, көзкөрген=koezkoergen,көзмөлшер=koezmoelcer, көккөл=koekkoel, көкжөтел=koekxoetel, көкөніс=koekoenis, көркемөнер=koerkemoener.

The National anthem of Republic of Kazakhstan

Current alphabet	Future alphabet
Алтын күн аспаны, Алтын дән даласы, Ерліктің дастаны, Еліме қарашы! Ежелден ер деген, Даңқымыз шықты ғой. Намысын бермеген, Қазағым мықты ғой ***Қайырмасы:*** Менің елім, менің елім, Гүлің болып егілемін, Жырың болып төгілемін, елім! Туған жерім менің - Қазақстаным! Ұрпаққа жол ашқан,	Altyn kvn (kuen) aspany, Altyn daen dalasy, Erliktih dastany, Elime qaracy! Exelden er degen, Dahqymyz cyqty foj. Namysyn bermegen, Qazafym myqty foj! ***Qajyrmasy:*** Menih elim, menih elim, Gvlih (Guelih) bolyp egilemin, Xyryh bolyp toegilemin, elim! Tufan xerim menih – Qazaqstanym! Urpaqqa xol acqan,

Кең байтақ жерім бар. Бірлігі жарасқан, Тәуелсіз елім бар. Қарсы алған уақытты, Мәңгілік досындай. Біздің ел бақытты, Біздің ел осындай!	Keh bajtaq xerim bar, Birligi xarasqan, Taewelsiz elim bar, Qarsy alfan waqytty, Maehgilik dosyndaj, Bizdih el baqytty, Bizdih el osyndaj!

КОМПЬЮТЕРЛІК ЖҮЙЕЛЕРДІ ҰЛТТЫҚ ЛОКАЛИЗАЦИЯЛАУ ЖӘНЕ ТЕРМИНОЛОГИЯ
НАЦИОНАЛЬНАЯ ЛОКАЛИЗАЦИЯ КОМПЬЮТЕРНЫХ СИСТЕМ И ТЕРМИНОЛОГИЯ
THE NATIONAL LOCALIZATION OF COMPUTER SYSTEMS AND TERMINOLOGY

Д.Ш. СУЛЕЙМАНОВ, А.Ф.ГАЛИМЯНОВ

НИИ "Прикладная Семиотика"
академии наук Республики Татарстан

СИСТЕМА ТАТАРСКИХ ТЕРМИНОВ В КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЯХ И ИНФОРМАТИКЕ[1]

Последние 15-20 лет ознаменовались появлением большого количества новых понятий и терминов в разных областях науки, образования и технологий. Это явление особенно остро ощущается в области вычислительной техники и инфокоммуникационных технологий.

Как известно, термины являются структурными элементами информации и составляют главное ее содержание. В отличие от многих других наук, термины компьютерной техники и информационных технологий являются наиболее быстро развивающимися и обновляющимися терминами. Очевидно, для развития и участия в информационном пространстве в качестве языка накопления и передачи информации любой язык должен иметь свою терминологическую систему, а не довольствоваться простым заимствованием терминов из английского и русского языков путем простой адаптации. Приведем обоснование данного тезиса для татарского языка.

Во-первых, в настоящее время компьютеры проникают повсеместно и становятся обычным незаменимым инструментом практически для всех, от домохозяйки до профессора. А вместе с компьютерами в татарский язык проникают новые чужие термины и понятия. Очевидно, что они своим звучанием и своими правилами изменения не обогащают язык, а наоборот, нарушают систему языка, приводят к тому, что существующие термины не подчиняются закономерностям татарского языка.

Во-вторых, очевидно, что татарский язык как один из двух государственных языков в Республике Татарстан, должен быть рабочим языком компьютеров. Для этого необходимо осуществить полную татарскую локализацию компьютерных систем и технологий, то есть необходимо создать татарскую систему терминов и разработать татарский интерфейс для широко используемых прикладных программ. Также общение через Интернет по-татарски является необходимым условием для развития прикладных возможностей языка. Таким образом, для расширения сферы применения татарского языка имеет большое значение сделать его языком общения с компьютером, создать в компьютере татарскую среду с татарской терминологией.

В третьих, как следует из исследований зарубежных и отечественных ученых [1, 2, 3], татарский язык основан на жесткой логике, закономерностях, его морфология и синтаксис обладают большой регулярностью, обладают богатыми возможностями для передачи фразы любой сложности и глубины и, в то же время, являются достаточно простыми по структуре. Это означает, что татарский язык, как один из тюркских языков, имеет большой потенциал, чтобы стать языком общения в компьютерных сетях, языком операционных систем, языком составления специальных программ для компьютеров, языком интеллектуальных технологий для накопления и обработки информации, языком искусственного интеллекта. Соответственно, актуальным является разработка терминов для новых систем и технологий на основе татарского языка, которые предоставят новые возможности для развития компьютерной науки и новых компьютерных систем.

В-четвертых, для развития языка необходимо, чтобы его терминологическая система включала термины по всем наукам, особенно, по прикладным. Причем, что особенно важно,

[1] Исследование выполнено в рамках научно-исследовательского проекта РФФИ («Разработка моделей и программных средств прикладного синтаксического анализа татарских предложений»), проект № 12-07-00735-а

возникает возможность использовать эти термины в научно-прикладных исследованиях и получать результаты, которые будут интересны не только для татарского языка и для татар, но и всему научному сообществу. А это, в свою очередь, работает на существование и развитие татарского языка.

1. Из истории татарской терминологии

Несмотря на то, что правила создания и применения татарских терминов имеют вековую историю [4], проблем в этой области меньше не становится. Меняется мир, рождаются новые науки, появляются новые понятия. Появляются новые термины, которые практически не осваиваются или слабо осваиваются татарским языком, что приводит к механическому наполнению татарских терминологических словарей заимствованиями без необходимой ассимиляции. Активность же татарских ученых и официальных властей в этом направлении оставляет желать лучшего. Особо следует отметить практически отсутствие упорядочивания правил словообразования, исходя из современных требований и контроля правильности применения их на практике.

Как известно, во многих случаях, создание новых терминов и понятий сильно зависит от политических условий и состояния общества. Например, вопросы татарской терминологии для разных областей широко освещаются в татарской прессе во времена возрождения 1905-07 и 1917 годах. Несмотря на это, первый официальный документ принимается лишь в 1920 году, во Всероссийской конференции восточных журналистов. В этом творческом процессе принимали участие и Научный центр во главе с Галимджаном Ибрагимовым, и различные организации, и общественность. В 1925 году была принята обновленная и дополненная правило-инструкция «О языке и наименованиях». После этого и до настоящего времени обновление терминов, изучение и уточнение правил и терминов нашли отражение всего лишь в двух официальных документах. Первый из них – “Основные принципы в деле упорядочивания, создания новых терминов и составления терминологических словарей сегодняшней терминологии татарского литературного языка и инструкция”. Данный документ обсуждался в заседаниях президиума Правительственной терминологической комиссии, был утвержден и издан в 1943 году. Второй документ – книжка объемом в 13 страниц, которую составили М.З. Закиев и И.М.Низамов в 1991 году и который был утвержден в 1995 году комитетом при Кабинете Министров Республики Татарстан по претворению в жизнь Закона Республики Татарстан “О языках народов Республики Татарстан” и издан в том же году. После принятия Государственной программы Республики Татарстан “По сохранению, изучению и развитию языков народов Республики Татарстан” в 1994 году работа по татарской терминологии активизировалась и были изданы десятки терминологических словарей. В их число входят солидные двуязычные, трехъязычные и толковые словари по математике, начертательной геометрии, механике, физике, химии, нефтехимии, экологии, политике, медицине, строительству, информатике и информационным технологиям и по другим наукам. В последние годы проводится большая работа по татарской локализации компьютерных систем. Полностью переведены на татарский язык и адаптированы для татарского языка последние версии операционной системы MS Windows и офисные программы. Татарскую локализацию информационных систем, а также составление словарей в настоящее время выполняют разные группы. Очевидно, при этом важным и критичным является обеспечение единообразия и идентичности терминов и понятий, включаемых в компьютерные системы и издаваемые словари. Одним из важных условий для этого является наличие четко прописанных принципов и критериев терминотворчества для татарского языка, которых придерживаются специалисты.

2. Формальный взгляд на образование новых слов (терминов, понятий) на татарском языке

Принципы создания новых терминов и понятий на татарском языке можно рассматривать, подразделяя их на две группы. Принципы первой группы отвечают на вопрос: как должны создаваться татарские термины (наименования), принципы второй группы отвечают на вопрос: какими должны быть татарские термины.

2.1. Принципы первой группы словообразования на татарском языке

I. Поиск готовых наименований (терминов) в самом языке

Например, санак (палка пастуха с засечками для расчетов) – компьютер (рус.), нишан – печать (рус.), мурта – (бал корты) – пчела (рус.), чүрә (литературного варианта нет) – изнанка (рус.), тәрәз (тәрәзә) – window (англ.) – окно (рус.) (применяется для технологий).

В настоящее время данный принцип, хотя и декларируется в соответствующих инструкциях, работает не очень активно, однако является весьма перспективным и имеет большие возможности для создания новых терминов, особенно, при привлечении слов из диалектов татарского языка.

II. Создание новых терминов, применяя правила татарского языка

1) <Корень>+<аффиксы>, т.е создание нового термина путем присоединения к корню аффиксов. Например, *бегунок – шудырма*: [*шудыр* гл. повел. накл., ед. число] + *ма* [словообр.афф.], *база данных – тамгасар*: *тамга* [корень] + *сар* [афф.мн.ч.], *папортниковые – абагасыманнар*: *абага* [корень] + *сыман* [афф.сравн.] + *нар* [мн.ч.], *меню – сайлак*: *сайла* [гл. повел.накл., 2 лицо] + *к* [словообр.суфф.].

В силу того, что татарский язык является агглютинативным языком, возможности образования новых терминов при помощи аффиксов весьма богаты. В особенности, если учесть, что некоторые аффиксы могут служить и для образования новых слов и для склонения, а некоторые могут преобразовать слово из одной части речи в другую (например, имя существительное в глагол, глагол в имя существительное, глагол в имя прилагательное, глагол в наречие и т.д.), то можно понять, какой богатый морфологический потенциал имеется в татарском языке для создания новых терминов. В настоящее время особенно активны аффиксы *–чы/-че, -лык,/-лек, -лы/-ле, -ла/-лә* и некоторые другие. В тоже время, ряд аффиксов, например, такие как - аффиксы *-ма/-мә,* преобразующие глагол в имя существительное (*ярма: яр+ма*), *-[ə]к/-[а]к* (*тарак: тара+к*), *чә/-ча* - имя существительное в имя существительное (*кулча: кул+ча, аркача: арка+ча*), *-чек/-чык* (*уенчык: уен+чык*) активно не применяются и изучаются лишь с точки зрения диахронии. Их можно назвать «спящим» потенциалом татарского языка.

2) <Корень>+<корень>, то есть соединением двух коренных слов создается слово, которое означает одно понятие. Например, *видеохәтер (видеопамять): видео* [корень] + *хәтер* [корень], *аспрограмма (подпрограмма)*: *ас* [корень] + *программа* [корень], *ачыкчак* (*выдержка* – в фотоделе): *ачык* [корень] + *чак* [корень], *ярымсүз (полуслово)*: *ярым* [корень] + *сүз* [корень].

3) Определение (обозначение) одного понятия через словосочетание. Например, *ЭХМ архитектурасы (архитектура ЭВМ), кара карга (ворона), файлны өстәп кертү (включение файла).*

4) Создание нового понятия применением искусственного парного слова. К первому слову присоединяется через дефис вспомогательное слово. В результате создается обобщенная форма слова, обозначенного первым коренным словом. Например, *савыт-саба (посуда), бала-чага (дети), тирә-юнь (окрестность).*

Конечно же, потенциал татарского языка по созданию новых терминов исходя из возможностей самого языка не исчерпывается указанными вариантами. Парные коренные слова, явления синонимии, применение словосочетаний и т.д. являются новыми, пока не до конца изученными возможностями для создания и определения новых понятий и терминов.

III. Заимствование терминов из тюркских языков.

Среди десятков тюркских народов на сегодня есть и такие, которые пишут научные труды на своем языке, применяют свой язык в компьютерных технологиях, создают новые термины. Самыми активными в этой области являются турки, татары, казахи,

азербайджанцы, киргизи, якуты, чуваши. Но несмотря на это, обмен терминами, заимствование терминов из других тюркских языков практически отсутствует. Соответственно, это ведет к отдалению тюркских языков друг от друга.

IV. Заимствование терминов из других языков

Это в настоящее время один из самых активно работающих принципов. Например, *утюг – үтүк, файл – файл, компьютер – компьютер, stack – стәк*. Заимствование терминов напрямую, без перевода, из языка оригинала, естественное явление. В большинстве случаев, для преподавания наук на татарском языке это является оптимальным, оправдывает себя. Особенно в условиях двуязычия, как в Республике Татарстан, это явление широко распространено при переводе содержания учебных предметов на татарский язык. И вполне естественно, что такие понятия, как интеграл, функция, дифференциал, анализ, синтез, заимствованные через русский язык, активно использующиеся в узкой предметной сфере, принимаются без изменения. Очевидно, при заимствовании понятий и терминов из других языков необходимо следить за тем, чтобы они ассимилировались в языке, т.е. адаптировались по правилам татарской орфоэпии, иначе, заимствования запутают татарский язык своим произношением, склонениями. К сожалению, это явление уже успело закрепиться в татарских словарях и даже в правилах орфоэпии. Как следует из исследований, проведенных специалистами научно-исследовательского института «Прикладная семиотика» АН РТ при создании синтезатора татарской речи, до 35-40% слов в орфографическом словаре татарского языка являются заимствованиями, не ассимилированными в языке. Так как татарский язык не является активным в официальной сфере, в науках и компьютерных технологиях, острота этого явления практически не ощущается большинством ученых, тем более широкой общественностью. Однако с каждым днем, при заимствовании новых понятий и терминов, это явление, при отсутствии ассимиляции, разрушает татарский язык изнутри, повреждает его главное свойство – внутреннюю логику и закономерности.

V. Прямой перевод из других языков при заимствовании

1) **Прямой перевод.** Перевод коренных слов: *mouse – мышь – тычкан*; перевод словосочетаний: *булево значение – буль кыйммәте.*

2) **Смысловой перевод. Создание нового термина по смыслу.** Например, если термин *mouse – мышь – тычкан* был введен сообразно форме, для “хвостатого” или проводного устройства, учитывая то, что сейчас имеют все большее распространение бесхвостые варианты, это устройство можно было бы назвать по-другому, например, «йомры» - кругляк, «бака» - лягушка, или даже «кабартма» - пирожок. Функциональный смысл данного устройства – управление курсором – стрелкой на экране. Исходя из этого его можно было бы назвать как “укйөрткеч” (водить стрелкой) или “укидарә” (управление стрелкой). Это устройство мы назвали в операционной системе Windows XP как «йомры», но учитывая, что вариант «тычкан» уже был широко распространен при преподавании информатики и информационных технологий в школах и вузах, в операционной системе Windows 7 обратно вернулись к термину «тычкан». Для примера, приведем еще несколько терминов, созданных по принципу смыслового перевода: *антивирусная программа – вирустан дәвалаучы программа, Save as – Саклау рәвеше.*

Пять принципов, описанные выше, это в основном те принципы, которыми в той или иной мере пользовались или пользуются ученые и специалисты для создания терминов на татарском языке, которые описаны в соответствующих справочниках, правилах и инструкциях по образованию татарских терминов и понятий.

Здесь мы предлагаем добавить еще три новых принципа создания новых терминов и понятий на татарском языке.

1) принцип «формального гнезда». Суть принципа: из татарского слова создаются формальные конструкции, состоящие из последовательности согласных букв, путем пропуска гласных букв, затем, вставляя гласные буквы между согласными, создаются новые слова.

2) принцип «возвращения», «восстановления» в языке. Суть принципа: слова, которые сохранились в других языках, будучи заимствованными ранее из тюркско-татарского языка, возвращаются в язык обратно. Сюда же относится и восстановление архаизмов, которые по тем или иным обстоятельством перестали активно применяться в языке.

3) «блендинг» - синтетический принцип. Суть принципа: соединением метафор создается новая метафора, новое понятие из нескольких метафор. Данный принцип активно используется в индо-европейских языках.

VI. Принцип формального гнезда.

При анализе татарских коренных слов можно прийти к выводу, что татарские слова созданы, заполнением гласными буквами схемы из согласных букв. Рассмотрим несколько примеров. Если учесть, что в татарском языке имеется 9 гласных (а-ә, о-ө, у-ү, ы-е, и) по схеме Т-З создаются следующие односложные слова: *таз, тәз, тоз, төз, туз, түз, тыз, тез, тиз*. Из этих 9 слов в татарском языке используются 7: *таз (тазик), тоз (соль), төз (стройный), туз (береста), түз (терпи), тез (строй), тиз (быстро).* Для формы “К-Н” из 9 возможных на татарском языке используется 6 слов: *кан (кровь), кон (кон), көн (день), кун (заночуй), күн (кожа), кын (ножны).* Хотя слово *кон* (используется в карточной игре) не является ассимилированным и не звучит по-татарски, он вошел в таком виде в словари. В то же время на основе формы “Җ-Л” образовано только одно слово *җил (ветер)*, не имеется ни одного слова, образованного на основе формы “Б-Н”.

Какой вывод можно сделать, исходя из данных примеров?

Во-первых, на основе некоторых «формальных гнезд» образуются до 7 слов из 9 возможных, от других еще меньше, а от третьих не может быть ни одного слова, как в последнем примере. Таким образом, некоторые слова, созданные методом формального гнезда, на сегодняшний день не употребляются.

Во-вторых, это явление само по себе может натолкнуть на интересные исследования: не сохранены ли эти «спящие» слова в диалектах, не обозначены ли они как анахронизмы в словарях? Также их поиск в других тюркских языках и выяснение их смысла может привести, с точки зрения диахронии, к интересным открытиям для татарского языка.

Тот же принцип диахронии можно применять и к структуре согласных, образующих несколько слогов. К примеру, рассмотрим форму «Б-С-К». Здесь формально образуются следующие слова: *басак, басык, басик, басук, босак, босук, босык, босик, бусак, бусык, бусик, бусук, бүсәк, бүсек, бүсик, бүсүк, бысак, бысук, бысик, бысык, бәсәк, бәсик, бәсүк, бәсек, бөсәк, бөсүк, бөсик, бөсек, бисәк, бисүк, бисик, бисек.* Из этих 32 слов только одно – *басак* применяется как неологизм в смысле “принтер”. В то же время, слово *бысак – нож* применяется в башкирском языке. Здесь, при заполнении схем гласными буквами также учитывалось и то, что в многосложных словах в последнем слоге в татарском языке не применяются гласные “О”, “Ө”. На основе формы «С-Н-К» из 28 возможных слов употребляются 3: *сүник (угаснем), сынык (сломанный), сәнәк (вилы),* а для формы “Т-Р-К” таких слов 7: *тарак (расческа), тырык (междометие), төрик (завернем), төрек (живой), тирәк (тополь)*. В настоящее время в компьютерных технологиях применяется возрожденное слово *санак (компьютер)*, которое можно образовать на основе формы “С-Н-К”. Таким образом, можно сказать, что хотя и редко, но слова, образованные по принципу “формального гнезда”, уже находят применение. Также к словам, образованным по этому принципу, можно отнести слова “күрәк” – “дисплей”, “күрсәр” – “курсор”, “сайлак” – “меню”, “буяк” – “тонер”, “турак” – “размелчитель бумаги”. Видно, что эти слова принадлежали бы к группам, автоматически созданным на основе форм “к-р-к”, “к-рс-р”, “с-йл-к”, “б-й-к”, “т-р-к”, соответственно.

Формальные конструкции, приведенные выше, отражают богатый, пока “спящий” словообразовательный потенциал татарского языка. Как видно из примеров, те слова, которые образованы добавлением гласных букв к формам, состоящим из согласных, даже если не употребляются и не содержатся в словарях, звучат и пишутся по-татарски. Соответственно, по мере развития наук и технологий, по мере появления новых понятий, те

слова, которых пока нет в употреблении, но которые можно создать автоматически по указанной схеме, можно будет использовать для обозначения новых терминов.

VII. Блендинг (Blending): гармоническое соединение метафор в одном понятии

Данный принцип означает принцип словообразования разнообразным объединением двух или более метафор в одну новую метафору. Например, соединением слов *пүчтәк* (рус. пустяк) и *күчтәнәч* (презент) можно образовать слово *пүчтәнәч* (пустяковый презент). Например, из слов *гафу үтенәм* (рус.: прошу прощения) и *итенәм* (рус.: кривляюсь) получается метафора *гафу итенәм*, то есть кривляясь, прошу прощения. Несмотря на то, что в татарском языке присутствуют похожие на блендинг слова - *таяныч* (*таяк+аяныч*), можно сказать, что данный принцип в татарском языке специально для образования новых понятий и терминов не используется.

VIII. Возвращение, восстановление

То, что татарский язык как один из тюркских языков является древним языком, уже факт, доказанный во многих научных трудах. Следов татарских слов можно увидеть не только в русских словарях, но также в греческой, арабской, английской, германской, китайской лексике, лексике американских индейцев. Перешли ли эти слова и понятия из тюркского языка, который когда-то был глобальным языком и занимал ведущее место в развитии цивилизаций, являясь языком описания мировых явлений и процессов, или некоторые из этих языков являются потомками древнетюркского языка, измененными до неузнаваемости? На этот счет у ученых не имеется единого взгляда и разные источники отвечают на этот вопрос по-разному.

В качестве примера возвращенных слов приведем слова «эт» и «айкен». При написании адреса электронной почты используется символ @. Этот символ американцы вводили сначала как коммерческий знак, англоязычные пользователи называют его - «эт» (собака). Если учесть, что этот знак и в самом деле похож на собачку, есть основание думать, что данный знак является глифом, наследованным у народа майя, обозначавшим собаку - “эт” (по-татарски). Многие пользователи так и называют данный символ собачкой (кстати, турки его называют “көчек” – маленькая собака, собачка). Для возвращения древнетюркского слова нам остается лишь прочитать этот знак как “эт”.

Пиктограмму – небольшое растровое символическое изображение, используемое в графическом интерфейсе пользователя для выбора того или иного инструмента (программы) или файла и управления им называют по-русски *иконка*, по-английски - *icon* (айкен). На языке майя слово ай-кен означает луна-солнце. Также и по-татарски. Таким образом, предполагая, что слово айкен является древнетюркским, пиктограммы естественно называть термином *айкен*.

К восстановленным словам относятся слова, которые ранее были активными, например, применялись в прошлых веках в медресе, в повседневной жизни, а потом вышли из употребления, стали «архаизмами».

В качестве примера можно привести математический термин “хасилә” – производная. Данный термин встречается в малоизвестном русско-татарском математическом словаре, изданном в 20-х годах прошлого столетия еще на арабской графике. К таким словам в той или иной мере можно отнести слова *хисап (расчет), мөгаллим (учитель), әсбап (предмет).*

2.2. Вторая группа принципов словообразования на татарском языке

Как отмечалось выше, вторая группа принципов определяет, какими должны быть новые понятия и термины.

Первый принцип: понятие, термин должны обозначаться как можно более коротким словом. Лучше, если это слово корневое.

Второй принцип: выгоднее всего термины и понятия обозначать одним словом (особенно, с точки зрения технологий).

Третий принцип: надо стараться избегать омонимии. Например, слово *печать* - *бастыру* лучше чем слово *язу*, так как писать можно и на экране, а печать (бастыру) возможна только на принтере.

Четвертый принцип: понятия и термины должны быть понятными и ясными. (То есть они должны быть распространенными, принятыми большинством, хотя и могут звучать не совсем по-татарски – *компьютер, дифференциал, функция*).

Пятый принцип: термин и понятие должны быть благозвучными (“красивое звучание”) (*cash* – по-русски произносится *кэш*, но по-татарски *кәш, stack – стек – стәк, tag – тег – тәг*).

Шестой принцип: в качестве термина и понятия брать синонимы, которые употребляются редко (возможно, диалектный вариант) (*окно – тәрәз)*.

Седьмой принцип: заимствования из иностранных слов брать напрямую, не через другой язык.

Восьмой принцип: максимально избегать неологизмов. Не нужно создавать слова, которые являются непонятными, сложными для употребления и понимания.

Девятый принцип: термины и понятия из других языков должны заимствоваться только в корневой форме. Грамматические, просодические склонения новых слов должны подчиняться правилам только татарского языка, но не правилам языка заимствования.

2.3. Многословные конструкции при образовании терминов на татарском языке

Естественно, все понятия в инфокоммуникационных технологиях нельзя выразить при помощи однословных терминов. Образуются, как и во всех языках, устойчивые обороты, многословные конструкции. При переводе с английского обычно многословные конструкции из английского превращаются в такие же на татарском языке, например, add-in memory - өстәмә хәтер и.т.д. Анализ и использование таких конструкций легко поддается алгоритмизациии. Но вместе с тем есть случаи, когда однословные конструкции из английского превращаются в устойчивые многословные конструкции на татарском языке, например, cancel – баш тарту. При образовании устойчивых многословных конструкций используются те же девять выше рассмотренных принципов.

3. Практические работы, выполненные с применением татарской терминологии

На основе терминологической системы, описанной выше, и на основе описанных основных принципов, в научно-исследовательском институте «Прикладная семиотика» Академии наук Республики Татарстан выполняются практические работы по созданию татарской терминологической системы в области информатики и информационных технологий. Приведем два примера такой деятельности. Первый – составление англо-татарско-русского толкового словаря по информатике и информационным технологиям, второе – татарская локализация операционных систем и офисных приложений фирмы Майкрософт.

Толковый терминологический словарь по информатике и информационным технологиям был издан в издательстве «Мәгариф” в 2006 году [5]. Этот словарь содержит более 7000 терминов.

При составлении этого трехъязычного толкового словаря мы опирались прежде всего на опыт, который был накоплен в совместной научно-исследовательской лаборатории искусственного интеллекта АН РТ и КГУ, и на словари, которые были разработаны в высших учебных заведениях Татарстана.

Термины в области информатики и компьютерных технологий условно были разделены на четыре большие группы: 1) термины из области компьютерной техники (hardware), 2) термины из области программного обеспечения (software), 3) термины сервиса и интерфейса, то есть термины, которые создают татароязычную среду при работе на компьютере, 4) термины, которые используются при преподавании информатики.

При создании терминов авторы опирались на принципы, описанные в данной статье.

Термины, используемые массово, в повседневной жизни, авторы старались обозначать словами, имеющимися в языке или образованными на основе словарных слов по правилам татарского языка. Например, к таким понятиям можно отнести понятия *компьютер, калькулятор, дисплей, меню, принтер, тонер.* Эти слова, во-первых, ни по произношению, ни по записи не совпадают с формой, общепринятой в мире, так как они вошли в татарский язык через русский. Во-вторых, все эти слова не отвечают закономерностям татарского языка, т.е. не ассимилированы. В то же время, учитывая, что эти понятия относятся к инструментарию, а в татарском языке такие слова образуются в основном присоединением к корню аффиксов –ак, -әк, -к, можно образовать следующие татарские термины: *компьютер – санак (санау+ак)*, неологизмы: *дисплей, монитор – күрәк (күрү+әк), меню – сайлак (сайлау+ак), принтер – басак (басу+ак), язак (язу+ак), тонер – буяк (буяу+ак), калькулятор – сансанак (сан санау + ак)*. При помощи аффикса *–сар*, по образцу понятия *каенсар* (березовая роща, много берез) можно образовать следующие термины: *тамгасар (тамга+сар: много знаков) – база данных, төймәсар – (төймә+сар: много кнопок) – клавиатура*. Слова, приведенные в качестве примера, не нарушают правил татарского языка, легкопроизносимы, благозвучны, по смыслу точно отражают объект. Исходя из этого, можно считать данный принцип терминообразования удачным для татарского языка. В данном словаре таких слов немного, но по мере раширения сферы применения татарского языка, по мере развития научного татарского языка, очевидно, их число будет увеличиваться.

Специальные термины, которые применяют только специалисты в области компьютерной техники или программисты, в основном были взяты без изменений, лишь в нужных случаях адаптируя фонетически.

Большое внимание уделялось и составлению словника, отбору терминов. Здесь важными считались следующие критерии: термины должны быть часто употребимы в широкой аудитории, они должны относиться именно к информатике и технологиям.

Толковый англо-татарско-русский словарь терминов информатики и информационных технологий, также другие специальные терминологические словари были активно использованы при татарской локализации операционных систем Windows XP, Windows Vista, Windows 7, Windows 8 и их офисных приложений. Из 5000 терминов, использованных при татарской локализации этих систем две тысячи – новые термины, созданные на основе принципов татарской терминологии, приведенных выше. В настоящее время в научно-исследовательском институте «Прикладная семиотика» АН РТ идет работа над обновленным вариантом толкового англо-русско-татарско-чувашского словаря по информатике и информационным технологиям, дополненным терминами, созданными при указанных локализациях.

Татарская локализация операционной системы MS Windows включает, в основном, следующее:

1) перевод интерфейсов операционных систем и ее офисных приложений – перевод текстов, которые отображаются на дисплее (например, текстов, которые отображаются при работе с e-mail, с приложениями Windows Word, Excel и т.д.)
2) перевод на татарский язык текстов на кнопках меню (Save, Open, Close, OK, Yes, Delete, Print и т.д.)
3) перевод файлов справок и помощи на татарский язык

Главные условия для локализации состоят в следующем:

1) правильность, естественность, точность татарских текстов (то есть перевод не прямой, не «калька»)
2) краткость, понятность и правильность текстов (команд, действий) на кнопках меню
3) понятность и компактность текстов в файлах справок

Национальная локализация операционной системы и офисных приложений – это не прямой перевод с английского и русского, а творческая адаптация программного продукта для обеспечения комфортной работы татароязычного пользователя в среде Windows. Важно,

чтобы версии Windows на разных языках имели одинаковый смысл, все тексты должны восприниматься одинаково, должно соблюдаться политкорректность по отношению к пользователю, особенно четко и понятно должна быть переведена информация о правах пользователя и фирмы-производителя. Татарская локализация компьютерных систем требует использования знаний из разных областей - технологий, лингвистике, татарскому языку, информатике.

Заключение

Данная статья посвящена вопросам формирования новых понятий и терминов в татарском языке и задаче построения терминологической системы в одной из наиболее быстро развивающихся научно-прикладных областей – области информатики и инфокоммуникационных технологий. Очевидно, чтобы вновь созданные термины стали неотъемлемой частью языка, обогатили язык и расширили горизонты его применения, еще недостаточно порождать на татарском языке новые понятия и термины, а необходимо, чтобы эти термины и понятия прошли, по крайней мере, три этапа. Во-первых, татарские термины должны активно применяться в науке, культуре, и в средствах массовой информации. Во-вторых, татарские термины должны использоваться в процессе получения и оформления новых научных результатов, которые будут интересны всему научному сообществу. В-третьих, татарские термины должны использоваться на других языках в научных публикациях зарубежных авторов.

Литература

1. Heintz J. and Schonig C. Turcic Morphology as Regular Language // Central Asianic Jornal (CFJ), 1989. -P.1-24.
2. Suleymanov D.S. Natural cognitive mechanisms in the Tatar language // In the Collection of the Vienna Proceedings of the Twentieth European Meeting in Cybernetics and Systems Research. Edited by Robert Trappel. Vienna, Austria, 6-9 April, 2010. – P. 210-213.
3. Правила создания, совершенствования и использования татарских терминов (Татар терминнарын ясау, камилләштерү һәм куллану кагыйдәләре) // Составители: Закиев М.З., Низамов И.М. – Казан, 1995. – 13 с.
4. Татарская грамматика. Т.2. Морфология. – Казань: Тат. кн. изд-во, 1993. – 397 с.
5. Сулейманов Д.Ш., Галимянов А.Ф., Валиев М.Х. Термины по информатике и информационным технологиям: англо-татарско-русский толковый словарь (Сөләйманов Җ.Ш., Галимҗанов Ә.Ф., Вәлиев М.Х. Информатика һәм мәгълүмат технологияләре терминнары: инглизчә-татарча-русча аңлатмалы сүзлек). – Казань: Магариф, 2006. -383 с.

А.К.ХИКМЕТОВ, О.Л.КАРУНА, К.К.КАРЖАУБАЕВ

Казахский Национальный Университет имени аль-Фараби, Алматы, Казахстан

АДАПТАЦИЯ LINUX-СИСТЕМ ДЛЯ ИХ ИСПОЛЬЗОВАНИЯ В РЕСПУБЛИКЕ КАЗАХСТАН

Необходимость создания высокотехнологичной экономики РК ставит на первое место развитие науки и всех её структурных оснований по производству новых знаний, приборов и ПО. Прошлое десятилетие послужило толчком к разработке большого количества приложений на казахском языке, что существенно продвинуло казахскую научную школу на международную арену. Неотъемлемую часть формирования научно-производственной инфраструктуры составляют операционные системы (ОС), на основе которых

функционируют вычислительные машины, обеспечивающие делопроизводство компаний, работу различной техники на заводах и т.д. Наиболее популярной в Казахстане является ОС Windows, однако надежность и дороговизна данной ОС оставляет желать лучшего. В связи с чем, более приемлемой считается ОС семейства Unix, бесплатная лицензия, многозадачность, а также надежность, которых являются решающим аргументом в выборе ОС, особенно при работе на кластерных системах. Системы на базе UNIX показывают большие функциональные возможности, позволяют достичь более высокой степени защиты информационной системы, позволяют создавать автономную информационную среду, сохраняя при этом возможность интегрирования в другие системы с использованием стандартных протоколов обмена данными.

Unix-подобная операционная система Linux повсеместно используется в Европе, России, США, Японии и т.д. Применимость казахских шрифтов в Linux возможна при условии создания нового стандарта кодирования. Прозрачность документации Unix-подобных ОС позволяет создавать любые драйверы до требуемой глубины детализации, создавать собственные библиотеки (стандартные подпрограммы, используемые в различных приложениях). Авторы данной статьи в рамках проекта «Разработка защищенной операционной системы с поддержкой казахского языка на основе Linux-платформ» осуществляют адаптацию ОС Linux для казахстанских пользователей в соответствии со следующими этапами:

1. Разработка 8-битной кодировочной системы для консоли.
2. Создание и внедрение новой раскладки клавиатуры для консоли.
3. Создание шрифтов консоли ОС LINUX.
4. Создание шрифтов для графической среды ОС LINUX.
5. Перевод на казахский язык интерфейсов популярных программ среды Linux.
6. Создание векторных шрифтов для графической среды Linux.
7. Разработка кодировки Unicode для Linux.

Адаптация начинается с создания файла kz.map, который содержит настройки раскладки клавиатуры. Переключение с одного языка на другой осуществляется с помощью правой клавиши Ctrl. Далее производится задание букв казахского алфавита в соответствии с клавишами клавиатуры keycode 2, keycode 3 - keycode 9, keycode 0.

Для консоли ОС Linux был разработан шрифт Cyrkza8x16.psf на основе следующих разработанных программ: CONVERT - выводит на экран изображение букв казахского языка и символов находящихся в psf файле; DRAW - редактирование бинарных файлов; PSFCREATE - осуществляет сбор всех бинарных файлов в один psf файл (шрифт).

Загрузкой раскладки в консоль занимается утилита loadkeys. Ей на вход подаётся файл раскладки *.map, в котором описано поведение каждой клавиши. Для использования внедренных казахских букв был взят за основу и изменен файл соответствия ru.map, где были назначены коды казахских букв к клавишам 2, 3, 4, 5, 8, 9, 0, -, =. После загрузки раскладки в консоль становиться возможным создание в консоли файлов и папок на казахском языке.

Пошаговое внедрение шрифтов в консоли осуществляется в следующей последовательности:

Загрузка шрифта (setfont /usr/share/kbd/consolefonts/Cyrkza8x16.psfu)

Загрузка кодировки KOI-8rk (mapscrn /usr/share/kbd/consoletrans/koi8rk)

Загрузка соответствия между вводом (клавиатура) и выводом (экран) \\ (loadkeys /usr/share/kbd/keymaps/i386/qwerty/kz.map)

Менеджеры окон (Window managers) — часть графического пользовательского интерфейса, позволяющая управлять размерами и расположением окон на экране, сворачивать и разворачивать окна, а также отвечающая за внешний вид окон (например, вид заголовков, рамок и т.д.) – также были преобразованы в соответствии с казахскими названиями используемых кнопок.

При создании *.bdf шрифтов использовалась программа Font Forge. Было создано 60 казахских шрифтов. При создании которых в каждом шрифте были прорисованы казахские буквы и расставлены соответствующие ссылки на Юникод в соответствующих ячейках шрифта. crox1c.bdf. crox1cb.bdf, crox1cbo.bdf, crox1co.bdf, crox1h.bdf, crox1hb.bdf, crox1hbo.bdf , crox4tb.bdf, crox4tbo.bdf, crox4to.bdf, crox5h.bdf, crox5hb.bdf, crox5hbo.bdf , crox5ho.bdf, crox5t.bdf, crox5tb.bdf, crox5tbo.bdf, crox5to.bdf , crox6h.bdf, crox6hb.bdf, crox6hbo.bdf, crox6ho.bdf, kz-koi10x20-20.bdf, kz-koi12x24-24.bdf, kz-koi12x24b-24.bdf, kz-koi5x8-8.bdf, kz-koi6x10-10.bdf, kz-koi6x13-13.bdf, kz-koi6x13b-13.bdf, kz-koi6x9-9.bdf, kz-koi7x14-14.bdf, kz-koi8x13-13.bdf, kz-koi8x16-16.bdf, kz-koi8x16b-16.bdf, kz-screen8x16-16.bdf, kz-screen8x16b-16.bdf – название некоторых созданных казахских шрифтов.

Следующим этапом стала разработка комбинированных символов, содержащихся в некоторых позициях UCS. Стандарт Unicode 3.0, опубликованный Unicode Consortium, содержит полный уровень реализации UCS Basic Multilingual Plane – уровень 3, как описано в стандарте ISO 10646-1:2000. К Unicode 3.1 также добавлены дополнительные уровни ISO 10646-2. Стандарт Unicode и технические сообщения, публикуемые Unicode Consortium, обеспечивают много дополнительных рекомендаций по использованию разных символов. Также поясняются руководящие принципы и алгоритмы для редактирования, сортировки, сравнения, нормализации, преобразований и выводе строк Unicode. Все это потребовало разработки настроек кодовых преобразований и локалей для kz-utf.map.

Адаптация ОС Linux для Казахстана поможет жителям нашей Республики быстрее и эффективнее осваивать новые технологии, позволит сократить время на адаптацию сотрудников к программному обеспечению и созданию специальных отраслевых решений, которые будут учитывать специфику местного рынка.

Литература

1. Bach M. J., «The Design of the UNIX Operating System», Englewood Cliffs, NJ, Prentice Hall, 1987.
2. Alexander Mikhailian, Belarusian-HOWTO, TLDP, 2001.
3. Tomohiro KUBOTA, «Introduction to i18n », Official debian documentation, 1999.
4. Бектаев К., Большой казахско-русский, русско-казахский словарь, 2007.
5. Сыздыкова Р.Г., Қазақша-орысша сөздік. Казахско-русский словарь, Дайк-пресс, 1008 стр., 2002.
6. http://www.gnu.org/software/gettext/manual/gettext.html

Т.СУЛЕЙМЕНОВ, Р.С.НИЯЗОВА, Л.Т.УРАЗБАЕВА.

Л.Н.Гумилев атындағы Еуразия Ұлттық университеті, Астана, Қазақстан

МӘТІНДІК ӘРІПТЕРДІ АУЫСТЫРУШЫ БАҒДАРЛАМАЛЫҚ ҚАМТАМАЛАР ЖҮЙЕЛЕРІНІҢ ВЕРИФИКАЦИЯСЫНДАҒЫ СЕНІМДІЛІК МӘСЕЛЕЛЕРІ

Мақсаттық жүйе ешқашанда монолитті болмайды да ол бірнеше компоненттерден тұрады. Яғни бұл жағдайда жүйенің сыртпен әсерлесуі сол компоненттердің өз ара әсерлесуімен жалғасып жатады. Соңғысы ішкі процесс ретінде боладыда сырттан бақылауға көнбеуі мүмкін. Ендеше біз жүйелерінің компоненттерінің сенімділігін болжай білуіміз керек.

Сенімділік деп тасымал, сақтау, жөндеу және программалық қамтамалық қызмет көрсетуде, берілген режимде және қолдану шарттарында талап етілетін функцияларды орындау мүмкіндігін бейнелейтін барлық параметрлердің мәндерінің белгіленген шегінде уақыт бойынша объектінің қасиетін сақтауды атаймыз. Пайдалану шарттарын кеңейту,

радиоэлектронды құрылғылармен орындалатын функциялардың жауапкершілігін жоғарылату, олардың күрделенуі өнімнің сенімділігіне деген талаптың жоғарылауына алып келеді [1].

Сенімділік күрделі қасиет болып табылады, және тоқтаусыздық, ұзақ мерзімділік, қайта қалпына келу және сақталыну сияқты құрамалардан қалыптасады. Мұндағы негізгісі тоқтаусыз жұмыс істеу қасиеті – уақыт ағымында бұйымның жұмысқа қабілеттілік жағдайын үздіксіз сақтау қабілеті. Сол себепті программалық қамтамалық құралдардың сенімділігін қамтамасыз етуде оның тоқтаусыздығын жоғарылату анағұрлым маңызды болып табылады.

Сенімділік мәселелерінің ерекшелігі оның программалық қамтамалық құрылғыларының «өмірлік циклінің» барлық этаптарымен байланысы болып табылады, құру идеясының пайда болуынан бастап сипатталуына дейін: өнімді есептеуде және жобалауда оның сенімділігі жобаға салынады, дайындау кезінде сенімділік қамтамасыз етіледі, пайдалану кезінде − жүзеге асырылады. Сол себепті, сенімділік мәселесі – кешенді мәселе және де оны барлық кезеңде, сонымен қатар түрлі құралдармен шешу қажет. Өнімді жобалау кезеңінде оның құрылымы анықталады, таңдау немесе элементтік базаны әзірлеу орындалады. Сондықтан мұнда программалық қамтамалық құрылғылардың талап етіліп отырған деңгейде анағұрлым жоғары мүмкіндікті сенімділігі қамтамасыз етіледі. Бұл есепті шешудің негізгі әдісі болып жобаны тізбекті қажетті түзетуі бар, оның құрамдас бөліктерінің сипаттамалары мен объектінің құрылымына тәуелді сенімділікті есептеу, бірінші кезекте − тоқтаусыз жұмыс істеу болып табылады.

Сенімділікті жоғарылатуды талап ететін себептердің бірі программалық қамтамалық жүйелердің күрделенуінің, оларға қызмет көрсететін аппаратуралардың өсуі, оларды пайдаланудағы шарттардың және тапсырмалардың жауапкершілігінің қатаңдығы болып табылады.

Программалық қамтамалық жүйелердің [2]жеткіліксіз сенімділігі жобалауға, өндіріске және осы жүйелерді пайдалануға кеткен жалпы шығынмен салыстырғанда эксплуатациялық шығынның үлесінің өсуіне алып келеді. Мұнымен қоса, программалық қамтамалық жүйелердің эксплуатациясының құны оны өңдеуге және дайындауға кеткен бағадан бірнеше есе асып түсуі мүмкін. Бұдан басқа, программалық қамтамалық жүйелердің тоқтап қалуы әр түрлі салдарға алып келеді: ақпаратты жоғалту, программалық қамтамалық жүйелермен жанасқан басқа құрылғылардың және жүйелердің бос тұрып қалуы, апаттың болуы және т.б.

Сонымен қатар, ақырғы есепте программалық қамтамалық жүйелердің сенімділігі іріктеліп жиналған элементтердің сенімділігімен анықталады. Сол себепті сенімділіктің элементтік қорының негізгі сұрақтарын білу қазіргі таңда табысты жұмыстың қажетті шарты болып табылады.

Бұл жұмыста программалық қамтамалық жүйенің тоқтаусыздығының сандық сипаттамалары, олардың жалпы сипаттамалары, сонымен қатар программалық қамтамалық жүйенің құрылымдық−логикалық анализін, құрылымдық сенімділігін есептеу қарастырылған. Қазіргі шақта бізде программалық қамтамалық жүйелердің сенімділігін арттыру әдістері қарастырылып жатыр.

Бұл жұмыста жоғарыда қарастырылып келген жүйенің құрылымды сенімділігін есептеу, программалық қамтамалық жүйелердің сенімділігін арттыру әдістерін бағдарламалық қамтамада жүзеге асыру жүргізілген, яғни программалық тілде жобалау немесе жүзеге асыру бөлімі қамтылған.

Әдебиеттер

1. Шарипбаев А.А., Ефимкин К.Н.,Задыхайло И.Б.Об одном подходе к верификации программ обработки символьной информации. Тез.док.Всесоюзной конференции «Методы искусственного интеллекта», Паланга,1980,с.67-70

2. Шарипбаев А.А. Редукция проблемы верификации программ к проблеме выполнимости логических формул. Доклады национальной академии наук РК, №6, Алматы,1994, с.15-21

ТҮРІК ТІЛДЕРІНІҢ ЭЛЕКТРОНДЫ КОРПУСТАРЫ
ЭЛЕКТРОННЫЕ КОРПУСЫ ТЮРКСКИХ ЯЗЫКОВ
ELECTRONIC CORPORAS OF TURKIC LANGUAGES

Р.Я.ГИБАДУЛИН[1], Я.Н.ГИБАДУЛИН[1], А.Р.САКАЕВ[1], М.З.ЗАКИЕВ[2], И.М.САЛАМАТИН[3]

[1]НКО "Инсан" г.Москва,
[2]РФ, ИЯЛИ, г.Казань, Татарстан,
[3]РФ, ОИЯИ, г.Дубна, РФ

ЭЛЕКТРОННЫЕ СЛОВАРИ ТЮРКСКИХ ЯЗЫКОВ

Ключевые слова: электронные словари, корпус тюркских словарей, компьютерная лексикография, мультимедиа.

В настоящее время известны и развиваются два вида электронных словарей:

1) работающие при поддержке Интернета on-line и 2) автономные, не нуждающиеся в использовании Интернета, off-line словари.

Статья посвящена созданию электронных off-line словарей тюркских языков.

1. Введение

В советскую эпоху издательства "Советская энциклопедия" и "Русский язык", выполняя государственную программу создания словарей на языках народов СССР, в сотрудничестве с учеными из Академий наук Союзных и Автономных республик, подготовили и издали большое количество различных тюркологических словарей. Авторами и составителями многих из них были видные ученые лингвисты-тюркологи и авторитетные авторские коллективы, включавшие известных ученых того времени из национальных республик, а также из центра. Труды этих ученых не потеряли свою актуальность и ценность и для нашего времени. Можно сказать, что созданные ими словари, являются культурным наследием советской эпохи. Некоторые из этих словарей перечислены ниже[1-6]. Эти словари использовали алфавиты на кириллической основе, т.е. буквы русского алфавита и их графические модификации. Алфавиты не были унифицированы по тюркским языкам и, в результате, в советскую эпоху в каждой тюркоязычной республике бывшего СССР использовался свой кириллический алфавит.

В связи с переходом с кириллицы на латиницу в ряде новых независимых тюркоязычных государствах стала актуальной задача переиздания на латинице ранее изданных кириллических бумажных словарей .Переиздание их на латинице по известной технологии офсетной печати весьма трудоемкий, длительный и дорогостоящий процесс. Более целесообразнымм представляется переиздание устаревших бумажных словарей в электронном виле. В этом случае кириллические тексты могут быть автоматически (программным способом) перекодированы в латиницу на уровне интерфейса пользователя, либо перекодировка может быть выполнена полностью для всего словаря.

2. О словаре исторически однокоренных слов татарского языка

Словарь разработан в издательстве ИНСАН в период с2009 по 2013 гг.[7]. Содержит около 36 тысяч слов, объединённых в гнёзда по родству. Является одной из первых работ по исследованию общетюркских корневых слов на примере татарского языка. В отличие от существующих словарей однокоренных слов, гнёзда эти – не только словообразовательные. Например, традиционная тюркская лингвистика не считает однокоренными слова *китап* и *мәктәп*, которые в данном словаре, в силу их исторического родства сводятся в одно гнездо как производные от арабского глагола *qtb* (qataba) «читать».

При составлении словаря использовались материалы исследований многих компаративистов и их оппонентов. Результатом проведенной многолетней работы стал опыт презентации близких и дальних связей лексики татарского языка, как одного из тюркских

языков. При этом дальнее родство выявляется уже на евразийском уровне. Например, в одно гнездо (группу) попали такие казалось бы совершенно разнящиеся по смыслу татарские слова как *ятарга* «лежать» (общетюрк.), *диван* (перс.), *фөрьяд* «вопль, стенание» (перс.) и *кәнфит* «конфета» (из русского языка через немецкий из латыни). Все они восходят к праевразийскому (ностратическому) корню *д’ ~ *дғ «класть». В сносках-комментариях даны этимологии слов с возведением их (где возможно) к древнейшему корню и показаны иные производные того же корня, пришедшие в татарский язык другим путем.

Отметим, что ссылки играют важнейшую роль в данном словаре, составляя половину его объема. Возможно, именно ссылки будут представлять особый интерес, поскольку здесь приведены мнения крупнейших лингвистов (тюркологов, арабистов, индоевропеистов) о происхождении того или иного слова. В ряде случаев в комментариях приведены народные этимологии и спорные версии происхождения слов с анализом этих мнений учеными-тюркологами.

Большая часть словника – общее тюркское наследие. В словарь включены и диалектные слова, если они имеют интересные параллели в общетюркской лексике. Это слова среднего (казанского) и западного (мишарского) диалектов, уральских говоров, но не сибирско-татарского диалекта (по сути, отдельного языка). Например, лексика кряшен (крещеных татар) любопытна тем, что в ней доля языческого явно преобладает. При этом для нужд народной религии приспособлены не только древнетюркские реликты, но и заимствованные у татар-мусульман арабизмы. Сравните, тат.-кряш. *кереметь* «языческое капище; священная роща» от тат. *кәрамәт* «чудо».

Попутно отметим, что приведённый в словаре материал привлечет внимание и к решению ряда орфографических проблем. Например, к назревшей необходимости введения общетюркских норм написания сложных слов (например, *тимер казык – тимерказык* «Полярная звезда; север») и заимствований (тат. *сурәт / сүрәт*, каз. *сурет*, уйг. *сүрәт* «изображение»).

Задача данного словаря – показать лексические связи татарского языка как с тюркским миром от древнейших времен до современности, так и с внешними языками, оказавшими на него влияние. В первую очередь, это языки арабский, персидский и русский. Сходные процессы заимствования характерны для большинства тюркских языков, что нашло отражение в словаре.

Словарь исторически однокоренных слов татарского языка предназначен для широкого круга читателей, интересующихся историей тюрков и их языков. В качестве справочного пособия словарь поможет этимологическим исследованиям тюркологов и, можно надеяться, будет способствовать раскрытию белых пятен в истории тюркских языков.

Словарь реализован в виде мультимедийного программного продукта с использованием программной технологии создания электронных словарей.

3. О программной технологии создания электронных словарей

За период от разработки в НКО "ИНСАН" первых вариантов электронных словарей с 2007 г.до настоящего времени было испытано несколько версий программной реализации технологической цепочки создания словарей. Это был естественный процесс. За это время существенно изменилась компьютерная база, изменились операционные системы, трансляторы программ, появились более мощные редакторы текстов, в инструментальных программных средствах обеспечена возможность работать с UNICOD-ом. Это стимулировало принятие решения о разработке новой технологической цепочки программных средств для реализации электронных мультимедийных словарей для тюркских языков и приложений на их основе. Этот процесс непрерывного обновления, развития и совершенствования программных технологических средств объективно закономерен и продолжается по настоящее время.

Создание электронного словаря проходит ряд этапов. На подготовительном этапе формируется исходный текст словаря. Словари могут быть различного назначения –

дву(много)язычные, фразеологические, толковые и др. На этом этапе осуществляется в основном лингвистическая проработка словаря: составление словарных статей, выбор их структуры, информационных полей и пр., производится их заполнение соответствующими данными. Отметим, что при этом структура словарной статьи остается неизменной для всех словарных статей данного словаря. Естественно, она может измениться для других типов словарей. Исходный текст нового словаря может создаваться в отсутствии прототипа с "чистого листа", как в случае со словарем [7]. Часто исходный текст словаря заимствуется из ранее изданных «бумажных» словарей. В электронном переиздании «бумажного» словаря структура словарных статей при необходимости может быть изменена, например, добавлены новые информационные поля. Лингвистическая проработка словарных статей на подготовительном этапе производится с помощью специальных сервисных программ, призванных максимально облегчить подготовительную работу лингвиста, в частности, обеспечить быстрый доступ к справочной информации. Результатом подготовительного этапа является лингвистически выверенный текст словаря.

На следующих этапах программная технологическая цепочка включает программы разбора текста словаря (парсер), заполнения базы данных, подготовки звуковых файлов, выполнения других операций по проверке целостности и защите базы данных. Программа разбора текста имеет ряд режимов работы, предназначенных для проверки текста с целью выявления и устранения нарушений принятой структуры словарных статей. В этих режимах выводятся фрагменты текста с обнаруженными ошибками, указывается их местоположение. Нарушения могут быть самыми различными, например, отсутствие данных в некоторых информационных полях словарной статьи, перевода, ключевого слова, служебных символов разметки текста и т.д. Помимо этого, парсер подготавливает таблицу входов для записи звукового файла перевода.

Заполнение базы данных выполняется после завершения коррекции текста всех словарных статей процедурой, которую вызывает парсер. Разработчику предоставляется возможность контроля состояния базы данных, редактирования словарных статей и другие операции коррекции.

Запись звуковых переводов производится в диалоговом режиме. Диктору- оператору на экране предоставляется таблица входов. В таблице отмечаются входы, для которых имеются подготовленные для озвучивания тексты из словарных статей. Диктор выбирает в таблице вход, для которого нужно записать звуковой файл перевода, инициирует запись и голосом прочитывает(произносит) показанный ему в диалоговом окне текст. Процесс записи останавливает также диктор, после чего программа записи автоматически формирует название файла и записывает созданный звуковой файл в формате MP3 в базу данных. Затем операция записи повторяется для другого входа. Для удобства контроля записи на экране все время индицируется уровень шума в помещении. После завершения записи можно прослушать записанное и при необходимости перезаписать текущий или любой из ранее записанных звуковых файлов. Можно также воспользоваться специальным фирменным программным обеспечеиием редактирования звуковых файлов и подавления помех[8].

4. Заключение

1. Разработаны программные технологии создания электронных словарей как вновь разрабатываемых, так и воссоздаваемых на основе "старых", традиционных бумажных словарей. Последние в электронном издании обретают "новую" жизнь и обладают зачастую функциональными характеристиками недоступными для традиционных словарей.

2. При использовании разработанной технологии реализованы автономные (off-line) русско-татарский [5] ,татарско-русский [7] и русско-башкирский [6] электронные словари. Подробное описание этих словарей и руководство пользователя приведены в [9].

3. Описанная программная технология может быть использована и для создания других электронных словарей различных типов, в том числе электронного переиздания тюркских

словарей [1-4] на латинице. При этом учет особенностей вновь создаваемого или переиздаваемого словаря производится при его лингвистической проработке на подготовительном этапе. На остальных этапах технологическая цепочка остается практически неизменной.

4. Словарь (приложение) может быть использован как основа для создания других приложений, например, словарей обучения произношению, переводу и автоматическому чтению текстов, создания мультимедийных учебников и др.

Литература

1. Кенесбаев С.К. Фразеологический словарь казахского языка //изд-во "Гылым", Алма-ата, 1977, 712 с. (Более 10 тыс. фразеологических единиц).

2. *Юдахин К.К.* Киргизско-русский словарь // изд. "Сов. энциклопедия",1965, 976 с.(Около 40 тыс. слов)

3. *Чарыяров Б., Алтаев С.* Большой русско-туркменский словарь // В 2-х томах,т.1 816 с, т.2 752с., 1986, изд-во "Русский яз."

4. *Хамзаев М.Я.* (ред.) Толковый словарь туркменского языка // Ашхабад, 1962.

5. *Ахунзянов Э.М., Газизов Р.С., Ганиев Ф.А. и др.* Русско-татарский словарь // Изд-во " Русский яз.",736 с.,1984,1985, 1991 - (Около 47 тыс. слов).

6. *Ураксин З.Г.* Русско-башкирский словапрь // В 2-х томах. т.1 808 с, т.2 680 с, изд-во "Башкирская энциклопедия", Уфа, 2002.

7. *Сакаев А.Р.* Татарско-русский словарь исторически однокоренных слов // Рукопись словаря , 2013 (в печати).

8. Sony Sound Forge Pro 10 User's Manual.

9. Сайт www.tatar-tele.info.

ТАШПОЛОТ САДЫКОВ[1], БАКЫТ ШАРШЕМБАЕВ[2]

[1]К.Карасаев атындагы Бишкек гуманитардык университети,
[2]Кыргыз-түрк Манас университети, Кыргызстан

«МАНАС» ЭПОСУНУН УЛУТТУК КОРПУСУН ТҮЗҮҮ ЖӨНҮНДӨ

Кыргыз элинин улуттук сыймыгы, көөнөрбөс көрөңгөсү, улуу мурасы жана соолбос руханий булагы болгон Манас дастаны миңдеген жылдар бою атадан балага, муундан муунга өтүү аркылуу биздин күнгө жетип олтурат. Дастаныбыз көлөм жактан дүйнөдө теңдешсиз, мазмун жактан элибиздин көөнө тарыхын, алмустактан берки рухий жана заттык маданиятын чагылдырган, поэтикалык жактан көркөм сөз өнөрүнүн эң жогорку деңгелине жеткирилген жалпы адамзаттык маанидеги эстелик экени талашсыз. Муну «*Белес-белден бороондоп, Беш удургуп өткөн Сөз. Баласына атасы Мурас кылып кеткен Сөз. Эли сактап жүрөккө, Биздин күнгө жеткен Сөз»* деп дастаныбыз өзү бир тастыктаса, *«Манас» эпосу жалпы адамзаттык маанидеги терең маани-мазмунга сугарылган көркөм сөз өнөрүнүн туу чокусу, дүйнөлүк көчмөндөр цивилизациясынын кенчи, кыргыз элинин улуттук аң-сезиминин манифести жана көркөм идеологиясы, Алатоолук ак калпак калктын турмушунун энциклопедиясы болуп эсептелет. Поэтикалык күчү, эпикалык арымы жана көлөмү жагынан дүйнөдө теңдеши жок улуу дастан. «Манас» - байыркы кыргыз рухунун туу чокусу»* деп залкар жазуучубуз Чыңгыз Айтматов дагы бир ирет тастыктайт.

Бүгүнкү күндө 2,5 миллиондон ашуун ыр сабын камтыган эпостун токсонго жуук варианты Улуттук Илимдер академиясы Ч.Айтматов атындагы тил жана адабият институнун колжазмалар фондунда сакталып турса, миллиондон ашуун ыр сабын камтыган текст

кытайлык кандаштарыбыздан катталганы анык. Эбегейсиз зор көлөмдөгү 3,5-4 миллион ыр сабын ичине катыган бул казынабызда элибиздин эчен кылым карыткан улуттук маданияты, дүйнө кабылдоосу, менталитети, чарбачылыгы, устачылык, аңчылык, саяпкерлик, сынчылык өнөрү, үрп-адат, салт-санаасы, адеп-ахлак, жүрүм-туруму, ишеними, диний, мифологиялык, философиялык түшүнүктөрү, экологиялык, астрономиялык, географиялык, медициналык билими, жоокерчилик өнөрү, курал-жарак, буюм-тайымдары, үй эмеректери, кийим-кечектери, аш-той, тамаша-зоок, шаң-салтанаттары, коңшу тайпалар менен болгон мамилелери, атажурттун ажайып кооздугу, аска-зоо, тоо-таш, жайлоо-төр, өрөөн-өңүр, талаа-түз, өзөн-сууларынын көрк-касиети, каармандардын кулк-мүнөзү, кыймыл-аракетти, келбет-көрүнүштөрү, болочок урпактарды атажурттун атуулу, эрктүү, күчтүү, кайраттуу, чапчаң кылып тарбиялоодо эрсайыш, балбанкүрөш, оодарыш, көкбөрү, аламан байге, жорго салыш сыяктуу элдик оюндардын ролу таамай сүрөттөлүп, таасирдүү көркөм сөз менен берилген.

Улуттук тилибиздин мартабасын мамлекеттик деңгелге көтөрүүдө, адабий норманы эне тилдин төл кыртышында өркүндөтүүдө, тилибизди илим-техника-технология, башкаруу-өндүрүш-бизнес тилине айлантып, дүйнөлүк маалыматтар мейкиндигине алып чыгарууда да даңазалуу дастаныбыздын мааниси баа жеткис. Манас эпосу, акыйкатта да, кыргыз тилинин көөнөрбөс алтын казынасы, анда катылган эбегейсиз сөз байлыгы, көркөм сөздүн асыл берметтери, аңыз-уламыш, жөө жомок, санжыра, макал-лакаптары эне тилибиздин дүйнөдөгү эң бай, кооз, таасирдүү, элестүү жана туюнтуу кудурети мол тилдердин катарына жатарын кадиксиз тастыктайт.

Колжазмалар фондундагы эпостун негизги деп табылган нускалары ондон ашуун. Ыр саптарынын саны боюнча булардын көлөмү төмөнкүдөй:

Нускалар	**Манас**	**Семетей**	**Сейтек**	**Бардыгы**
Сагымбай	180 378	-	-	180 378
Саякбай	84 830	218 787	196 936	500 553
Шапак	24 588	42 338	14 718	81 644
Тоголок Молдо	53 045	24 390	-	77 435
Багыш	141 147	67 704	5 594	214 445
Молдобасан	57 718	43 102	2 760	103 580
Ибраим	3 731	23 364	7 839	34 934
Мамбет	106 002	52 059	43 333	201 394
Шаабай	8 368	-	3 842	12 210
Жакшылык	-	52 136	145 959	198 095
Мамбеталы	26 952	-	-	26 952
Ысса	-	14 763	-	14 763
Жаңыбай	19 445	66 454	-	85 899
Бардыгы	**706 204 сап**	**605 097 сап**	**420 981 сап**	**1 732 282 сап**

Ал заманыбыздын залкар жазма манасчысы атанган жана кытайлык кыргыздардын өкүлү болгон Жусуп Мамай атабыз тарабынан жазылган Манастын сегизилтиги 200 миңге жуук ыр сабынан турат экен. Демек, жакынкы биздин максат – корпустук лингвистиканын

жетишкендиктерине таянуу менен «Манас» эпосунун улуттук корпусун түзүү жумушун колго алуу.

«Манас» эпосунун академиялык басылышын басмадан чыгаруу иши толук бүткөрүлбөй, учурда улантылып жаткандыктан, улуттук корпуска улуу манасчыларыбыз Сагымбай Орозбак уулу менен Саякбай Карала уулунун мурда жарык көргөн варианттары киргизилди. Алар төмөнкүлөр:

Сагымбай Орозбак уулу. Манас. 1 китеп. Ф: Кыргызстан, 1978.

Сагымбай Орозбак уулу. Манас. П китеп.Ф: Кыргызстан, 1980.

Сагымбай Орозбак уулу. Манас. Ш китеп. Ф: Кыргызстан, 1981.

Сагымбай Орозбак уулу. Манас. IV китеп. Ф: Кыргызстан, 1982.

Саякбай Каралаев. Манас. I китеп. Ф: Кыргызстан, 1984.

Саякбай Каралаев. Манас. II китеп. Ф: Кыргызстан, 1986.

Саякбай Каралаев. Семетей. I китеп. Ф: Кыргызстан, 1987.

Саякбай Каралаев. Семетей. II китеп. Ф: Кыргызстан, 1989.

Саякбай Каралаев. Сейтек. Ф: Кыргызстан, 1991.

Буга кытайлык манасчыбыз Жусуп Мамай тарабынан жазылган нускасы (Манас. Шинжаң эл басмасы: 2004, 1782 б.) кошумчаланды.

Ошентип, корпуска жүктөлгөн өйдөкү текстердин негизинде эпосто катталган сөздөрдүн грамматикалык формаларынын толук тизмесин түзүү, ар бир сөздүн грамматикалык формаларын ошол сөздүн уясына бириктирүү, сөздүктө камтылган бардык бирдиктерди алфавит тартибинде жайгаштыруу, сөздүн лексикалык маанилерин түркчө которуп берүү, ар бир сөз менен анын бардык грамматикалык формаларынын кайсы вариантта, канчанчы бетте, кайсы сапта колдонулгандыгын тастыктаган даректерин көрсөтүү иштери аткарылды. Демек, мындай сөздүк эпостун китеп түрүндө даярдалган маалымыттар банкы, алфавит тартибинде жайгаштырылган сөз аркылуу текстке чыгуучу ачкычы катары кызмат кылып, кыргыз элинин тарыхын, улуттук тилин, этномаданиятын, этнографиясын, менталитетин, мифологиясын, философиясын, фольклорун, этнопедагогикасын, ата мурастарын, нарк-дөөлөттөрүн изилдөөгө өбөлгө түзүп, көмөк көрсөтөрү анык.

Сөздүк түзүү үчүн тандалып алынган Сагымбай менен Саякбайдын тексти корпуска жүктөлгөн соң түпнускадагы ар бир бет көрсөтүлүп, ар бир сапка катар номур ыйгарылды. Бул, алибетте, сөздүктөгү сөздөн текске чыгуунун төтө жолу.

Иштин экинчи этабында тексте кездешкен ар бир сөз формасы жалпы тизмеде алфавит тартиби боюнча жайгаштырылып, бардык тексте колдонулган даректерине улам шилтеме берилип туруу аркылуу эпосто колдонулган сөз формаларынын алфавит тартибиндеги тизмесин түзүү жана ал тизмедеги ар бир сөздүн дарегин көрсөтүү менен аяктады.

Иштин үчүнчү этабында лематизация маселесин чечүү максаты көздөлдү. Лематизация деп компьютердин жардамы менен текстеги сөздү (= сөз формасын) анын сөздүктөгү турпатына (= лексемага, сөзгө) келтирүү процесси аталат. Бирок, тилекке каршы, кыргыз тили боюнча лематизатор алигиче жасалбагандыктан, текстеги сөз формасын сөздүктөгү турпатына келтирүү ишин кол менен жасоого туура келди. Бул ишти аткарууда ар кандай тыбыштык өзгөрүүлөрдөн улам бир сөздүн ар башка грамматикалык формалары алфавит тартиби боюнча катар жайгашпай, жалпы тизменин баш-аягына чейин чачылып кеткен учурлар арбын кездешти.

Аларды бир уяга топтоо көп эмгекти жана убакытты талап этти. Маселен, *азап* сөзүнүн *азабы, азабым, азабын* сыяктуу формалары *азада, азазил, азай-, азамат, азан* сөздөрүнүн уясынан, *ак-* сөзүнүн *агып, агыптыр* сыяктуу формалары *адам, адат, адаш, адис, ажар, азан, азил, азоо, айбан* сөздөрүнүн уясынан мурда келет.

Ошентип, текстеги сөздү сөздүктөгү турпатына келтирүүдө бир сөздүн ар башка грамматикалык формалары бир жерге топтолуп, сөздүн мааниси түркчөгө которулуп, кийинки сапта сөздүн өзү баш тамгасына чейин кыскартылып, баш тамгадан кийин чекит коюлуп, сөздүн грамматикалык формаларын уюштуруучу мүчө же мүчөлөрдүн айкашы алфавит тартибине келтирип, андан соң булардын баарынын колдонуш даректери

көрсөтүлүп берилди.

Буга мисал кылып ***аккаңкы*** сөзүнүн төмөнкү беренесин келтирүүгө болот:

аккаңкы **eyerin bir türü**
К1:229-27, К2:62-58, 148-7, 243-44,
О3:287-27
а.га К3:201-17, К4:13-94, 124-49, 187-29
а.ны К1:65-10, 176-90, 189-110, К2:62-61,
К5:66-95, 69-64, 257-14
а.нын К3:42-68, 100-52, К4:89-14, 96-66, 123-61.

Мында ээрдин бир түрүн билдирген ***аккаңкы*** сөзү эпос текстинде өз алдынча сөздүктөгү турпатында да, ***-га, -ны, -нын*** мүчөлөрү уланган жөндөмө формаларында да колдонулгандыгы ачык көрүнүп турат. Ал эми бул формалардын даректери мындайча чечмеленет:

а) кош чекитке чейинки К тамгасы Саякбай Карала уулунун, О тамгасы Сагымбай Орозбак уулунун вариантын билдирсе, андан кийинки сан варианттын канчанчы тому экендигин билдирет,

б) сызыкчанын сол жагындагы сан бет номурун, оң жагындагы сан сап номурун көрсөтөт.

Эми өйдөкү беренге сереп салсак, анда, маселен, ***аккаңкы*** сөзү ушул турпатында Саякбайдын Каралаевдин 1-томунун 229-бетиндеги 27-сапта, 2-томунун 62-бетиндеги 58-сапта, 148-бетиндеги 7-сапта, 243-бетиндеги 44-сапта, Сагымбай Орозбак уулунун 3-томунун 287-бетиндеги 27-сапта колдонулгандыгы айкын болот. Аталган сөздүн калган формаларынын кайсы даректерде кездешкени ушундай эле жол менен тастыкталат.

Иштин жогоруда белгиленген этаптарын ишке ашырууда кыргыз жана түрк лексикографиясынын калыпташкан салттары эске алынып, бул багытта аткарылган изилдөөлөр жана сөздүктөр кеңири пайдаланылды. Ошону менен катар эпостун сөз байлыгын мүмкүн болушунча толук каттоо максатында орфографиянын эски нормаларынан четтеп, 2002-жылкы кыргыз тилинин жазуу эрежелеринин жаңы редакциясы сунуштаган нормаларга артыкчылык берилди.

Ошентип, сөздүктү түзүүдө төмөнкү жоболор жетекчиликке алынды:

1. Сөздүккө чыгарылган сөздөр алфавит тартибинде жайгаштырылып, алардын маанилери түркчөгө которулуп берилет.

2. Кош сөздөр өзүнчө сөз катары сөздүккө чыгарылып, алфавит тартибинде эмес, биринчи түгөйүнүн уясынан кийинки катарда берилет. Мисалы:

акыл akıl, us, zihin, zeka
а.га
а.дан
а.ы
а.ына
акыл-айла hile, kunazlık
акыл-насаат tavsiye, öğüt.

3. Эпос текстинде айрым кошмок сөздөр бириктирилип да, ажыратып да жазылган.

Кубаты кетип тайтактап,
Куруп турат ал ***ак куу***
К4:19-81

Кабыландын Акшумкар
Аккуу менен алышып
К4:19-96

Ак куу кебин кийинип
Айчүрөк учуп кетти эми
К4:209-83

Мындай сөздөр сөздүктө бир бүтүн сөз катары бириктирилип берилди.

4. Энчилүү аттар жалпы аттардан бөлүнбөй, алар менен чогуу алфавит тартибинде сөздүккө алынат да, адам аты, тулпар аты, жер аты, суу аты сыяктуу түркчө белгилер, кыскача түшүндүрмөлөр менен коштолот. Мисалы:

албарсты dev nevinden kadın varlık
а.дай
а.нын
Албүбү Kanıkey'in uşağı
албыр- yüzünden nur saçmak, parlamak
а.ып
Алгара Koñurbay'ın küheylânı
а.га
а.ны
а.сын.

5. Жер-суу аттары адам жана айбан аттарынын үлгүсүндө бириктирилип берилет. Мисалы:

Аксарай Kanıkey'in sarayı
Аксаргыл Manas'ın savaş atı
Аксеңир yer adı
Аксур küheylân adı
Аксуу yer adı
Акталаа yer adı
Актелки küheylân adı
Актүз yer adı
Баркөл göl adı
Ботомойнок yer adı
Итөлбөс yer adı.

6. Омонимдер тексten териштирилип, омонимдик катарга топтолот да, маанилери түркчө которулуп жайгаштырылат. Эгерде омонимдик катарда этиш сөз бар болсо, анда этиш сөз бул учурда да, калган бардык учурларда да аягына сызыкча коюлуп берилет. Мисалы:

бута çalı
бута hedef, nişan
бута kumaş türü
бута- budamak.

7. Эпос текстинде колдонулган бир катар сөздөрдүн грамматикалык формалары өз ара омонимдик катышка кириптер болот. Мисалы:

атам (менин атам) **атам** (мен атам)
атты (атты мин) **атты** (ок атты).

Айрым учурларда бир сөздүн грамматикалык формасы менен экинчи бир сөздүн өзү текстеги омонимдик катышты түзөт. Мисалы:

асыл (туюк мамиле) **асыл** (кымбат баалуу)
атым (менин атым) **атым** (бута атым).

Мындай учурлар омонимдик катыштын булагы болгон мүчөлөрдү бөлүп алып, тийиштүү сөздүн уясына жайгаштыруу же сөздүн маанисине түшүндүрмө берүү жолу менен чечилет:

асыл yüce soylu
асыл- asılmak, takılmak
ат at
а.ым
а.ты
ат- atmak, fırlamak
а.ам
а.ты

ата baba
а.м
атым bir ok atımlık yer.

8. Четчил каткалаң үнсүз менен аяктаган сөздөр таандык формаларда (сал. *аспап–аспабы, белек–белеги*) соңку үнсүзүн жумшартат. Бирок буга карабастан алар негизги сөздүн уясында берилет. Мисалы:

аспап alet
а.ың
белек armağan, hediye
б.и
б.им
б.ин
б.ке
б.ти.

9. Айрым сөздөргө мүчө жалганганда соңку муундагы кууш үндүүсүн жоготуп, сөздүн негизи тыбыштык жактан өзгөрөт. Мындай сөздөрдүн кыскарган турпаты да, толук турпаты да которулуп, өзүнчө сөз катары алфавит тартибинде берилет. Мисалы:

айб suç, kahabat
а.ы
а.ым
айл köy
а.ы
а.ым
а.ыңа
айыл köy
а.га
а.дан
а.ын
а.ынан
айып suç, kahabat
а.ка
а.ы
а.ын.

10. Эпос текстинде *арк-нарк, айза-найза, амыз-намыс, араб-арап, арабыча-арапча, бери-пери, байгамбар-пайгамбар* сыяктуу бир сөздүн ар башка диалектилик варианттары кездешет. Бул варианттардын ар бири бирдей котормо менен коштолуп, өзүнчө сөз катары алфавит тартибинде реестрге чыгарылат.

11. Эгерде сөздүккө алынган айрым сөздөрдүн маанисин ачып берүү кыйынчылык туудурса, анда мындай сөздөн кийин толкун сызыкча коюлат да, айкашкан сөзү кошо көрсөтүлөт. Мисалы:

андис ~ мерген keskin nişancı
анжы беш ~ beşe dallanma
байман ~ күрүч pirincin türü
байбайлуу ~ кундуз su samuru
бас ~ кел- denk gelmek.

12. Тууранды сөздүн маанисин ачып берүү үчүн ал сөздөн кийин толкун сызыкча коюлуп, аны менен айкашкан жардамчы этиш кошо берилет. Мисалы:

булт ~кой- fırlamak
бүлк ~эт- silkinmek
былк ~эт- kımıldamak.

13**.** Ат атоочтун туура эмес формалары өзүнчө сөз катары алфавит тартибинде берилет. Мисалы:

ага ona
анын onun
буга buna
буган buna.

14. Башкы муунун кайталоо жолу менен жасалган күчөтмө сөздөр реестрге чыгарылат. Мисалы:

аппак bembeyaz
бүпбүтүн büsbütün.

15. Мааниси түшүнүксүз болгон айрым сөздөр болжолдоп которулуп, керектүү учурда контексти чогуу көрсөтүлүп, андан соң суроо белгиси менен коштолуп берилет. Мисалы:

батын~ ы ачык эр Кошой cesur Koşoy (?)
бүйөнгө (~ тийип мойнуна) ok, mermi (?)
бөздө -аркар атса ~п ал- dağ koyunu ateş etse beze sararak (?) almak.

Ошентип, жогорудагы принциптердин негизинде түзүлгөн «Манас» эпосунун чоң көрсөткүч сөздүгү Түркияда Түрк тил куруму тарабынан басылып, дүйнөлүк коомчулукка сунушталганын жана Жусуп Мамай нускасы боюнча түзүлгөн көрсөткүч сөздүк да басмага даяр турганын белгилей кетмекенибиз оң. Эгерде бул үч нусканы салыштыра келсек, анда төмөнкүдөй статистикалык даректерге күбө болобуз:

1. Нускалардын көлөмү. Жусуп Мамай нускасынын текстин компьютерге жүктөп, анын көлөмүн аныктоо маселесин койгонубузда аталган варианттын көлөмү 750904 сөз колдонушун камтыры айкын болду. Бул, алибетте, эбегейсиз зор көлөм. Анткени ал Сагымбай менен Саякбайдын өйдөкү варианттарын кошуп эсептегендеги 598956 сөз колдонуш көлөмдөн да ашып түштү. Сал.:

Жусуп Мамай	Сагымбай	Саякбай	Сагымбай+Саякбай
750 904	163 962	434 994	598 956

2. Сөз формасынын жалпы саны. Эпосто колдонулган сөз формасынын тизимин алып, тизимге камтылган бирдиктерди эсептей келгенибизде Ж.Мамайдын вариантында 58690 сөз формасы катталганы анык болду. Эгерде бул көрсөткүчтү салыштыра келсек, Сагымбай менен Саякбайдын варианттары сөз формасы байлыгы жагынан Ж.Мамайдын нускасынан алда канча төмөн турары аныкталды. Сал.:

Жусуп Мамай	Сагымбай Саякбай (бардыгы)	Сагымбай	Саякбай Манас+Семет.+ Сейтек	Сагымбай Саякбай (бирдейлери)
58 690	52 703	27 424	39 089	13 808

Көрүнүп тургандай, Жусуп Мамай нускасы, омонимдерди ажыратпаган учурда да, сөз формаларынын молдулугу жагынан Сагымбай менен Саякбайдын варианттарынын ар биринен болжол менен 1,5-2 эсеге ашып түшсө, эки вариантты бириктиргендеги бардык сөз формаларынан 6000 ге ашып түштү.

3. Сөз байлыгынын жалпы саны. Сагымбай менен Саякбайдын варианттарынын негизинде түзүлгөн чоң көрсөткүч сөздүктө 20 миң чамалуу сөз камтылганы анык болду. Ал эми Жусуп Мамай нускасынын сөз байлыгын аныктоо үчүн дагы кошумча илик абзел.

Жыйынтыктап айтканда, «Манас» эпосунун улуттук корпусун түзүү жумушу учурда улантылууда. Бул корпус келечекте кыргыз тилинин улуттук корпусунун бир бөлүгү болору анык. Учурда 4 млн.дон ашуун сөз колдонушту камтыган ар кандай жанрдагы тексттер корпуска жүктөлүүдө. Мындай олуттуу ишти аркалоодо корпустук лингвистикада топтолгон бай тажрыбаны чыгармачылык менен өздөштүрүү керектиги айдан ачык.

О. МАХАМБЕТОВ, А. МАКАЖАНОВ, Ж. ЕСЕНБАЕВ, Б. МАТКАИМОВ, И. АБЫРГАЛИЕВ, А. ШАРАФУДИНОВ

Nazarbayev University Research and Innovation System, Astana, Kazakhstan

КОРПУС КАЗАХСКОГО ЯЗЫКА: МЕТОДИКА СБОРА, СТРУКТУРИРОВАНИЯ И РАЗМЕТКИ ДАННЫХ

Аннотация

В данной работе мы представляем вашему вниманию *Текстовый корпус казахского языка* (КК – казахский корпус), что является одной из первых попыток, предпринятой местным научным сообществом составить подобный корпус. Корпус содержит более 135 миллионов различных словоформ и состоит из более чем 445 тысяч документов, сгруппированных по пяти стилистическим жанрам: *художественный, публицистический, официально-деловой, научный* и *разговорный*. Наряду с основной частью КК содержит: (1) аннотированный под-корпус, содержащий сегментированные документы в формате eXtensible Markup Language (XML), в котором закодированы полная морфологическая, синтаксическая и структурная разметки текста; (2) под-корпус с аннотированными аудио данными. КК оснащен электронной навигационной системой, доступной через Интернет, что облегчает поиск и обработку искомой информации. Корпус является открытым в обоюдном порядке: (1) данные корпуса являются свободно доступными для некоммерческого использования; (2) каждый желающий может помочь советом по улучшению, а также пожертвовать текст.

1 Введение

Эта статья описывает теоретические и практические вопросы испытанные во время разработки Корпуса казахского языка. Казахский язык – агглютинативный язык с высоким фактором инфлекции (генерации) словоформ, который относится к тюркской группе. Это - официальный государственный язык Республики Казахстан и родной для более чем 10 миллионов людей во всем мире. Но тем не менее, вплоть до начала 90-х годов 20-го века , в связи с историческими событиями в советский период , русский язык был преобладающим языком в устной и письменной коммуникации в Республике Казахстан. Этот факт в свою очередь создал проблемы в представлении казахского языка в различных областях таких как наука, развлечения, официальная документация и т.д. По этой причине , при сборке корпуса, мы должны были сгруппировать категорий, которые обычно представляются в виде отдельных корпусов, на пять стилистических жанров. Кроме того, в отличие от других корпусов (Aksan и др., 2012 . ; Chen , 1996), мы включили тексты в том виде в каких они были доступны, то есть мы не пытались заполнить предопределенный набор категорий. Значительная часть материалов была собрана с использованием веб краулеров (программа для автоматизированного сбора информации) , адаптированных под специализированные источники и пожертвованных текстов.

КК также содержит, аннотированный вручную, суб-корпус с морфо-синтаксическими и структурными наценками, которые кодируется в XML, следуя общим понятиям, изложенным в CES (IDE,1998). Наши синтаксические множества тегов содержат набор синтаксических категорий, четко определенных в классической казахской грамматике, и множество тегов частей речи (POS) основаны на позиционной системе, в которой теги образованы конкатенациями POS слова форм и цепями кодированных лингвистических свойств, таких как количество , случай, голос и т.д. Аннотации были проведены вручную студентами факультета филологии, специализирующихся в морфологии и синтаксисе. Пытаясь сделать процесс аннотации максимально комфортным, мы разработали веб инструмент с удобным

интерфейсом для аннотации . Мы позаботились о качестве аннотации, и для этого разработали систему рекомендации, которая впоследствии увеличела скорость разметки.

В рамках КК мы также скомпилировали аннотированный корпус чтения речи (RSC), которая включает в себя аудио записи слов, фраз, предложений (от всех жанров), новостных статей и отрывков из книг, которые были тщательно отобраны из основной части корпуса. Все текстовые материалы были прочитаны добровольцами разных возрастов, полов, уровней образования из разных регионов. Каждый аудиофайл сопровождается файлом этикетки и соответствующим транскриптом текста. Кроме того, некоторые транскрипты были грамматически аннотированными, т.е. в дополнение к словоуровневой сегментации аудиоинформации часть наших данных имеет лексические и морфо-синтаксические аннотации. В общей сложности RSC содержит 10GB или более чем 40 часов речи.

Эта статья организована следующим образом. Раздел 2 рассматривает существующую работу. Раздел 3 предоставляет подробную информацию о первичном корпусе. Разделы 4 и 5 подробно описывают аннотированный текст и речь суб-корпуса соответственно. Наконец, мы делаем выводы и обсуждаем будущую работу в разделе 6.

2 Схожие работы

Корпусная лингвистика стала популярной областью исследований вслед за работой Francis & Kucera(1979) с Brown University по построению первого корпуса. За последние два десятилетия исследователи по всему миру построили множество корпусов, в том числе известный Британский Национальный Корпус (БНК) (Бернард, 2007), разработанный в 1991-94 годах, и содержащий более 100 миллионов слов письменной и устной речи из различных видов источников(Ide and Macleod, 2001; Al-Sulaiti and Atwell, 2006). Все материалы выбирались на основе трех независимых критериев(носитель, жанр и временной период) и заранее определены количественные пропорции между этими критериями. Разговорная часть состоит из транскрипций неофициальных бесед и разговорного языка в различных контекстах. В БНК проведена работа по аннотации на части речи с помощью инструмента CLAWS, разработанного в университете Lancaster. БНК считается сбалансированным корпусом, и большинство исследователей используют ее модель для собственных разработок, такие как: Turkish National Corpus (Aksan et al., 2012), Korean National Corpus (Kim, 2006).

Национальный корпус русского(РНК) языка был создан группой специалистов из различных сфер под руководством Института русского языка им. В. В. Виноградова РАН (Ruscorpora, 2003). Корпус охватывает писменные (художественная и религиозная литература, мемуары, научные публикации и другие) и аудиоматериалы(публичные выступления и частные беседы) периода середины XVIII века до начала XXI века. В данный момент корпус содержит более 350 мил. лемматизированных и размеченных частями речи словоформ. Корпус также включает семантические тэги для слов и текстов (Apresjan et al., 2006). Кроме основной части в РНК имеются следующие подкорпусы: глубоко аннотированный (синтаксический) корпус, содержит тексты снабженные морфо-синтаксической разметкой основанной на лингвистической модели «Смысл <=> Текст» И. А. Мельчука и А. К. Жолковского; корпус параллельных текстов - англо-русский, немецко-русский, украинско-русский, белорусско-русский; корпус диалектных текстов; корпус поэтических текстов и другие.

3 Основная часть корпуса

КК представляет собой первую попытку построить масштабный корпус общего значения, который описывает текущее состояние Казахского языка. Корпус содержит более 135 миллионов различных словоформ и состоит из более чем 445 тысяч документов, сгруппированных по пяти стилистическим жанрам: (1) *художественный* (Казахские литературные тексты, охватывающие период с начала XX века до современности)*;* (2) *публицистический* (периодика и новостные статьи из Интернет-источников, опубликованные

за последнее десятилетие)*;* (3) *официально-деловой* (приказы, акты, и другие официальные документы, опубликованные в период с 2009 до 2012гг.)*;* (4) *научный* (книги, монографии, и работы на различные научные темы); (5) *разговорный* (популярные блог-посты, изданные с 2009 года поныне). Обращаем внимание, что мы намеренно не ставили жестких ограничений на конкретные источники данных, жанры и временные периоды. Это объясняется относительной скудностью данных и причинами, озвученными во введении.

Основными источниками данных послужили веб-сайты, а также оцифрованные книги и статьи, полученные из общественных и частных библиотек. Для каждого веб-сайта мы адаптировали веб-краулер, что увеличило точность извлечения служебной информации (автор, дата, категория, и т.д.)

Данные корпуса распространяются по лицензии, которая согласно закону РК позволяет распространять некоторые данные целиком (официальная документация, новостные статьи), и некоторые частично (литература, научные работы, аналитика), при условии, что источники должным образом указаны.

3 Аннотированный под-корпус

3.1 Разработка тэгсетов

Руководствуясь мировым опытом по созданию тэгсетов, и учитывая специфику Казахского языка, мы разработали синтаксический (члены предложения) и лексический (часть речи) тэгсеты.

Синтаксический тэгсет описан в таблице 1, содержащей наименование и описание тэгов, а также эквиваленты из широко употребляемого тэгсета, Penn tagset.

Таблица 1. Синтаксический тэгсет

№	Тэг	Описание	Эквиваленты Penn tagset
1	S	Простое предложение	S
2	BSS	Главное предложение	S
3	BGS	Зависимое предложение	SBAR, SBARQ
4	BAS	Подлежащее	NP
5	BND	Сказуемое	VP
6	TOL	Дополнение	NP, WHNP
7	ANT	Определение	ADJP
8	PYS	Обстоятельство	PP, WHP, ADVP, WHADVP
9	X	Пустой/неоднозначный член	X

Синтаксическим тэгсетом также предусмотрена разметка фразеологизмов, путем присвоения тэгу соответствующего бинарного атрибута.

Лексический тэгсет. Казахский язык относится к агглютинативным Тюркским языкам, в которых словоформы образуются путем присоединения к корню цепочки морфем. Морфемы характеризуют различные грамматические свойства (лицо, падеж, и т.д.) и несут в себе важную контекстную информацию, без учета которой лексический разбор может оказаться не полным.

Сравним варианты лексического разбора одного и того же предложения на трех языках:

Мектепке/существительное бардым/глагол ./.
I/местоимение went/глагол to/предлог school/существительное ./.
Я/местоимение пошла/глагол в/предлог школу/существительное ./.

Как видим, в Казахском варианте отсутствуют местоимение и предлог, которые переданы морфемами в лице глагола и падеже существительного соответственно:

Мектеп/сущ. + ке/(вин. падеж = предлог «В»)
бар/глагол + ды/(прошед. время) + м/(I лицо = мест. «Я») ./.

Для интеграции грамматических свойств в лексический тэгсет, был разработан позиционный тэгсет, согласно которому, лексическая метка состоит из основного тэга (развернутая часть речи) и закодированной строки грамматических свойств.

Таблица 2 содержит список грамматических свойств учтенных при создании лексического тэгсета, где под кардинальностью понимается количество возможных значений принимаемых данным свойством (например: одушевленность имеет два значения).

Таблица 2. Грамматические свойства, рассматриваемые в лексическом тэгсете

№	Грамматические свойства	Код	Кардинальность
1	Одушевленность	A	2
2	Число	N	2
3	Принадлежность	S	10
4	Лицо	P	8
5	Падеж	C	7
6	Отрицание	G	2
7	Время	T	3
8	Наклонение	M	4
9	Залог	V	5

Наконец, Таблица 3 содержит полный лексический тэгсет, состоящий из 36 базовых тэгов (исключая пунктуацию), сгруппированных по частям речи. Каждому тэгу соответствует цепочка грамматических свойств (ГС), а также генеративная емкость (Емк.), т.е. количество тэгов получаемое от всех возможных комбинаций ГС и базового тэга. Таким образом, полный тэгсет состоит из 3844 различных тэгов.

3.2 Разработка структуры для разметки текстовых данных

Руководствуясь международными стандартами хранения размеченных данных, учитывая особенности разработанных тэгсетов, а также специфику собранных данных, был разработан XMLшаблон разметки.

Согласно шаблону разметка хранится вместе с текстом, но при необходимости может быть легко отделена. Структура документа подчинена следующей иерархии, выраженной в наборе правил. Каждое правило состоит из отношения и двух аргументов, где отношения между аргументами выражают понятие «включать себя» и обозначаются стрелкой, а аргументами являются конструкции документа. Пример разметки дан в приложении I. Ниже приведены правила описывающие иерархию конструкций документа:

Фразеологизм → предложение, токен (слово, пунктуация)

Предложение → предложение (главное/зависимое), фразеологизм, токен, прямая речь, перечисление

Прямая речь → прямая речь, предложение, токен

Перечисление → перечисление, предложение, токен

Раздел → перечисление, прямая речь, предложение, фразеологизм, токен

Документ → раздел.

Таблица 3. Лексический тэгсет

№	Тэг	Описание	ГС	Емк.	№	Тэг	Описание	ГС	Емк.
		Существительное:					Местоимение:		
1	ZEP	нарицательное	ANSPC	314	20	SIMZ	личное	NSPC	229
2	ZEQ	собственное	ANSPC	314	21	SIMU	указательное	NSPC	157
		Глагол:			22	SIMS	вопросительное	NSPC	157
3	ET	основной	GTMVP	840	23	SIMD	возвратное	NSPC	157

4	ETU	инфинитив	GSC	196	24	SIMB	безличное	NSPC	157
5	ETK	вспомогательный	P	8	25	SIMY	отрицательное	NSPC	157
6	ETB	вспом., отрицание	P	8	26	SIMP	собирательное	NSPC	157
7	KEL	вспом., желательный	GT	6			Частица:		
8	ESM	причастие	GNSPC	314	27	KOM	вспомогательное имя	C	7
9	KSE	деепричастие	G	2	28	SHS	предлог	-	1
		Прилагательное:			29	SHZ	союз	-	1
10	SE	основное	P	8	30	SHD	частица	-	1
11	SES	сравнительное	P	8			Междометие:		
12	SEA	превосходное	P	8	31	OSP	обращение	-	1
		Числительное:			32	OSQ	рассуждение	-	1
13	SN	количественное	NSPC	157	33	OSO	восклицание	-	1
14	SNR	порядковое	NSPC	157					
15	SNS	собирательное	NSPC	157	34	ELK	Звукоподражание	-	1
16	SNB	дробное	NSPC	157	35	MOD	Модальное слово	-	1
		Числительное:							
17	US	основное	-	1	36	BOS	Иностранное слово	-	1
18	USS	сравнительное	-	1					
19	USA	превосходное	-	1			Суммарная емкость:		3844

5 Речевой корпус

Многие современных систем обработки речи требуют большое количество аудио и текстовых данных для создания акустических и языковых моделей. В зависимости от типы приложений данные меняются от высококачественных микрофонных начитанных записей (Garofalo et al., 2007) до разговорной телефонной речи (Godfrey and Holliman, 1997; Canavan and Zipperlen, 1996), от непрерывной речи (Garofolo et al., 1993) до отдельных слов и фраз (Leonard and Doddington, 1993; Pitrelli et al., 1995). В данной работе мы собрали более 40 часов высококачественной микрофонной казахской речи, начитанной 169 носителями языка, для задач распознавания непрерывной речи.

5.1. Текстовые материалы

Текстовые материалы для озвучивания были тщательно отобраны из основной части текстового корпуса и разделены на два раздела: предложения и статьи. Раздел «Предложения» содержит более 12000 различных предложений, равномерно и случайным образом извлеченных из пяти стилистических жанров корпуса. Предложения выбраны таким образом, что они содержат более 120 тысяч наиболее часто встречаемых слов, которые покрывают 95% всех текстов корпуса. Дополнительно, предложения сгруппированы по количеству содержащих слов так, что первая группа содержит шесть слов, вторая – семь, и так далее до длины предложения в 15 слов. Раздел «Статьи» содержит онлайн новости, извлеченные из раздела публицистического жанра корпуса. Каждая статья состоит из не более чем 300 слов.

Все материалы были разделены на непересекающиеся наборы, состоящих из 75 предложений и одной статьи. Из 75 предложений 50 представляют короткие предложения (по 10 предложений из первых пяти групп), а 25 – длинные предложения (по 5 предложений из последних пяти групп).

5.2. Дикторы

Основными критериями отбора дикторов были: регион, в котором диктор освоил казахский или провел большую часть своей жизни, пол, возраст и способность читать на казахском.

Первый критерий позволил нам уловить различные типы говора, связанные с физическим регионом проживания, как местного, так и зарубежного. С точки зрения регионального признака дикторы разбиты на 15 групп: 14 областей Казахстана и одна группа для зарубежья. Далее дикторы разбиты на следующие четыре возрастные категории: 1) 18-27 лет; 2) 28-37 лет; 3) 38-47 лет; 4) 48 лет и старше. Мы намеренно не старались балансировать дикторов по половому признаку, в виду сложностей нахождения добровольцем, но все же пытались ограничиться только не более тремя дикторами одного пола в каждой возрастной и региональной группе. Соотношение женского к мужскому полу составило 57% к 43%.

Еще одним немаловажным критерием отбора дикторов было умение свободно читать на казахском языке, так как это является общей проблемой для двуязычных стран как Казахстан. Дополнительно, мы вели информацию об образовании дикторов: наличие среднего, незаконченного высшего или высшего образования.

Дикторам присваивался шифр в соответствии со следующей кодировкой: «Регион»-«Пол»-«Год рождения»-«Инициалы»-«Образование», где «Регион» принимает значения 1-15; «Пол» - F или M; «Год рождения» последние две цифры года рождения диктора; «Инициалы» - инициалы диктора; «Образование» - 1 для школы, 2 - для колледжа/незаконченное высшее, 3 - высшее (например, 06F70ZK3).

Всего записано 169 дикторов. В табл. 7 показано распределение дикторов по возрастному, половому и региональному признакам. Пустые ячейки означают отсутствие дикторов с соответствующим профилем. В большинстве случаев, это соответствует наиболее удаленным регионам и мужским группам.

5.3. Условия записи

Фактическая запись дикторов проводилась в звукозаписывающей студии университета с участием звукооператора. Перед записью дикторы были зарегистрированы и проинструктированы, также им было дано время на подготовку. Каждый диктор заполнял соглашение о передаче исключительных прав на использование аудио данных с их голосом. При озвучивании материала от дикторов не требовалось четкой дикции и особой манеры произношения, кроме как правильного чтения материала. Среднее время записи на одного диктора заняло около 40-45 минут, хотя были и случаи достигавшие двух часов. Аудио данные были получены с помощью профессионального микрофона Neumann TLM 49 и оцифрованы на звуковой карте LEXICON I-ONIX U82S. Данные сохранены в wav-формате с одним каналом, частотой дискретизации 44,1 кГц и 16-битовой PCM-кодировкой. Все аудио файлы были обработаны вручную так, что каждое предложение и статья хранятся в отдельных аудио и текстовых файлах. Размер речевого корпуса на диске составил около 8,5 Гб, а общая продолжительность записи – более 40 часов.

5.4. Разметка и транскрипция

Каждый аудио файл сопровождается соответствующей орфографической транскрипцией и сегментацией на уровне слов аналогичной базе TIMIT, а также морфо-синтактической разметкой. Все разметка осуществлялась вручную обученными лингвистами. К примеру, орфографическая транскрипция содержит развернутые значения сокращений, чисел и дат, в соответствие с тем, как их прочитали дикторы. Дополнительно, каждое предложение начинается со специальных символов начала и конца предлдожения. Для сегментации мы использовали программное обеспечение WaveSurfer (2013), которое поддерживает аннотацию TIMIT.

6 Заключение

В данной работе мы описали процесс создания Корпуса казахского языка. ККЯ ориентирован на широкий круг пользователей, и мы верим, что он будет полезен научного

сообщества, учитывая то, что корпус имеет богатую разно-уровневую разметку текстовых и аудио данных. Более того, данная эти данные уже были использовании в наших экспериментах по морфологической сегментации и автоматической корректировке слов. Желающие могут пройти по ссылке http://kazcorpus.kz, чтобы ознакомиться с корпусом.

В качестве будущей работы мы планируем использовать данный корпус в решении таких проблем как: 1) автоматическое определение частей речи; 2) снятие морфологической омонимии; 3) машинный перевод текстов. Для последней задачи уже начаты работы по сбору параллельных текстов на русском и английском языках.

Литературы

1. G.T. Bekmanova and B.Zh. Ergesh. 2010. A system for automatic alternation of Kazakh words: word forms generator module. In Proceedings of Lomonosov 2010 international conference.
2. G.T. Bekmanova. 2010. On the approaches to automated word alternation and morphological analysis of Kazakh language. In Proceedings of the second international conference on informatics society, pages 466–469.
3. Thorsten Brants. 2000. Tnt: a statistical part-of speech tagger. In Proceedings of the sixth conference on Applied natural language processing, pages 224–231. Association for Computational Linguistics.
4. E. Brill and R. Moore. 2000. An improved error model for noisy channel spelling correction. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong.
5. Eric Brill. 1992. A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language, pages 112–116. Association for Computational Linguistics.
6.
7. Eugene Charniak. 2000. A maximum-entropy inspired parser. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.
8. Grzegorz Chrupala. 2006. Simple data driven context sensitive lemmatization. Procesamiento del Lenguaje Natural, 37:121–127.
9. K. Church andW. Gale. 1991. Probability scoring for spelling correction. Statistics and Computing, 1(2):93–103.
10. Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th annual meeting on Association for Computational Linguistics, pages 184–191. Association for Computational Linguistics
11. Mathias Creutz and Krista Lagus. 2004. Induction of a simple morphology for highly inflecting languages. In Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology, pages 43–51. Association for Computational Linguistics.
12. Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of speech tagger. In Proceedings of the third conference on Applied natural language processing, pages 133–140. Association for Computational Linguistics
13. Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. Commun. ACM, 7(3):171–176.
14. David Elworthy. 1995. Tagset design and inflected languages. In In EACL SIGDAT workshop From Texts to Tags: Issues in Multilingual Language Analysis, pages 1–10.
15. Anna Feldman. 2008. Tagset design, inflected languages, and n-gram tagging. Editors: Paul Robertson and John Adamson, 3(1):151.
16. Sheila A. Greibach. 1964. Formal parsing systems. Commun. ACM, 7(8):499–504, August.
17. Jan Haji?c and Barbora Hladk?a. 1998. Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational

Linguistics - Volume 1, ACL '98, pages 483–490, Stroudsburg, PA, USA. Association for Computational Linguistics.

18. Dilek Z Hakkani-Tur, Kemal Oflazer, and Gokhan Tur. 2002. Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities, 36(4):381–410.

19. Jirka Hana and Anna Feldman. 2010. A positional tagset for russian. Proceedings of LREC-10. Malta.

20. James Henderson. 2004. Discriminative training of a neural network statistical parser. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL'04, Stroudsburg, PA, USA. Association for Computational Linguistics.

21. Dan Klein and Christopher D Manning. 2002. Conditional structure versus conditional estimation in NLP models. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 9–16. Association for Computational Linguistics.

22. Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, pages 423–430. Association for Computational Linguistics

23. Oskar Kohonen, Sami Virpioja, and Mikaela Klami. 2009. Allomorfessor: Towards unsupervised morpheme analysis. In Evaluating Systems for Multilingual and Multimodal Information Access, pages 975–982. Springer.

24. Kimmo Koskenniemi. 1983. Two-level model for morphological analysis. In Proceedings of the 8th International Joint Conference on Artificial Intelligence, pages 683–685.

25. V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady., 10(8):707–710, February.

26. Bao-Liang Lu, Qing Ma, Michinori Ichikawa, and Hitoshi Isahara. 2003. Efficient part-of speech tagging with a min-max modular neural network model. Applied Intelligence, 19(1-2):65–81.

27. Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the kazakh language corpus. In Empirical Methods in Natural Language Processing (to appear).

28. Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In Computational Linguistics and Intelligent Text Processing, pages 171–189. Springer.

29. Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. Comput. Linguist., 19(2):313–330, June

30. E. Mays, F. Damerau, and R. Mercer. 1991. Context based spelling correction. Information Processing & Management, 27(5):517–522.

31. Microsoft. 2010. Microsoft Office 2010, Kazakh language pack.

32. Akmaral Mussayeva. 2008. Kazakh language spelling with hunspell in openoffice.org. Technical report, The University of Nottingham.

33. nlpub.ru. 2013. A small directory of linguistic resources for processing Russian language: nlpub.ru.

34. Anthony G Oettinger. 1961. Automatic syntactic analysis and the pushdown store. American Mathematical Society.

35. Kemal Oflazer and Cemaleddin Guzey. 1994. Spelling correction in agglutinative languages. In ANLP, pages 194–195.

36. Kemal Oflazer. 1994. Two-level description of Turkish morphology. Literary and linguistic computing, 9(2):137–148.

37. Praharshana Perera and ReneWitte. 2005. A self-learning context-aware lemmatizer for German. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 636–643. Association for Computational Linguistics.

38. Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. Mach. Learn., 34(1-3):151–175, February.

39. Has?im Sak, Tunga G?ung?or, and Murat Sarac?lar. 2009. A stochastic finite-state morphological parser for turkish. In Proceedings of the ACLIJCNLP 2009 Conference short papers, pages 273–276. Association for Computational Linguistics.

40. Helmut Schmid. 1994a. Part-of-speech tagging with neural networks. In Proceedings of the 15th conference on Computational linguistics-Volume 1, pages 172–176. Association for Computational Linguistics.

41. Helmut Schmid. 1994b. Probabilistic part-of speech tagging using decision trees. In Proceedings of international conference on new methods in language processing, volume 12, pages 44–49. Manchester, UK.

42. Claude E. Shannon. 1948. A mathematical theory of communication. The Bell system technical journal, 27:379–423, July.

43. A.A. Sharipbayev and A.K. Buribayeva. 2010a. Kazakh speech synthesis on a hardware level. In Proceedings of the second international conference on building information-aware society, pages 557–558.

44. A.A. Sharipbayev and A.K. Buribayeva. 2010b. Kazakh speech synthesis on a hardware level in the Quartus II environment. pages 197–203.

45. A.A. Sharipbayev, G.T. Bekmanova, B.Zh. Ergesh, A.K. Buribayeva, and M. Kh. Karabalayeva. 2012. Intellectual morphological analyzer based on semantic networks. In Proceedings of the OSTIS-2012, pages 397–400.

46. Rustem Takhanov and V. Kolmogorov. 2013. Inference algorithms for pattern-based crfs on sequence data. In International conference on machine learning

47. Scott M Thede and Mary P Harper. 1999. A second-order hidden Markov model for part-of speech tagging. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 175–182. Association for Computational Linguistics.

48. Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 173–180. Association for Computational Linguistics.

МӘТІНДІ МОРФОЛОГИЯЛЫҚ ЖӘНЕ СИНТАКСИСТІК ӨҢДЕУ ЖҮЙЕЛЕРІ
СИСТЕМЫ МОРФОЛОГИЧЕСКОЙ И СИНТАКСИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ
SYSTEMS OF MORPHOLOGICAL AND SYNTACTIC PROCESSING OF TEXTS

А.М.ГАЛИЕВА, А.Р.ГАТИАТУЛЛИН

НИИ "Прикладная Семиотика"
академии наук Республики Татарстан

ОБОЗНАЧЕНИЕ МОРФОЛОГИЧЕСКИХ КАТЕГОРИЙ ГЛАГОЛА В МОДЕЛЯХ ОКОНЧАНИЙ ТЮРКСКИХ СЛОВОФОРМ[2]

Введение

В НИИ "Прикладная семиотика" Академии наук РТ ведется работа над проектом по созданию комплексных моделей данных на основе ситуационного анализа текстов. В рамках этого проекта решаются задачи создания модели окончаний и базы данных со словарями окончаний для татарского, казахского и турецкого языков. На базе этой модели реализуется программа морфологического анализа, которая на вход получает словоформу на одном из указанных тюркских языков, а на выходе выдает структуру этой словоформы в виде последовательности морфем и в виде последовательности морфологических категорий. Одна из причин такого двойного представления результата, то, что одна и та же морфологическая категория в этих тюркских языках может быть представлена разными морфемами. Например, категория инфинитива в татарском языке представляется аффиксальной морфемой -[Ы]РГА, в турецком языке морфемой –мАк, а в казахском морфемой –У.

В данной статье рассматривается система обозначений для морфологических категорий татарского глагола со сравнением этих категорий в казахском и турецком языках.

1. Тюркский глагол

Из всех частей речи глагол выделяется лингвистами как самая сложная и самая емкая, а система тюркского глагола отличается особой сложностью и разветвленностью форм.

Для тюркского глагола характерно наличие, следующих морфологических категорий:

- категория аспекта;
- сложная система времен и наклонений, включающая синтетические и аналитические формы;
- развитая и многочленная система глагольных имен - имена действия, причастия, субстантивно-адъективных форм, деепричастные формы;
- глагольные финитные формы с обстоятельственными значениями;
- сложная система залоговых форм глагола (взаимно-совместный, понудительный, страдательный, возвратный залоги), способность показателей залоговых форм комбинироваться друг с другом в пределах словоформы;
- разнообразные формы выражения категории каузатива, причем в словоформе могут присутствовать два, три и более показателей каузатива, модифицирующих действие, выраженное знаменательной частью лексемы, располагающейся слева от каузативного аффикса.

При разработке системы обозначений для грамматических категорий татарского глагола нами изучены системы обозначений в словарях разного типа и грамматиках тюркских языков, система грамматической аннотации в Национальном корпусе русского языка, работы по общей морфологии и другие исследования. Особо следует выделить Лейпцигские правила глоссирования (The Leipzig Glossing Rules), которые были разработаны в отделе лингвистики

[2] Исследование выполнено в рамках научно-исследовательского проекта РФФИ («Разработка комплексных моделей данных на основе ситуационного анализа текстов в задачах многоязычного поиска»), проект № 13- 07-00494-А.

в Институте эволюционной антропологии имени Макса Планка и в отделе лингвистики Лейпцигского университета (http://www.eva.mpg.de/lingua/resources/glossing-rules.php). Данную систему правил можно считать своеобразным общепризнанным стандартом у лингвистов мира, в первую очередь у специалистов по типологии. Обозначение результатов поморфемного анализа в программах морфологического анализа татарских словоформ приближено к данным правилам. В ходе работы над системой обозначений привлекались и другие источники, в частности, изучена система категорий базы данных Verbum, которая отражает состав и структуру элементарных глагольных значений, выявляемых путем сопоставления форм глаголов на материале большого количества языков [13].

Рассмотрим более подробно примеры интерпретации глагольных грамматических форм и категорий в тюркских языках.

2. Темпоральные категории

По темпоральному признаку глаголы в тюркских языках выражают значения прошедшего, настоящего и будущего времени, но каждая форма обычно совмещает несколько значений, что ставит вопрос о том, какие семы выделять в качестве основных, доминирующих и фиксировать в названиях категорий.

2.1. Прошедшее время

Форма прошедшего времени на -ДЫ в разных работах по грамматике называется:

- прошедшее определенное,
- прошедшее категорическое [9],
- прошедшее очевидное.

Существующая многозначность глагольных форм приводит к тому, что в тюркских грамматиках часто предпочтительными оказываются формулировки типа «прошедшее время на -ды», «причастие на -ып», «имя действия на -у» и т. п. Такая терминология для нашей системы неприемлема с учетом того, что морфологический анализатор уже выделяет соответствующие аффиксы, и дублировать их в названиях категорий нет никакого смысла.

Если эмпирические факты свидетельствуют о том, что грамматические значения существуют лишь в «связанном» виде, то возникает вопрос: что именно фиксировать в системе обозначений? При семантическом подходе возможно выявление элементарных значений – «семантических атомов» путем межъязыкового сопоставления. Однако в нашей системе обозначений отправными точками классификации являются аффиксы, которые объединяют комплексы значений (темпоральность + модальность), при этом интерпретация зависит от контекста.

Например, в работе [15] указывается, что парадигматическим значением прошедшего категорического времени в татарском языке является «выражение очевидного, целостного, однократного действия в прошлом». Однако добавление эта однократность пропадает при добавлении в словоформу морфемы –ГАлА или –[Ы]штЫр.

Например:

барды ‘ходил’ – баргылады ‘хаживал’ – барыштырды ‘хаживал’.

При этом в различных контекстах может идти речь о временной локализации действия, о динамике развертывания действия, законченности/незаконченности действия, длительности, повторяемости действия и т.д.

Прошедшее категорическое как в татарском, так и во многих тюркских языках обычно противопоставлено прошедшему результативному или результативным, поскольку в казахском языке есть еще две формы прошедшего, выражаемых с помощью морфемы –ҒАн и -[Ы]п. Прошедшее категорическое и результативное противопоставляются не по аспектуальным параметрам, а по вовлеченности субъекта высказывания. Оно выражает действие очевидное, несомненное, свидетелем которого являлся сам субъект (субъективная модальность), когда субъект сам являлся свидетелем действия и осознает это.

Например:

яңгыр яуды ‘*шел дождь’* - субъект видел, как он шел.

При прошедшем некатегорическом субъект может наблюдать результат действия в прошлом, но не само действие

яңгыр яуган ‘*шел дождь’* - субъект видел не сам дождь, а может наблюдать лужи и мокрую землю или о дожде сообщает другое лицо. Русский перевод не отражает специфику противопоставленных форм. В татарском языкознании не выделяется специализированная грамматическая категория эвиденциальность, но значение очевидности присутствует в связанном виде со значением прошедшего времени и является отличительным признаком прошедшего категорического времени.

Отсюда прошедшее категорическое время можно было бы маркировать и как PST_EVID (прошедшее эвиденциальное) - формально это выглядит как нечто непривычное, но такая помета вполне укладывается в русло тех тюркских грамматик, которые прошедшее категорическое называют «прошедшим очевидным» или «прошедшим определенным».

В своей системе обозначений мы выбрали вариант PST_DEF(«прошедшее определенное»).

2.2. Будущее время

Синтетические формы будущего времени в работах по татарской грамматике наиболее часто называются: будущее неопределенное и будущее категорическое.

Будущее неопределенное – выражается с помощью морфемы –[Ы]р: -ыр/-ер/р/-с. В казахской грамматике оно называется – будущим предположительным временем [9]. А в турецком языке с помощью морфемы -[I]r выражается настоящее-будущее время, которое кроме значения неопределенности в совершения действия еще означает и регулярно повторяющееся действие.

Будущее категорическое время – в татарском языке выражается с помощью аффиксальной морфемы –АчАк: -ачак/-әчәк/-ячак/-ячәк. В грамматике турецкого языка оно также называется будущим категорическим временем, а в казахском языке такая категория отсутствует [9].

Дифференциация этих форм связана с модальными значениями определенности/неопределенности и категоричности/некатегоричности.

Наличие осложнения темпорального значения модальными делает возможными конкретизирующие тэги, указывающие на это: FUT_PROB (Probabilitive) или FUT_POTEN (Potentialis).

В базе данных Verbum термин Probabilitive используется для обозначения форм, когда говорящий оценивает ситуацию Р как вероятную - как правило, в будущем ("возможно", "вероятно", "скорее всего"), при этом степень уверенности говорящего специально не различается: от "может быть" до "скорее всего" (Verbum). Конкретизирующий тэг potentialis также указывает на возможность совершения действия.

Предложенная нами, система грамматической разметки для глагольных времен татарского языка включает следующие тэги:

Таблица 1. Темпоральные категории

Обозначение	Наименование	Морфемы
PRES	настоящее время	-Й: -й/-а/-ә
PST_DEF	прошедшее категорическое	-ДЫ: -ды/-де/-ты/-те
PST_INDF	прошедшее результативное (перфект)	-ГАн: -ган/-гән/-кан/-кән
FUT_DEF	будущее категорическое	-АчАК: -ачак/-әчәк/-ячак/-ячәк
FUT_INDF	будущее неопределенное	-[Ы]Р: -ар/-әр/-ыр/-ер/-р/-с

Для использования этих обозначений для других тюркских языков требуется некоторая

доработка.

3. Залоговые категории

Залоговые формы в тюркских языках также являются полисемантичными и практически полностью совпадают во всех трех перечисленных тюркских языках.

В татарских грамматиках для взаимно-совместного залога, выражаемого морфемой –[Ы]ш: -ыш/-еш/-ш выделяют 3 основных значения, которые могут быть обозначены при помощи терминов, используемых в современных работах по типологии:

1. Взаимное значение (реципрок) — «действие представляется не только исходящим от его производителя, но и направленным на него со стороны другого лица» [15]. В этом случае субъект ситуации одновременно является ее объектом.

2. Значение содействия — действие осуществляется в помощь другому лицу, Обычно это глаголы деятельности, а синтаксическая конструкция включает адресата в форме направительного падежа. В базе данных Verbum для значения содействия используется термин Assistive, однако он там обозначает подтип каузативной актантной деривации, когда «субъект каузирует ситуацию P, помогая ее осуществлению ("помогать", "способствовать")» [13].

3. Значение совместности действия - совместность осуществляемых действий (обычно в этом значении функционируют непереходные глаголы.). У В. Плунгяна это тип актантной деривации Associative, когда субъект участвует в ситуации P совместно с другими участниками ("вместе с", "совместно") [13].

Возникает вопрос — какое из этих трех значений считать базовым и зафиксировать в названии? Логически — значение простой совместности Associative, которое можно вычленить в реципроке и ассистиве, но термин Associative практически не используется в тюркологии и известен лишь узкому кругу типологов. С другой стороны, «реципрок», хотя и не покрывает всего разнообразия значений взаимно-совместного залога, используется традиционно и знаком пользователю, поэтому в качестве прототипа был выбран именно этот термин.

Таблица 2. Залоги

Наименование	Морфемы	Обозначение
Действительный (основной) залог	-	ACT
Страдательный залог (пассив)	-Ыл:-ыл/-ел/-л	PASS
Возвратный залог (рефлексив)	-Ын:-ын/-ен/-н	REFL
Взаимно-совместный залог (реципрок)	-Ыш:-ыш/-еш/-ш	RECP
Понудительный залог (каузатив)	-Д[Ыр]: -т/-тыр/-тер/-дыр/-дер	CAUS

4. Императив

Еще одной серьезной проблемой при разработке системы морфологических обозначений стали формы императива и разграничение их от желательного наклонения (оптатива). Как правило, в тюркских языках фиксируется шестичленная парадигма императива, хотя среди тюркологов продолжаются дискуссии о включении в ее состав форм 1-го и даже 3-го лица [16].

Так, формы татарского глагола 3 лица на *–сЫн: барсын ‘пусть пойдет’* могут быть интерпретированы как «императив 3 лица», либо «юссив».

В современной лингвистике используются разные названия для категорий, выражающих

побуждения не второго лица. Так, в известной монографии по типологии императива [14] приняты термины «императив 1-го лица», «императив 3-го лица» . Однако императив 3 лица имеет свою специфику — это повеление дистантное и опосредованное, косвенное, а не прямое, как в случае 2-го лица. Для обозначения косвенного побуждения, адресованного третьему лицу, в современных работах по типологии используется термин «юссив». Юссив «выделяется из других императивных форм тем, что, адресуя побуждение к 3-му лицу, говорящий не имеет возможности контролировать ситуацию, поскольку адресат побуждения находится вне коммуникативной ситуации. Степень контроля над ситуацией при побуждении 2-го лица (слушающего) значительно выше» [6].

С другой стороны, встает вопрос и об интерпретации форм побуждения 1 лица. В грамматиках последних десятилетий выделена категория желательного наклонения [14, 15]. При этом отмечается дефектность этой категории: имеется только форма 1 лица единственного и множественного числа. Но основная проблема заключается в том, что значение желательного наклонения в татарском языке и желательного наклонения (оптатива) в других языках имеют принципиальные несовпадения. Подробнее об этом написано в работе [7].

В более старых работах по тюркологии также прослеживается тенденция рассматривать формы типа *барыйм, барыйк* в парадигме императива. Так, В.Н.Хангильдин не признает синтетических форм желательного наклонения. По его мнению, значение желания передается глаголами повелительного наклонения, выражаемого с помощью аффиксальной морфемы -АйЫм:-ыйм/-им [17].

Формы 2 лица получают пометы императива единственного и императива множественного числа (IMP_SG и IMP_PL); формы 1 и 3 лица обозначаются на основе терминологических прототипов «гортатив» и «юссив» (HOR и JUSS), однако в пользовательском интерфейсе разрабатываемых программ будут предлагаться альтернативные варианты («гортатив», «желательное наклонение» и «императив 1 лица», соответственно – «юссив», «императив 3 лица»).

Таблица 3. Императив

Название	Обозначение	Морфемы
1 лица (гортатив) ед.числа	HOR_SG	-ЫЙм: -ыйм/-им
1 лица (гортатив) мн.числа	HOR_PL	-ЫЙк: -ыйк/-ик
2 лица ед.числа	IMP_SG	-
2 лица мн.числа	IMP_PL	-[Ы]гЫз: -ыгыз/-егез/-гыз/-гез
3 лица (юссив) ед.числа	JUS_SG	-сЫн: -сын/-сен
3 лица (юссив) мн.числа	JUS_PL	-сЫн+ЛАр: -сыннар/-сеннәр

Подобная специфика обозначений: движение от реальных аффиксов, то есть по существу — обозначение (разметка) аффиксов, выведение грамматических категорий и значений через материально выраженные аффиксы, проявляется и при обозначении некоторых других категорий.

Заключение

Эту систему обозначений планируется также использовать в других компьютерных разработках, которые ведутся в институте “Прикладная семиотика”, в частности в татарском электронном корпусе “Туган тел” для осуществления морфологической разметки корпуса.

Предлагаемая система обозначений подходит для представления морфологических категорий глаголов татарского языка, а для возможности ее использования в других тюркских языках требуется небольшая доработка.

Литература

1. *Володин А.П., Храковский В.С.* Об основаниях выделения грамматических категорий // Проблемы лингвистической типологии и структуры языка / Отв. ред. В.С.Храковский. - Л.: Наука, 1977. - С.42-54.
2. *Гильмуллин Р.А., Невзорова О.А., Хакимов Б.Э.* Корпус татарских текстов: проблема репрезентативности // Труды международной конференции «Корпусная лингвистика-2011». 27-29 июня 2011 г., Санкт-Петербург. – СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011. — С. 125-130.
3. *Гузев В.Г., Бурыкин А.А.* Общие строевые особенности агглютинативных языков //Acta linguistica Petropolitana. Труды ИЛИ РАН. - Т. 3. - Ч. 1. - СПб., 2007. - С. 109-117) // http://www.philology.ru/linguistics1/guzev-burykin-07.htm
4. *Кибрик А. Е., Архипов А. В., Даниэль М. А., Кодзасов, Майерс Т., Нахимовский А. Д.* Технологии обработки языковых данных в документировании малых языков // http://www.dialog-21.ru/digests/dialog2007/materials/html/35.htm
5. Краткий справочник по современному русскому языку / под редакцией П.А.Леканта. - М.: Высшая школа. 1995. - 382 с.
6. *Добрушина Н.Р.* Семантическая зона оптатива в нахско-дагестанских языках // Вопросы языкознания. 2009. — № 5. — С. 48-75. hse.ru›data/2011/10/11/1270415648/Dobrushina.pdf.
7. *Добрушина Н.Р.* Оптатив или императив? //Мишарский диалект татарского языка: Очерки по синтаксису и семантике / Под ред. Е.А.Лютиковой, К.И.Казенина, В.Д.Соловьева, С.Г.Татевосова. — Казань: Магариф, 2007. – С. 252-266
8. *Жеребило Т.В.* Фрагмент словаря лингвистических терминов.
9. *Мусаев К.М.* Казахский язык: учебник / К.М. Мусаев; ин-т стран Азии и Африки МГУ имени М.В.Ломоносова. — М.: — Вост. Лит., 2008. — 367с,
10. Национальный корпус русского языка. Семантика // http://www.ruscorpora.ru/corpora-sem.html (дата обращения:).
11. *Невзорова О.А., Салимов Ф.И., Хакимов Б.Э., Гатиатуллин А.Р., Гильмуллин Р.А., Галиева А.М., Якубова Д.Д., Аюпов М.М.* Семантико-грамматическая аннотация в русско-татарской лексикографической базе данных / О.А. Невзорова, Ф.И. Салимов, Б.Э. Хакимов, А.Р. Гатиатуллин, Р.А. Гильмуллин, А.М. Галиева, Д.Д. Якубова, М.М. Аюпов. // Филологические науки. Вопросы теории и практики. – Тамбов: Грамота, 2012. - №7 (18): в 2-х ч. Ч. I., с. 141- 146.
12. *Плунгян В.А.* Общая морфология: Введение в проблематику. - М.: Едиториал, 2003. - 384 с.
13. *Плунгян В.А.* Классификация элементарных глагольных значений, используемых в БД “Verbum” // http://www.mccme.ru/ling/verbum.html
14. Татар грамматикасы. – Т.2. – М.: ИНСАН, Казан: ФИКЕР, 2002. — 448 б.
15. Татарская грамматика: В 3т. Т.2, Морфология / Рос. АН, АН Татарстана, Ин-т яз., лит. и ист. им. Г.Ибрагимова; Казан. Науч. центр; Редкол.: М.З.Закиев и др..—Казань: Татар. кн. Изд-во, 1993.—397 с..
16. Типология императивных конструкций / Отв. ред. В.С.Храковский. - СПб.: Наука, 1992. - 301 с.
17. *Хангилдин В.Н.* Татар теле грамматикасы (морфология һәм синтаксис) / В.Н. Хангилдин; СССР ФА, Казан филиалы, Тел, әдәбият һәм тарих ин-ты.— Казан: Татарстан китап нәшрияты , 1959.—641б.

С. Ж.КАРАБАЕВА, А.И.ИМАНАЛИЕВА

Кыргызский государственный университет строительства, транспорта и архитектуры им. Н.Исанова, Бишкек, Кыргызстан

ИСПОЛЬЗОВАНИЕ ГРАММАТИЧЕСКИХ ПРАВИЛ В ПРОЛОГе

(на примере кыргызского языка)

This article describes the three stages of technology, parsing sentences in the Kyrgyz language. The first stage provides the basic definitions and concepts that allowto record a formal language, as a certain mathematical object. In the second stage we create a model with details of the list word, with each link and their dependence relations. In the third stage we construct a two-stage level of words in the Kyrgyz language.

Настоящая статья описывает три этапа техники разбора предложений кыргызского языка. На первом этапе даются основные определения и понятия, позволяющие записать формальный язык как некоторый математический объект. Навтором этапе мы создаем модель с детали головы списка, каждое звено и их зависимость отношений.На третьем этапе построим двухступенчатый уровень слов кыргызского языка.

Ключевые слова: формальный язык, формальные модели языка, список, цепочка, синтаксис, семантика.

Известный кыргызский ученый Касым Тыныстанов в своих научных исследованиях поставил задачу создать список всех кыргызских слов.

Но, как известно из психологических исследований, если даже очень умный человек попытается вспомнить информацию одного типа, то одна информация повторится, а другая может забыться.

Поэтому, Касым Тыныстанов составил из нескольких вертикальных таблиц кыргызский алфавит (на основе латыни), он выписывал слова, которые получались путем сдвига таблиц. Конечно, чтобы ускорить подбор К. Тыныстанов применил некоторые грамматические особенности.

Ему удалось создать остроумный алгоритм, использовав, который он придумал простое и оригинальное техническое средство, названное «технической таблицей». С помощью этого средства за короткий промежуток времени, Касым Тыныстанов собрал богатый лексический запас кыргызского языка, содержащий около ста тысяч слов.

Известный профессор-полиглот Е. Д. Поливанов в отзыве о научной деятельности Касыма Тыныстанова в 1935 г. писал «…1) начав с задания лексикологического характера, К.Тыныстанов самостоятельно изобрел оригинальный способ (и технический прибор) для исчерпывающего обследования словарного запаса в индивидуальном языковом мышлении (изобретение это может иметь большое теоретическое и прикладное значение); 2) работа над словарем привела К. Тыныстанова к вопросам так называемой морфонологии…» .

В первой половине XX века из идей и трудов таких мыслителей появилась нынешняя наука информатика.

I. Рассмотрим основные понятия теории формальных языков.

Язык (искусственный и естественный) состоит из трех основных компонентов: словаря, синтаксиса и семантики.

Формальные языки, так же как и языки естественные, можно рассматривать с точки зрения их формы, структуры, иначе говоря, синтаксиса, и с точки зрения смысла, вкладываемого в приложения языка, т.е. семантики. Синтаксический анализ формальных языков во многом напоминает известный по школе грамматический разбор предложений.

Словарь языка содержит множество лексем.

Словарь – это конечное множество элементов, называемых символами.

Пусть задан словарь V.

Цепочка над словарем V- это произвольная упорядоченная последовательность символов словаря. Например, V={a, b, c} – это словарь. a=aabc – цепочка, β – bbaaca – другая цепочка.

Пустая цепочка – это цепочка, не содержащая символов (ε).

Пусть V– некоторый словарь. V^*-множество всех возможных цепочек, составленных из символов словаря V, включая пустую цепочку .

Необходимой составляющей описания любого языка является его алфавит, т.е. непустое конечное множество элементов (символов), из которых состоят предложения языка. Будем обозначать алфавит следующим образом:

A=$\{A, B, C, \ldots Y, Z\}$.

Всякая конечная последовательность символов алфавита называется цепочкой (строкой или списки). Допускается существование пустой цепочки (или списки).

Длина цепочки записывается как $|X|$ и равняется количеству символов, составляющих данную цепочку. Таким образом, если ABC, AABB, [] есть цепочки, то $|ABC|$ = 3, $|AABB|$=4, []=0. В заглавные буквы обозначают символы языка, а строчные – цепочки символов, т.е.

$x = ABC;\ \ y = Y;\ \ Z = [\,]$.

Конкатенацией (сцеплением) цепочек x и y называется цепочка Z, состоящая из цепочки ***x*** и дописанной в след цепочкой ***y.*** Конкатенация цепочек *x* и *y* обозначается как *xy*. Если *x=ABC, y=DEF*, то *xy=ABCDEF*. Если *z=xy*, то *z=xy,* то *x-голова, а y –хвост цепочки (списка) z*. Будем обозначать множество цепочек через $\tilde{A}, \tilde{B}$и т.д. Произведением $\tilde{A}\tilde{B}$ двух множеств цепочек $\tilde{A}\ u\ \tilde{B}$ является множество, состоящее из всех комбинацией цепочек $\tilde{A}\ u\ \tilde{B}$.

$\tilde{A}\tilde{B}=\left\{\frac{xy}{x} \in \tilde{A},\ \ y \in \tilde{B}\right\}$

Если $\tilde{A} = \{a, b, ab\},\ \tilde{B} = \{c, d, cd\}$, то

$\tilde{A}\tilde{B}$ =	***c***	***dcd***
a	*ac*	*adacd*
b	*bc*	*bdccd*

ab*abcabdabcd*

Видно, что произведение множеств не коммуникативно, т.е. $\tilde{A}\tilde{B} \neq \tilde{B}\tilde{A}$. Множество, состоящее из пустой цепочки $\emptyset$, можно обозначить, как []. Для этого множества справедливо $[]\tilde{A} = \tilde{A}[] = \tilde{A}$.

Теперь можно определить понятие степени множество цепочек: $\tilde{X}\tilde{X} = \{xx / x \in \tilde{X}\}$ и $\tilde{X}^n$=$\{xxx\ldots x, / x \in \tilde{X}\}$.

Если$\tilde{X}$=$\{x\}$, то можно говорить 0 степени цепочки

x^n=*xxx…x*(*n*раз)

x^0=*[]*.

Аналогично, понятие степени можно распространить и на алфавит

A^0=*{[]}? *A^1=*A,*A^n=AA^{n-1}для*n>0.*

Семантика – совокупность правил интерпретации лексем и языковых конструкций.

Семантическими интерпретации и представления текстов на естественном языке является важной задачей обработки естественного языка. Количество себе семантических представлений были использованы представляют собой естественное язык с помощью концептуальных представлений.

Основная сложность построения транслятора состоит в том, что число возможных программ на входе бесконечно. Транслятор должен обрабатывать любую программу из этого множества, сама программа может иметь любую сложность. Для построения транслятора необходимо иметь ведущую идею, позволяющую вычислить значение любой

входной программы, выражая это значение на выходном языке или в последовательности действий.

II. Одна из таких идей – метод синтаксически-ориентированной трансляции, основанный на работах американского ученого Ноэля Хомского. На основании изучения механизма понимания смысла естественного языка ученый пришел к выводу, что существенную роль в этом процессе играет этап построения структуры предложения, которую, в свою очередь, используют для «вычисления» смысла предложения. Например , «Порядок сменит хаос». Предложение имеет 2 смысла :

1-й – вместо хаоса будет порядок.

2-й – вместо порядка будет хаос.

Из гипотезы Хомского следует, что семантический анализ сводится к синтаксическому и состоит из двух процедур: распознавания структуры входного предложения; построения выходного текста (действий) на основе этой структуры.

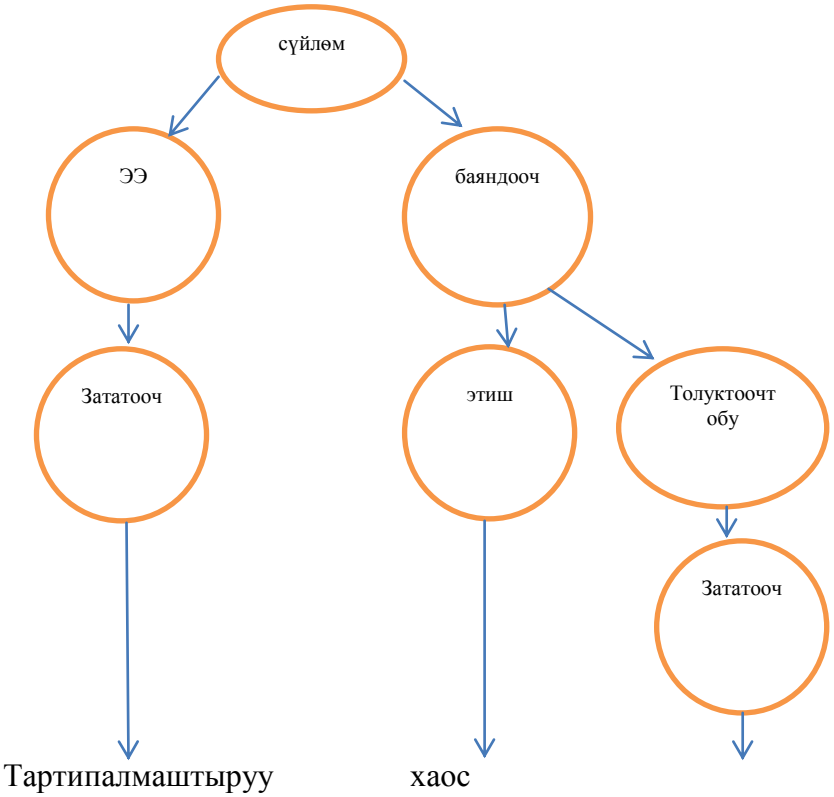


Рисунок 1. Структура предложения – вариант – 1.

Разбор помогает понять отношения между словами в предложении. Он играет важную роль в большом количестве приложений, таких как машинный перевод, неоднозначности смысла слова, поисковые системы, диалоговых систем и т.д. Анализаторы в основном подразделяются на две категории - грамматика и управляемых данными приводом.

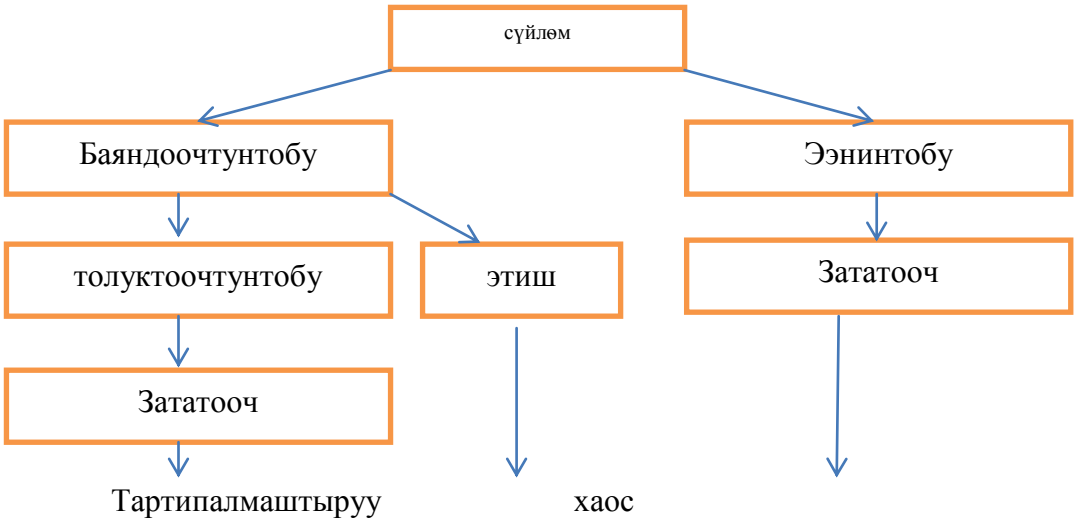


Рисунок 2. Структура предложения – вариант -2.

Чтобы применить математический подход к проблемам, связанным с языками и их обработкой, необходимо ограничиться множество цепочек, которые можно определить некоторым точным образом. Существуют различные способы точного задания таких множеств. Один из них заключается в задании языка как множества, допускаемого каким-нибудь распознавателем цепочек вроде конечного автомата. Другой – в использовании методов, которые можно считать грамматическими.

Термин «формальная грамматика» применим к любому определению формального языка, основанному на грамматических правилах, с помощью которых можно порождать и анализировать цепочки аналогично тому, как грамматики используют при изучении естественных языков.

III. Рассмотрим формальную грамматику, которая в какой-то степени напоминает фрагмент грамматики кыргызского языка и задает формальный язык, состоящий из четырех кыргызских предложений. Такой грамматике используют элементы, играющие роль членов предложение или частей речи:

Синтаксис – совокупность правил построения языковых конструкций (предложений) из лексем.

Синтаксический анализ является очень важным приложением Пролога и логического программирования. В действительности, происхождение Пролога связано с попыткой использовать логику для выражения грамматических правил и формализации процесса синтаксического разбора.

Элементы, приведенные в грамматике, такие как <подлежащие> или <существительное>, играющие роль членов предложения или частей речи, называют нетерминальными (вспомогательными) символами, или нетерминалами. При определении языка программирования нетерминалами служит такие элементы, как <оператор>, <выражение> и т.д.

Не терминалы – это конструкции языка.

Грамматика может содержать любое количество терминалов. В языках программирования терминалами являются используемые в них слова и символы, BEGIN, DO, + и т.д. Терминалы – это символы предложений порождаемого языка.

Наиболее распространенным подходом к реализации синтаксического разбора средствами Пролога является использование грамматик, задаваемых определительными предложениями (definiteclausegrammar, DCG). DC-грамматики являются некоторым обобщением контекстно-свободных грамматик. Контекстно-свободные грамматики определяются множеством правил вида

<нетерминал>→<тело>,

Где нетерминал является нетерминальным символом, а тело-последовательностью из одного или нескольких элементов, разделяемых запятыми. Каждый элемент–это либо нетерминальный символ, либо последовательность терминальных символов. Смысл правила в том, что тело есть возможная форма грамматической группы нетерминального типа. Нетерминальные символы записываются как атомы Пролога, а последовательности терминальных символов - в виде списков атомов. Это облегчает трансляцию грамматик в Пролог-программы.

Для каждого нетерминального символа S грамматика определяет язык, который представляет собой множество последовательностей терминальных символов, получаемых путем повторного недетерминированного применения правил грамматики, начиная с символа S.

Рассмотрим простую контекстно-свободную грамматику для небольшого подмножества кыргызского языка. Первое правило грамматики читается так: “Предложение состоит из группы существительного, за которым следует группа глагола”.

```
сүйлөм(с(Q,V)) -->зататооч(Q), этиш(V).
зататооч(з(кыз))--> [кыз].
зататооч (з(китеп))--> [китеп].
зататооч(з(кагаз))--> [кагаз].
зататооч(з(кой))--> [кой].
зататооч(з(ит))--> [ит].
этиш(эт(секирди)) --> [секирди].
этиш(эт(бакырды)) --> [бакырды].
этиш(эт(иштеди)) --> [иштеди].
этиш(эт(чуркады)) --> [чуркады].
1 ?- сүйлөм(S,L,[]).
S = с(з(кыз), эт(секирди)),
L = [кыз, секирди].
… … … … … …
S = с(з(ит), эт(чуркады)),
L = [ит, чуркады].
```

Грамматические правила: предложения -- >группа_существительного, группа_глагола.

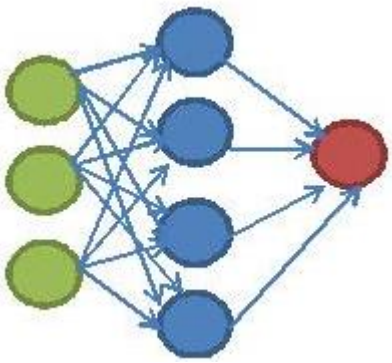


Рис.3. модель простой контекстно-свободной грамматики

Примером для реализации восходящей грамматики с рекурсивными правилами является синтаксический анализатор математических выражений. Такая система начинает работу с данными и переходит к простым синтаксическим объектам, а затем и к более сложным. В то время как система нисходящего разбора управляется в основном гипотезами, система восходящего грамматического разбора управляется данными.

Разбор предложения можно рассматривать как нахождение зависимости отношения некоторой парыслов в предложении. Слова должны быть связаны таким образом, что они образуют ровную древовидную структуру, где узлы являются словами, ребра назначаются между парами слова,которые являются связанными уровнем именем отношений.

Заключение

Пролог обладает большими возможностями по сопоставлению объектов с образцом, поэтому данный язык программирования хорошо подходит для обработки текстовой информации. В статье приведен пример для разбора предложений (в программе Swi-Prolog).

Система грамматического разбора – это программа, которая распознает синтаксические объекты в потоке лексем, т.е. реализует какую-либо формальную грамматику. Каждый язык имеет ряд особенностей, которые требуют особого внимания в формальные описания.

Таким образом, мы предлагаем последовательную древовидную структуру для кыргызского языка, принимая во счет языковые возможности, чтобы гарантировать эффективные семантические и синтаксические разбора и утверждаем, что дерево семантических классов (структура предложения – вариант -2.) является универсальной для классификации всех языков кроме английского.

Литература

1. Тыныстанов К. Лексикон кыргызского языка. – 1934// КР ИУА Кол жазмаларбөлүмү. Инв. 177/37-182/42, кол жазмалар 79-84.
2. Панков П.С. Обучающая и контролирующая программа по словоизменению в кыргызском языке на ПЭВМ. – Бишкек: Мектеп, 1992. – 20 с.
3. Карабаева С.Ж. Синтаксический анализ текста в логическом программировании // вестник КазНПУим.Абая. – Алматы. 2008. – №2(22) – С.213-218.

У.А. ТУКЕЕВ, Д.Р. РАХИМОВА, К. БАЙСЫЛБАЕВА, Н. УМИРБЕКОВ, Б.ОРАЗОВ, М. АБАҚАН, С. КЫЗЫРКАНОВА.

Әл Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

КӨПМАҒЫНАЛЫҚ БЕЙНЕЛЕУ КЕСТЕ ТӘСІЛІ НЕГІЗІНДЕ ОРЫС ТІЛІНЕН ҚАЗАҚ ТІЛІНЕ МАШИНАЛЫҚ АУДАРМАСЫНЫҢ МОРФОЛОГИЯЛЫҚ АНАЛИЗБЕН СИНТЕЗІН ҚҰРУ

Кіріспе

Машиналық аудару жүйесі күрделі және өте ауқымды болып табылады: құрамына екі тілді және одан да көп сөздіктерді қосады, ол қажетті грамматикалық ақпаратпен жабдықталған (морфологиялық, синтаксистік және семантикалық), ол эквиваленттті, нұсқалық және трансформациялық аударма сәйкестіктерін, сонымен қатар грамматикалық талдаудың алгоритмдік құралдарын жіберуді қамтамасыз ету үшін қажет. Осы мақалада орыс тілінен қазақ тіліне және керсінше машиналық аудармадағы морфологиялық талдау мен синтездің жеке жағдайы қарастырылады.

Сөздер өзара байланыспай, бір-бірімен тіркеспей тұрғанда ойды және нақты мағынаны білдіре алмайды. Сөздердің бір-бірімен байланысуы және мағынаның толық болуы аффикстердің жалғауы арқылы іске асады.

Қазақ тіліне де, орыс тіліне де жалғанатын жұрнақтар мен жалғаулардың өзіндік жалғану реті мен жүйесі бар [1]. Бұл тілдің ішкі заңдылығына бағынатын күрделі тарихи процесс. Түбір мен қосымшалардың мағыналарының берілуінің де өзіндік тәртібі болады. Қазақ

тіліндегі жалғаулар 4 топқа бөлінетін болса: көптік, септік, жіктік, тәуелдік, орыс тіліндегі жалғаулар «род, число, падеж (септік)» сияқты грамматикалық мағынада бөліне алады. Соның ішінде орыс тілі мен қазақ тіліндегі септік жалғауларға қатысты мәселерді қарастырсақ. Орыс тіліндегі «книга» сөзін септеп көрейік: именетилеьный падеж- книга, родительный падеж - книги, дательный падеж - книге, венительный падеж - книга, творительный падеж - книгой, предложный падеж - о книге. Бұл жерде орыс тілінен қазақ тіліне машиналық аударма кезінде қиындық туындайды себебі, екі септікте де «е» жалғауы кездеседі және «и» жалғауы «родительный» септігінде және атау септіктің көпше түрінде де кездесе алады. Мысалы: «книги нет на месте» және «я забрала все свои книги». Бірінші жағдайда бір ғана кітап жайлы және сол кітаптың орнында жоқтығы жайында сөз қозғалса, екінші жағдайда бірнеше кітап және соларды алып кеткендігі жайлы баяндалады. Бұның мағынасын біз жеңіл түсінгенімізбен, машинаның түсінуі екіталай.

Машиналық аудармадағы талдау тізбектелген және өзара байланасатын процесс болып саналады. Морфологиялық талдаудың нәтижесі синтаксистік талдауға әсерін тигізеді, ал морфология пен синтаксистік мәтіннің семантикасына ықпал тигізеді. Сондықтан мәтін талдауының бастапқы кезеңіне дұрыс көңіл бөлуді қажет етеді.

Біздің зерттеу жұмысымыздың негізгі бағыты: қазақ жалғауларының (аффикстерінің) түрлері мен байланыстардың атқаратын рөлі және мәтіннің аудармасына тигізетін әсерін көрсету. Орыс-қазақ машиналық аудармадағы морфологиялық талдаудағы қазақ тілінің жалғауларын (аффикстерін) қолайлы жүйеге түрлендірудің әдісін ұсыну болып табылады.

1 Көпмағыналық бейнелеу кесте тәсілі

Машиналақ аудармадағы негізгі мәселердің бірі көпмағыналық болып саналады. Оған қатысты көптеген әдістер бар. Қазіргі қолданыста ең көп тарағаны статистикалық және ережелер негізіндегі әдістер болып табылады. Бірақ олардың өзіндік артықшылықтары мен кемшіліктері бар.

Бұл жұмыста біз табиғи тілдерді бір біріне машиналық аударма күрделі мәселелерін шешу жолдарын іздеу бағытында келесі пікір(ұйғарым) еңгізіп жатырмыз. Әр сөзді тек екі бөлімнен құрылады деп санаймыз – негізден және жалғаулардан. Осы ұйғарымды компьютерлік көрсетуілінде біз **көпмағыналық бейнелеу кесте** түрінде қолдануын ұсынып жатырмыз. Көпмағыналы бейнелеу теория соңғы 10-15 жылдары қатты дамыған [2,3]. Машиналық аударма бағытында **бейнелеу** тәсілі қолданған мысалдары бар [4,5], бірақ **көпмағыналы бейнелеу** тәсілдерін қолданғанын қарастыра алмадық.

Көпмағыналық бейнелеу кестелер машиналық аударма процесін жүйректендіруге және жүйелендіруге ықпал тигізеді.

Машиналақ аударманы жасау барысында деректер қорына көпмағыналық бейнелеу кесте құру тәсілімен тілдер сөздігі мен қосалқы элементтер (жұрнақ, жалғау) кестелері құрылады. Қолайлық үшін орыс-қазақ электрондық сөздігін сөз табтарының түрлеріне сай кестелерге бөлуді ұйғардық. Электрондық сөйлем сөздігі тек қана сөйлем негізінен (түірінен) құралады. Және кестелерге әр сөз табына сәйкес қасиеттерді қосалқы атрибуттар ретінде еңгіземіз. Кестелердегі атрибуттар қарапа йым сандар ретінде белгіленеді. Мысалы: орыс тілдегі зат есімге "род", "число", "склонение"; қазақ тілінің етістіктеріне уақыт шағы, жақ және т.с.с.

Сөздер өзара байланыспай, бір-бірімен тіркеспей тұрғанда ойды және нақты мағынаны білдіре алмайды. Сөздердің бір-бірімен байланысуы, тіркесуі аффикстердің жалғауы арқылы іске асады. Грамматикада сөздер байланысы негізгі бес түрге (қиысу, меңгеру, матасу, қабысу, жанасу)бөлінеді , соның үшеуі (қиысу, меңгеру, матасу) жалғаулар негізінде құрастырылған. Соған қарағанда сөйлемдегі сөздердің мағыналық байланысы аффикстерге тәуелді.

Қазақ тіліне де, орыс тіліне де жалғанатын жұрнақтар мен жалғаулардың өзіндік жалғану реті мен жүйесі бар. Бұл тілдің ішкі заңдылығына бағынатын күрделі тарихи процесс. Түбір мен қосымшалардың мағыналарының берілуінің де өзіндік тәртібі болады.

Қазақ тіліндегі жалғаулар (аффикстер) 4 топқа бөлінетін болса: көптік, септік, жіктік, тәуелдік. Орыс тіліндегі жалғаулар «род, число, падеж (септік)» сияқты грамматикалық мағынада бөліне алады. Жалғаулардың сөз құрамына және құрылуына әсер ететіні мәлім. Орыс тілінен қазақ тіліне машиналық аударманы жасау барысында морфологиялық талдауда және синтезде көптеген мәселелерге және қиыншылықтарға тап болдық. Машиналық аударамаға морфологиялық талдауы мен синтезі үшін аффикстер көпмағыналық бейнелеу кестелер тәсілімен құрылады. Оларға да сәйкесінше қасиеттік атрибуттар еңгізіледі.

Мысалы, орыс тілі етістік сөздердің жалғаулар кестесінің құрылуын келесі суреттен көруге болады.

RecNo	id	okonch	chislo	vremia	lico	ch_r
Click here to define a filter						
1	1	ю	1	1	1	2
2	2	ешь	1	1	2	2
3	3	ишь	1	1	2	2
4	4	ет	1	1	3	2
5	5	ем	2	1	1	2
6	6	ете	2	1	2	2
7	7	ите	2	1	2	2
8	8	ют	2	1	3	2
9	9	л	1	2	4	2
10	10	ла	1	2	4	2
11	11	ло	1	2	4	2
12	12	ли	2	2	4	2
13	13	ать	5	5	5	2
14	14	ыть	5	5	5	2
15	15	ить	5	5	5	2
16	16	еть	5	5	5	2
17	17	ять	5	5	5	2
18	18	уть	5	5	5	2
19	19	у	1	1	1	2
20	20	лю	1	1	1	2

1 сурет. Орыс тілі етістіктердің аффикстер кесетесі.

Жоғары суреттегі етістік аффикстеріне төрт атрибут (число, время, лицо, часть речи) жалғанған, осы атрубуттарға сәйкес машиналық аударманың екінші (аудару) тілінің аффиксі ізделінеді. 1 суретте көрсетілгендей талдау және іздеу кезінде қызылмен ерекшеленген көпмағыналық бейнелеулерге тап боламыз.

Ал аударма процесінде көпмағыналы жағдайда біз бірмағыналы шешімді табуымыз қажет. Сол себептен әр көпмағыналы жағдайға әдейі арналған бірмағыналы шешімдер шығаратын ережелер құрамыз.

2 Орыс және қазақ тілдерінің септілік сәйкестендіру кестелері

Қазақ тілін орыс тілімен салыстырғанда септіктердің мағынасына қарай салыстыру жүйесін жасадық. Бірақ мұнда өзіндік қиыншылықтар туындады, себебі орыс тілінде алты, ал қазақ тілінде жеті септік бар және олар бір біріне сәйкес емес. Олардың сөз түрлендіруі бір бірінен өзгеше. Қазақ тілінің 7 септік түрі бар және кестеге еңгізу үшін біз оларға реттік нөмірлерді бердік: атау-1, ілік-2, барыс-3, табыс-4, жатыс-5, шығыс-6, көмектес-7. Орыс тілімен қазақ тілін сәйкестендіре отырып септіктің байланысын орнаттық (1 кестеде бейнеленген).

Кесте 1. Екі тіл септіктерінің сәйкестігі.

Орыс тіліндегі септік түрлері	**Қазақ тіліндегі септік түрлері**
Именительный падеж	Атау септік
Родительный падеж	Ілік септік
Дательныый падеж	Барыс септік
Винительный падеж	Табыс септік

Творительный падеж	Көмектес септік
Из (предлог)+ родительный падеж	Шығыс септік
От (предлог)+ родительный падеж	Шығыс септік
На (предлог)+ предложный падеж	Жатыс септік
В (предлог)+ предложный падеж	Жатыс септік

3 Орыс тілінен қазақ тіліне машиналық аудармасының морфологиялық талдау мен синтезін шешу

Орыс тілінен қазақ тіліне машиналық аударма бағытында септіктерге қатысты мәселеларды шешу үшін біз **көпмағыналық бейнелеу кестелерін** енгіздік, яғни деректер базасы кестелерден тұрады және әрбір кестеде өзіндік атрибуттары болады. Әр кестеде атрибуттар тізімі екі топқа бөлінеді: ену тобы және шығу тобы. Мысалға орыс тілінен мынадай бір сөз тіркесі келіп түсті делік: «гордиться городом» - қаламен мақтану. Алдымен жалғауды орысша сөз үшін қарастырады, ол орыс тілінің септіктерінің жалғауларынан қарайды. Кестеде 5 суретке сәйкесінше былайша болады:

RecNo	id	okonch	ch_r	padezh	chislo	skl
Click here to define a filter						
30	30	я	1	2	1	2
31	31	у	1	3	1	2
32	32	ю	1	3	1	2
33	33	<null>	1	4	1	2
34	34	о	1	4	1	2
35	35	е	1	4	1	2
36	36	ом	1	7	1	2
37	37	ем	1	7	1	2
38	38	ём	1	7	1	2
39	39	е	1	5	1	2
40	40	ы	1	1	2	2
41	41	и	1	1	2	2
42	42	я	1	1	2	2
43	43	ов	1	2	2	2
44	44	ей	1	2	2	2
45	45	ам	1	3	2	2
46	46	ям	1	3	2	2
47	47	ы	1	4	2	2
48	48	ей	1	4	2	2
49	49	я	1	4	2	2
50	50	ы	1	6	1	1
51	51	и	1	6	1	1

Сурет 2. Орыс тіліндегі «творительный» септігін анықтау мысалы

Бұл бейнелеу кестеде ену атрибуттар тобы- «okonch» (орыс тілі жалғауы), ал шығу атрибуттар тобына қалғандары кіреді- ch_r, padezh, chislo, skl. Бұл жерден «город+ом» жалғауын тапты, енді оның атрибуттарын қарастырады, ch_r- сөз табы, okonch- жалғау, padezh- септік. Сол сандар, яғни негізгі сипаттаушы параметрлер бойынша енді қазақ тілінің жалғаулар кестесіне барады(Сурет 3):

RecNo	id	okonch	ch_r	padezh	chislo	dauysty	dauyssyz
Click here to define a filter							
44	44	терде	1	5	2	2	3
45	45	нан	1	6	1	1	2
46	46	нен	1	6	1	2	2
47	47	дан	1	6	1	1	1456
48	48	ден	1	6	1	2	1456
49	49	тан	1	6	1	1	3
50	50	тен	1	6	1	2	3
51	61	лардан	1	6	2	1	14
52	62	лерден	1	6	2	2	14
53	63	дардан	1	6	2	1	256
54	64	дерден	1	6	2	2	256
55	65	тардан	1	6	2	1	3
56	66	терден	1	6	2	2	3
57	67	мен	1	7	1	1	1245
58	68	бен	1	7	1	2	6
59	69	пен	1	7	1	1	3
60	70	лармен	1	7	2	2	14
61	71	лермен	1	7	2	1	14
62	72	дармен	1	7	2	2	256
63	73	дермен	1	7	2	1	256
64	74	тармен	1	7	2	2	3
65	75	термен	1	7	2	1	3

Сурет 3. Қазақ тілінің көмектес септігіне сәйкестік мысалы

Бұл кестеде шығу атрибуты – ‘okonch’(қазақ тілі жалғаулары), ал ену атрибуттар тобы-қалған атрибуттар. 3 суреттен байқап отырғанымыздай бұл жерден орыс тіліндегі «ом» жалғауының {1 7 1} деген сандары (атрибуттары) қазақ тілінің «мен»-{ 1 7 1} деген жалғауымен сәйкес келіп тұр. Сондықтан осыны алып қазақ тіліне аударылған түбір сөзге жалғайды. Сол уақытта «город**ом**»- «қала**мен**» деп дұрыс аударманы шығарады.

Келесі кезекте етістік үшін морфологиялық синтезді қарастыралық. Бұл жағдайдағы жалғаудың жалғану заңдылығы да дәл сондай принциппен жұмыс істейді. Мысалы кіріс тіліндегі сөз тіркесін тағы да «гордиться городом» деп алайық. Бұл жағдайда «гордиться» сөзі (етістігі) етістіктің алғашқы формасында (неопределенная или начальная форма глагола) тұр. Ал етістіктің алғашқы формасы қазақ тіліндегі тұйық етістік ұғымымен сай келеді. Ондай сөздерді базаға дәл сол қалпында енгізуді жөн көрдік. Себебі олардың түбірі кейбір жағдайларда өзгеріп отырады, мысалы «гордиться» сөзі «горжусь» деген формада да бола алады, ал бұл істі қиындата түседі, сол үшін алғашқы формасы базада тұрады. Яғни аударма мынадай түрде болады: «гордиться городом» - қала**мен** мақтан**у.**

Басқаша мысалды алып қарайтын болсақ, «читаешь книгу». Бұл жерде зат есімнің жалғауын қалай анықтайтындығы жайлы жоғарыда сипаттап өттік, енді етістігін қалай анықтайдығын айтайық. Сөз тіркесі келіп түскеннен соң, базадан сөздерді іздейді, сол кезде «читаешь» сөзінің етістіктер кестесінде жатқандықтан етістік екендігін анықтап біледі. Алайда кестеде бұл сөздің тек түбірі ғана болады: «чита». Сол сөздің аудармасының негізі «оқ» екендігін тапқаннан соң, 4-суретке сәйкес

RecNo	id	rus	osnova_rus	kaz
Click here to define a filter				
698	1034	разрядить	разряд	жасандыр
699	1035	наряжаться	наряжа	жасан
700	1036	лежа	леж	жат
701	1037	советовать	совет	кеңес ет
702	1038	читать	чита	оқ
703	1039	помнить	помн	еске ал
704	1040	взять	возьм	ал
705	1041	воевать	вою	соғыс
706	1042	воровать	ворова	ұрла

Сурет 4. Етістіктер сөздігінің кестесі

«ешь» жалғауын жалғаулар кестесінен іздейді, ол 5-суретте келтірілген:

RecNo	id	okonch	chislo	vremia	lico	ch_r
Click here to define a filter						
1	1	ю	1	1	1	2
2	2	ешь	1	1	2	2
3	3	ишь	1	1	2	2
4	4	ет	1	1	3	2
5	5	ем	2	1	1	2
6	6	ете	2	1	2	2
7	7	ите	2	1	2	2

Сурет 5. Орыс тілі етістіктің жалғаулар кестесі

Көріп тұрғанымыздай «ешь» жалғауын тапты, енді оның атрибуттары бойынша, яғни okonch (жалғау)-«ешь», zhak (жақ)- 2, shak (шақ)- 1 және tur(түр)- 1 «жекеше» екендігін тауып алды. Сол бойынша etistik_kaz кестесінен дәл сондай сандарды (атрибуттарды) іздейді. Ондай атрибуттар саны 6-суретке сәйкес екеу болып тұр:

RecNo	id	okonch	zhak	shak	tur	dauysty	dauyssyz
Click here to define a filter							
1	1	мын	1	1	1	<null>	<null>
2	2	мін	1	1	1	<null>	<null>
3	5	пын	1	1	1	<null>	<null>
4	6	пін	1	1	1	<null>	<null>
5	7	сың	2	1	1	1	4
6	8	сің	2	1	1	2	4
7	9	мыз	1	1	2	<null>	<null>
8	10	міз	1	1	2	<null>	<null>
9	11	быз	1	1	2	<null>	<null>

Сурет 6. Қазақ тілі жалғаулары бойынша екі сәйкестік табылған жағдай

Алайда ол атрибутқа сай келетін жалғау екеу болып тұр. Оның бірі «сың» сингармонизм заңы бойынша буын жуан болған жағдайда, ал екіншісі «сің» буын жіңішке болған жағдайда жалғанады. Соны компьютерге түсінікті ету үшін қосымша тағы бір атрибуттар қосылады.

Бұл жердегі дауысты 1- ол жуан, дауысты 2- жіңішке, ал дауыссыз 4- «р, й, у» әріптерінен соң осындай жалғау жалғанатындығын көрсетуші. Ол әріптер комбинациясы программада енгізілген. Яғни осы ережелер бойынша «ешь» жалғауына қазақша сәйкесі «сың» болады. Сондағы алатын нәтижеміз «чита**ешь** книгу»- «кітап оқып отыр**сың**». Бұл жерде тағы бір ескертетін жайт, қазақ тілінде көмекші етістік деген ұғым бар. Ол негізгі етістікке көсемше жалғауын тіркеген соң қосылады. Ал көсемше жалғаулары программада айтылады. Жуаннан кейін «ып», жіңішкеден кейін «іп».

Біздің деректер базасында, зат есімдерге жалғанатын көптік жалғаулар кестесі екеу, бірі орыс тілінің көптік жалғаулары және екіншісі қазақ тілінің көптік жалғаулары. Біздің мақсатымыз кіріс тілінде келген жалғаудың көптік екендігін анықтап, оны атрибуттар арқылы қазақ тіліндегісімен сәйкестендіру. Мысалы мынадай сөз тіркесі берілсін: «книги лежат на полке». Бұл жердегі «книги» сөзінің көптік жалғау ма я септік жалғауы ма екендігін алдымен программа анықтап алуы керек. Егер, сөз тіркесі не сөйлемдегі етістік көпше түрде тұрса, онда бұл жалғау да сәйкесінше көптік жалғауы болады. Ал біздің жағдайда «леж**ат**» сөзіндегі жалғау көпше түрдегі заттың әрекетін білдіріп тұр. Сол себепті зат есім де көпше түрде. Мұны анықтап алғасын енді енді «книг» сөзін кестелерден жүгірте шығып, іздеп

тапқасын аудармасын аламыз. Кейіннен оның жалғауы «и» тағы кестелерден жүгірте отырып ізделінеді, табылғасын оның атрибуттарын қарап 3-суреттедгідей, қазақша балама жалғауды атрибуттар сәйкестігі бойынша табамыз:

RecNo	id	okonchanie	ch_r	padezh	chislo
Click here to define a filter					
1	1	и	1	1	2
2	2	ы	1	1	2
3	3	я	1	1	2
4	4	а	1	1	2

(а)

RecNo	id	okonchanie	ch_r	padezh	chislo	dauysty	dauyssyz
Click here to define a filter							
1	1	лар	1	1	2	1	3
2	2	лер	1	1	2	2	3
3	3	дар	1	1	2	1	2
4	4	дер	1	1	2	2	2
5	5	тар	1	1	2	1	1
6	6	тер	1	1	2	2	1

(ә)

Сурет 7. Орыс және қазақ тіліндегі көптік жалғаулар кестесі.

Енді қазақ тілінен {1 1 2 } атрибуттарымен жалғау іздейміз. Байқап отырғанымыздай бұл жерде барлық жалғау біз іздегендей атрибуттар тұр. Бұл жердегі дауысты 1- дауысты жуан, дауысты 2- дауысты жіңішке, дауыссыз 1- «б,в,г,д» және барлық қатаң дауыссыздар, дауыссыз 2- «м,н,ң» дауыссыздары деп белгіленген

Яғни бізге келіп түскен «книг**и**» сөзін не «кітаплар» немесе «кітапдер» деп емес, «кітап**тар**» деп дұрыс аударып беруі керек. Осылайша қажетті әрі дұрыс аударма шығарылады: «кітап**тар** сөреде жатыр».

Осындай салдарда лингвистика заңдарына жүгінеміз. Жалғауды дұрыс тіркеу үшін сингармонизм заңына бағынамыз. Қысқа айтып кетсек: сингармонизм заңы бұл сөздің түбірі қай әріпке аяқталанытына байланысты жалғауға қолданылатын үндестік ережелер жиынтығы. Мұндай ережелерді автоматандыруға қиындық соқпайды , себебі қазақ тілінің дыбыс үндестігі заңы тұрақты және барлық түркітілдес тілдерге тән.

Жоғарыда көріп тұрғандай аффикстердің жалғануы морфологиялық қасиеттерінен қана емес, сонымен қатар тіркеленетін сөздің түбіріне де байланасты. Бірақ кейде мұндай ақпарат жеткіліксіз болып қалады, себебі аффикстердің жалғануы мәтіндегі басқа тәуелді сөздерге де байланысты. Көрсетілген мысалда сөздің жалғауын дұрыс анықтау үшін «леж**ат**» етістістіктің көмегіне жүгіндік. Сонымен қатар бізге тек қана іс-әрекет байланысы ғана емес тұлғасын (субьект \ обьект) да иелендіру қажет. Мысалы ретінде жіктік және тәуелділік жалғауларды қарастыруға болады.

Орыс тілінде жіктік жалғау болмайды, тек қазақ тіліне аударма кезінде ғана жаққа байланысты жалғау қосылады, мысалы «я студент»- мен студент**пін** сол сияқты 2-3 жақтар үшін де түрлі формада болады. Оны анықтау үшін деректер базасындағы индекстік файл кестесін пайдаланамыз. Келген есімдіктің атрибуттарын тауып аламыз да 8-суретте келтірілгендей, сол бойынша сәйкес келетін жалғауды табамыз:

RecNo	id	rus	kaz	chislo	lico	padezh
Click here to define a filter						
1	1	я	мен	1	1	1
2	2	ты	сен	1	2	1
3	3	он	ол	1	3	1
4	4	она	ол	1	3	1
5	5	оно	ол	1	3	1
6	6	мы	біз	2	1	1
7	7	вы	сендер (сіздер)	2	2	1
8	8	они	олар	2	3	1

Сурет 8. Жіктік жалғауы және оның атрибуттары

«я» есімдігінің атрибуттары {1 1 1} екендігін анықтағасын енді кестесіне барып {1 1 1} іздейміз:

RecNo	id	zhalgau	chislo	lico	padezh	dauysty	dauyssyz
Click here to define a filter							
1	1	мын	1	1	1	1	1
2	2	мін	1	1	1	2	1
3	3	бын	1	1	1	1	<null>
4	4	бін	1	1	1	2	<null>
5	5	пын	1	1	1	1	3
6	6	пін	1	1	1	2	3
7	7	сың	1	2	1	1	123456
8	8	сің	1	2	1	2	123456

Сурет 9. Қазақ тіліндегі бірнеше сәйкес жіктік жалғаулары табылған жағдай

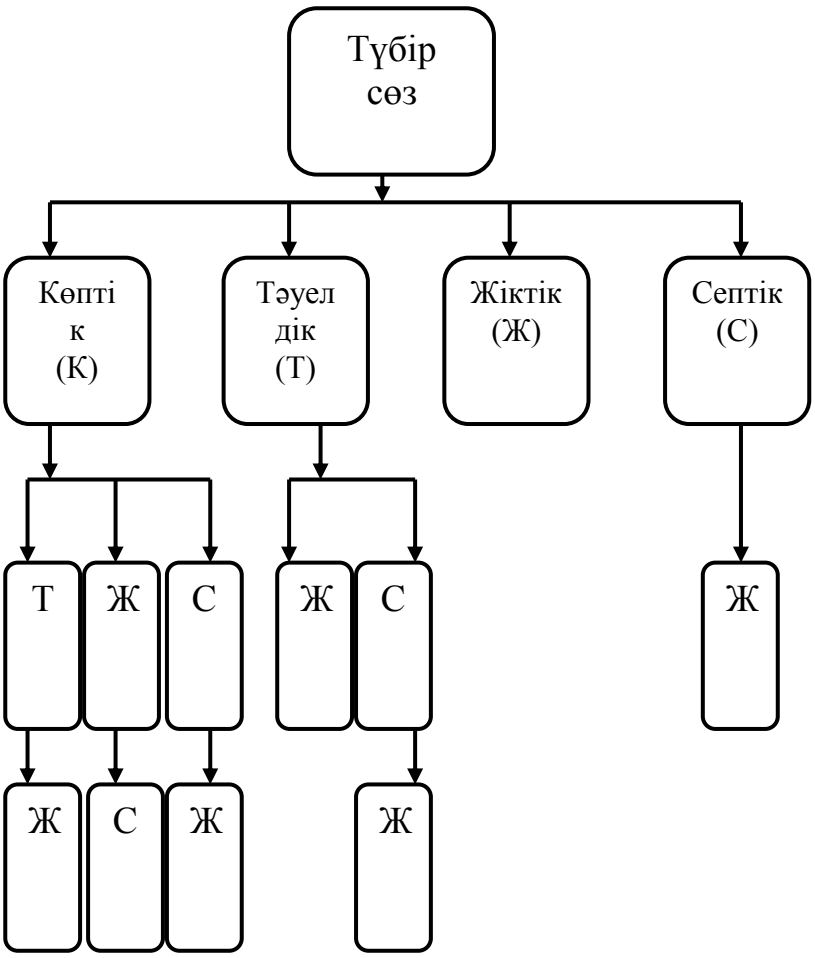
Ол атрибут алғашқы алты жалғаудың көпмағыналы бейнелеу болып, барлығында бірдей екендігі 9-суретте көрсетілген. Оларды ажырату мақсатында жоғарыда айтылып кеткен сингорманизм заңының ережелеріне сүйеніп қосымша дауысты дауыссыз атрибуттары (бағаналары) қосылды. Сол бойынша қажеттісі таңдалып алынады, біздің жағдайда ол «пін», себебі «студент» сөзі қатаң дауыссызға аяқталып тұр. Яғни алатын аудармамыз «я студент»- «мен студент**пін**»

Тәуелдік жалғау, әдетте, бір заттың басқа бір затқа тәуелді екенін білдіретін қосымша. Негізінде зат есімге тән қосымша бола тұрса да, зат есім қызметін атқаратын, демек, субстантивтенетін (заттанатын) сөздердің барлығына да жалғана береді. Бұл қосымшалар жалғанған сөздер, әдетте, өздерінен бұрын ілік септік жалғауда тұрып тіркесетін жіктеу есімдіктермен тікелей байланысты болады. Сол себептен тәуелдеулі сөздің жақ жалғаулары да жіктеу есімдіктерінің жақтарына сәйкес келіп отырады. Мысалы: менің қаламым; сенің қаламың; сіздің қаламыңыз, оның қаламы.Осылайша сәйкес әрі қажетті жалғау таңдалып алынып, қазақшаға аударылған түбір сөзге жалғанады. Біздің жағдайда «**моя** парта»- «менің парта**м**»

Жалпы келгенде машиналық аудармада аффикстер сөз құру және түрлендіру ролін атқарып сөз арасындағы байланысты орнататын құралдардың бірі болып табылады. Осындай маңызды морфологиялық процесске көпмағыналы бейнелеу кесте әдісі өте қолайлы болды.

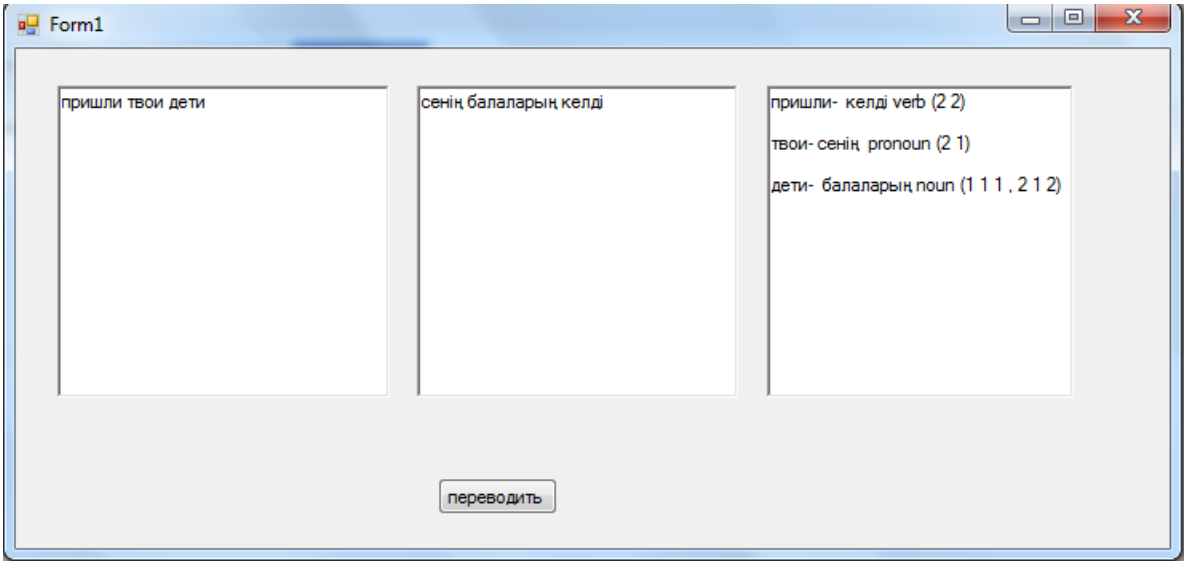
4 Көпмағыналық бейнелеу кестелер арасындағы байланыс

Көпмағыналы бейнелеу кестелердің құрылуы мен қолдануы жоғары тарауларда көрсетілген. Осы кестелер бойынша алынған мәлімметтерді машиналық аудармада шығыс тіліне дұрыс генерациялау білуі қажет. Морфологиялық талдау және синтез кезінде сөзге осындай көпмағыналы бейнелеу кестелер бірнешеу табылуы мүмкін. Себебі орыс тілінің аффикстері көп емес боландықтан сөздің лексикалық та, грамматикалық та мағынасы біріктірілген. Ал қазақ тілінің сөздер құрылымында екі немесе одан да көп аффикстер жалғануы мүмкін. және қазақ аффикстерінің өзіндік жалғану заңдылығы бар. 10-суретте келтірілген диаграммада қазақ тіліндегі жалғаулардың түрі және бір біріне тіркесуінің мүмкін болатын барлық жағдайлары келтірілген.



Сурет 10. Қазақ тіліндегі жалғаулар классификациясы мен комбинациясы.

Бұл өте қажетті талдаулардың бірі, себебі сөйлемді я тіпті сөзді біз белгілі бір заңдылықтарға сүйене отырып қана құрай аламыз, мысалы, септік жалғаудан кейін көптік жалғауды тіркей алмаймыз. Септік жалғаудан кейін тек қана жіктік жалғауы қойылады т.б. Мысалы: «пришли твои дети» мәтінін қазақ тіліне аударып, сөздердің жалғауларына талдау жасайық.



Сурет 11. Орыс тілінен қазақ тіліне машиналық аударманың мысалы

Қорытынды

Біз орыс тілінен қазақ тіліне машиналық аудармадағы морфологиялық талдау және синтез мәселелерін қарастырдық. Сонымен, көпмағыналық бейнелеу машиналық аудармаға көпмағыналық мәселелар туғызады, оларды шешу жолдары жоғарыда көрсетілгендей қосымша атрибуттарды еңгізу, немесе қосымша осы мәселелерді шешетін процедуларды құруды қажет етеді. Морфологиялық талдау мен синтез мәселелеріне тоқтатылып, сәйкес орыс және қазақ тілдерінің грамматикалық қасиеттеріне сүйеніп көпмағыналы бейнелеу кестелер және ережелер жиынтығы құрылған. Екі тілді сәйкестендіріп талдау жасадық. Алынған нәтижелер машиналық аударманың сапасын көтереді. Қазақ тілдегі жалғаулардың барлық мүмкін болатын комбинациялар зертелінген. Осы көпмағыналы бейнелеу кестелер әдісі арқылы грамматикалық заңдылықтарды пайдала отырып, орыс тілінен қазақ тіліне машиналық аудармадағы морфологиялық талдауы мен синтезі жоғары сападағы нәтижелерді көрсетті. Аудармашыдағы деректер қорындағы элементерді іздеу жылдамдығын және таңдау сапасын жоғарлатады. Практикалық нәтижеде алынған зерттеулер Microsoft Visual Studio10 бағдарламасының C# тілінде 10 000 дана сөзі бар орыс-қазақ аудармашыны жасау барысында қолданылған.

Әдебиеттер

1. Баскаков Н. А. Хасенова А.К. Исенгалиева И.А. Кордабаев Т.Р Сопоставительная грамматика русского и казахского языков. Морфология. Изд-во "Наука",1966.

2. Введение в теорию многозначных отображений. Составитель Гельман Б.Д., Воронеж, 2003.

3. Tomasz Kaczynski, Multivalued Maps As a Tool in Modeling and Rigorous Numerics. Departement de mathematiques,Universite de Sherbrooke, 2008.

4. Teruko Mitamura, Eric H. Nyberg, Hierarchical lexical structure and interpretive mapping in machine translation, Proceedings of International Conference COLING- 1992 Nantes, 1254-1258 pp.

5. Dilek Zeynep Hakkani, G¨okhan T¨ur, Kemal Oflazer, Teruko Mitamura, and Eric H. Nyberg, An English-to-Turkish Interlingual MT System, Proceedings of International Conference AMTA-1998, pp. 83-94.

Г.Т. БЕКМАНОВА, А. МАХИМОВ

Евразийский национальный университет им. Л.Н. Гумилева
Институт искусственного интеллекта

ГРАФЕМАТИЧЕСКИЙ И МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР КАЗАХСКОГО ЯЗЫКА

Автоматическая обработка текстов естественного языка является одним из актуальных направлений развития искусственного интеллекта и информатики в целом, так как результаты в этом направлении позволит решить проблему создания средств эффективного речевого взаимодействия человека с компьютером. Исследованием этой проблемы уже более 50 лет занимаются специалисты нескольких научных областей. С развитием современных естественно-языковых технологий появилась принципиальная возможность понимания естественно-языкового текста, то есть смысла текста компьютером. Сегодня можно с уверенностью говорить о том, что в Казахстане развивается компьютерная лингвистика, что позволяет надеяться на то, что в скором времени будут существовать лингвистические

процессоры – компоненты, составляющие структуру систем анализа текстов, которые последовательно обрабатывают входной текст. Вход одного процессора является выходом другого[1].

Выделяются следующие компоненты:

• графематический анализ — выделение слов, цифровых комплексов, формул и т.д.;

• морфологический анализ — построение морфологической интерпретации слов входного текста;

• синтаксический анализ — построение дерева зависимостей всего предложения;

• семантический анализ — построение семантического графа текста.

По результатам работы графематического анализатора при анализе повести «Көксерек» М. Ауезова были получены следующие результаты:

Количество абзацев: 231
Количество предложений: 871
Количество слов: 7396
Предложений длины 1: 4
Предложений длины 2: 32
Предложений длины 3: 58
Предложений длины 4: 75
Предложений длины 5: 92
Предложений длины 6: 80
Предложений длины 7: 85
Предложений длины 8: 76
Предложений длины 9: 72
Предложений длины 10: 50
Предложений длины 11: 55
Предложений длины 12: 45
Предложений длины 13: 27
Предложений длины 14: 26
Предложений длины 15: 22
Предложений длины 16: 11
Предложений длины 17: 14
Предложений длины 18: 11
Предложений длины 19: 10
Предложений длины 20: 11
Предложений длины 21: 6
Предложений длины 22: 2
Предложений длины 23: 1
Предложений длины 24: 1
Предложений длины 25: 1
Предложений длины 26: 1
Предложений длины 27: 2
Предложений длины 29:1

Данная статистика необходима для построения семантической модели текста. Поскольку можно предположить, что предложения длиной (под длиной предложения понимается количество слов в нем) меньше 4 слов является простым. А предложение длиной больше или равное 4 может не являться простым. Это простейшая проверка позволит не анализировать синтаксическим анализатором короткие предложения, что экономит время работы алгоритма.

Морфологический анализ казахских текстов – это задача обратная генерации (синтезу) словоформ и новых слов. Под словоформами понимаются измененные с помощью окончаний по падежам, числам, лицам и т.д. слова (флексии), под новыми словами понимаются слова, несущие новую смысловую нагрузку, образованные путем прибавления в суффиксов и т.д. Таким образом, при разработке морфологического анализатора был разработан морфологический синтезатор казахских слов, основанный на формальных правилах.

Морфологический синтез слов осуществляется с помощью эмулятора нейронной сети, который генерирует все словоформы на основе формальных правил (рисунок 1).

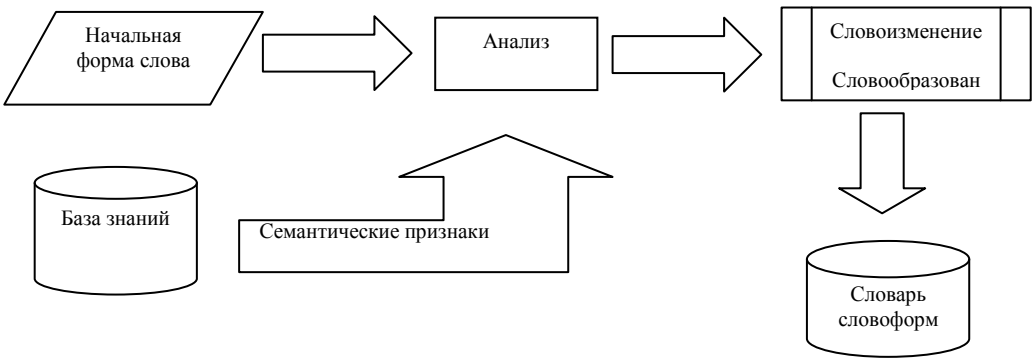


Рисунок 1. Процесс словоизменения

Процесс словоизменения и словообразования основывается на детальном анализе начальной формы слова с целью выделения его морфологических признаков и считывания его семантических признаков из базы знаний. Далее определяется траектория словоизменения, происходит сам процесс словоизменения на основе семантической нейронной сети и запись словоформы и его морфологической информации в словарь словоформ. В таблице 1 приведен пример словоизменения существительного «ізбасар». Фрагмент формальных правил словоизменения на примере существительного с учетом закона сингармонизма, который обуславливает добавления мягких или твердых окончаний в зависимости от мягкости или твердости основы. Данные формальные правила содержат и семантические категории.

Приведенный пример показывает фрагмент правил, где «зе» – зат есім (имя существительное), «жа» - жанды (одушевленность), «01» заканчивается на твердые гласные а, о, ұ, «))» между закрывающими скобками помещены окончания существительных, после «!» морфологическая информация[2].

Таблица 1. Словоизменение одушевленного существительного «ізбасар»

Словоформа	МИ	Словоформа	МИ
ізбасар	зежа	ізбасарымыздың	зежа#тә11іл
ізбасармын	зежа#жі11	ізбасарымызға	зежа#тә11ба
ізбасармыз	зежа#жі11	ізбасарымызды	зежа#тә11та
ізбасарсың	зежа#жі22	ізбасарымызда	зежа#тә11жс
ізбасарсыңдар	зежа#жі22	ізбасарымыздан	зежа#тә11шы
ізбасар	зежа#жі33	ізбасарымызбен	зежа#тә11кө
ізбасарым	зежа#тә11	ізбасарымызбенен	зежа#тә11кө
ізбасарымыз	зежа#тә11	ізбасарыңның	зежа#тә22іл
ізбасарың	зежа#тә22	ізбасарыңа	зежа#тә22ба
ізбасарыңыз	зежа#тә22	ізбасарыңды	зежа#тә22та
ізбасары	зежа#тә33	ізбасарыңда	зежа#тә22жс
ізбасарлар	зежа#кт	ізбасарыңнан	зежа#тә22шы
ізбасарлармыз	зежа#ктжі11	ізбасарыңмен	зежа#тә22кө
ізбасарларсыңдар	зежа#ктжі22	ізбасарыңменен	зежа#тә22кө
ізбасарлар	зежа#ктжі33	ізбасарыңыздың	зежа#тә22іл
ізбасарларым	зежа#кттә11	ізбасарыңызға	зежа#тә22ба
ізбасарларымыз	зежа#кттә11	ізбасарыңызды	зежа#тә22та
ізбасарларың	зежа#кттә22	ізбасарыңызда	зежа#тә22жс
ізбасарларыңыз	зежа#кттә22	ізбасарыңыздан	зежа#тә22шы
ізбасарлары	зежа#кттә33	ізбасарыңызбен	зежа#тә22кө
ізбасар	зежа#ат0	ізбасарыңызбенен	зежа#тә22кө
ізбасардың	зежа#іл	ізбасарлардың	зежа#ктіл
ізбасарға	зежа#ба	ізбасарларға	зежа#ктба
ізбасарды	зежа#та	ізбасарларды	зежа#ктта
ізбасарда	зежа#жс	ізбасарларда	зежа#ктжс
ізбасардан	зежа#шы	ізбасарлардан	зежа#ктшы
ізбасармен	зежа#кө	ізбасарлармен	зежа#кткө
ізбасарменен	зежа#кө	ізбасарларменен	зежа#кткө
ізбасарымның	зежа#тә11іл	ізбасарларымның	зежа#кттә11іл
ізбасарыма	зежа#тә11ба	ізбасарларыма	зежа#кттә11ба
ізбасарымды	зежа#тә11та	ізбасарларымды	зежа#кттә11та
ізбасарымда	зежа#тә11жс	ізбасарларымда	зежа#кттә11жс
ізбасарымнан	зежа#тә11шы	ізбасарларымнан	зежа#кттә11шы
ізбасарыммен	зежа#тә11кө	ізбасарларыммен	зежа#кттә11кө

ізбасарымменен	зежа#тә11кө	ізбасарларымменен	зежа#кттә11кө
ізбасарларыңның	зежа#кттә22іл	ізбасарларымыздың	зежа#кттә11іл
ізбасарларыңа	зежа#кттә22ба	ізбасарларымызға	зежа#кттә11ба
ізбасарларыңды	зежа#кттә22та	ізбасарларымызды	зежа#кттә11та
ізбасарларыңда	зежа#кттә22жс	ізбасарларымызда	зежа#кттә11жс
ізбасарларыңнан	зежа#кттә22шы	ізбасарларымыздан	зежа#кттә11шы
ізбасарларыңмен	зежа#кттә22кө	ізбасарларымызбен	зежа#кттә11кө
ізбасарларыңменен	зежа#кттә22кө	ізбасарларымызбенен	зежа#кттә11кө
ізбасарларыңыздың	зежа#кттә22іл	ізбасарларыңыздан	зежа#кттә22шы
ізбасарларыңызға	зежа#кттә22ба	ізбасарларыңызбен	зежа#кттә22кө
ізбасарларыңызды	зежа#кттә22та	ізбасарларыңызбенен	зежа#кттә22кө
ізбасарларыңызда	зежа#кттә22жс	ізбасарларымсыңдар	зежа#кттә11жі22
ізбасарларысыңдар	зежа#кттә33жі22	ізбасарларыңбыз	зежа#кттә22жі11

Казахский язык, относящийся к группе тюркских языков, очень хорошо поддается формализации. Далее существует три алгоритма работы морфологического анализатора: декларативный, процедурный, комбинированный.

Декларативный метод[3,4]. При таком методе реализации морфологического анализа в словаре хранятся все возможные словоформы каждого слова и соответствующая им морфологическая информация. В этом случае задача морфологического анализа заключается в поиске словоформы в словаре и переписывании из словаря морфологической информации. Так как количество различных словоформ каждого слова довольно велико, декларативный метод требует значительных затрат памяти вычислительной системы, что сопровождается рядом трудностей, связанных с созданием и поддержкой словаря, а также с избыточностью информации. К достоинствам данного метода следует отнести высокую скорость анализа и универсальность по отношению к множеству всех возможных словоформ.

Для эффективного поиска, по словарю словоформ, который в данное время содержит более 2 800 000 словарных статей, было построено в некоторых вариантах анализатора – дерево, в других использовалась хеш-функция.

Процедурный метод предполагает предварительную систематизацию морфологических знаний о естественном языке и разработку алгоритмов присвоения морфологической информации отдельной словоформе[3,4]. Процедурный морфологический анализатор состоит из следующих этапов: выделение в текущей словоформе основы, ее идентификация, приписывание словоформе соответствующего перечня морфологической информации. К недостаткам этого метода относятся высокая трудоемкость составления словарей совместимости, что является трудно решаемой и не автоматизируемой полностью задачей для языков, которым свойственно большое число слов-исключений. Реализация данного способа занимает значительно меньший объем памяти, но при этом увеличивается время морфологического анализа за счет разбиения словоформы на составляющие и применения процедур совместимости [4].

При использовании процедурного метода алгоритм морфологического анализа значительно усложняется. Дело в том, что, например, для существительных личное окончание первого лица «-м» входит в другие окончания «-мін», «-міз» и другие.

Рассмотрим пример для слова «бала - ребенок» (лемма) и двух его словоформ «баламның - моего ребенка», «баламын - я ребенок». В первом случае к основе присоединены два окончания личное окончание первого лица «-м» и падежное окончание «-ның». Во втором случае к основе добавлено одно личное окончание первого лица «-мын». Алгоритм поиска должен предусматривать любое возможное количество присоединенных окончаний и накопление морфологической информации. Ниже будет построен алгоритм поиска слова и его морфологической информации[81,82,83]:

1. Слово считывается;

2. Открывается словарь начальных форм и в нем выполняется поиск считанного слова;
3. Если слово найдено, то перейти к шагу 8, иначе шаг 4 ;
4. Слово в цикле посимвольно считывается, начиная с последнего символа, то, что получается, ищем в словаре окончаний;
5. Если окончание найдено, то остаток ищем в словаре начальных форм;
6. Запоминаем морфологическую информацию слова;
7. Если такое слово не найдено, то переходим к шагу 4, иначе к шагу 8 ;
8. Конец.

Комбинированный метод. В системах реальной степени сложности чаще используется комбинированный вариант морфологического анализа. При этом используется как словарь словоформ, так и словарь основ. На первом этапе проводится поиск по словарю словоформ, как при декларативном методе, и в случае успешного поиска анализ на этом завершается. В противном случае задействуется словарь основ и процедурный метод анализа.

На рисунке приведен пример работающего приложения морфологического анализатора на основе комбинированного метода.

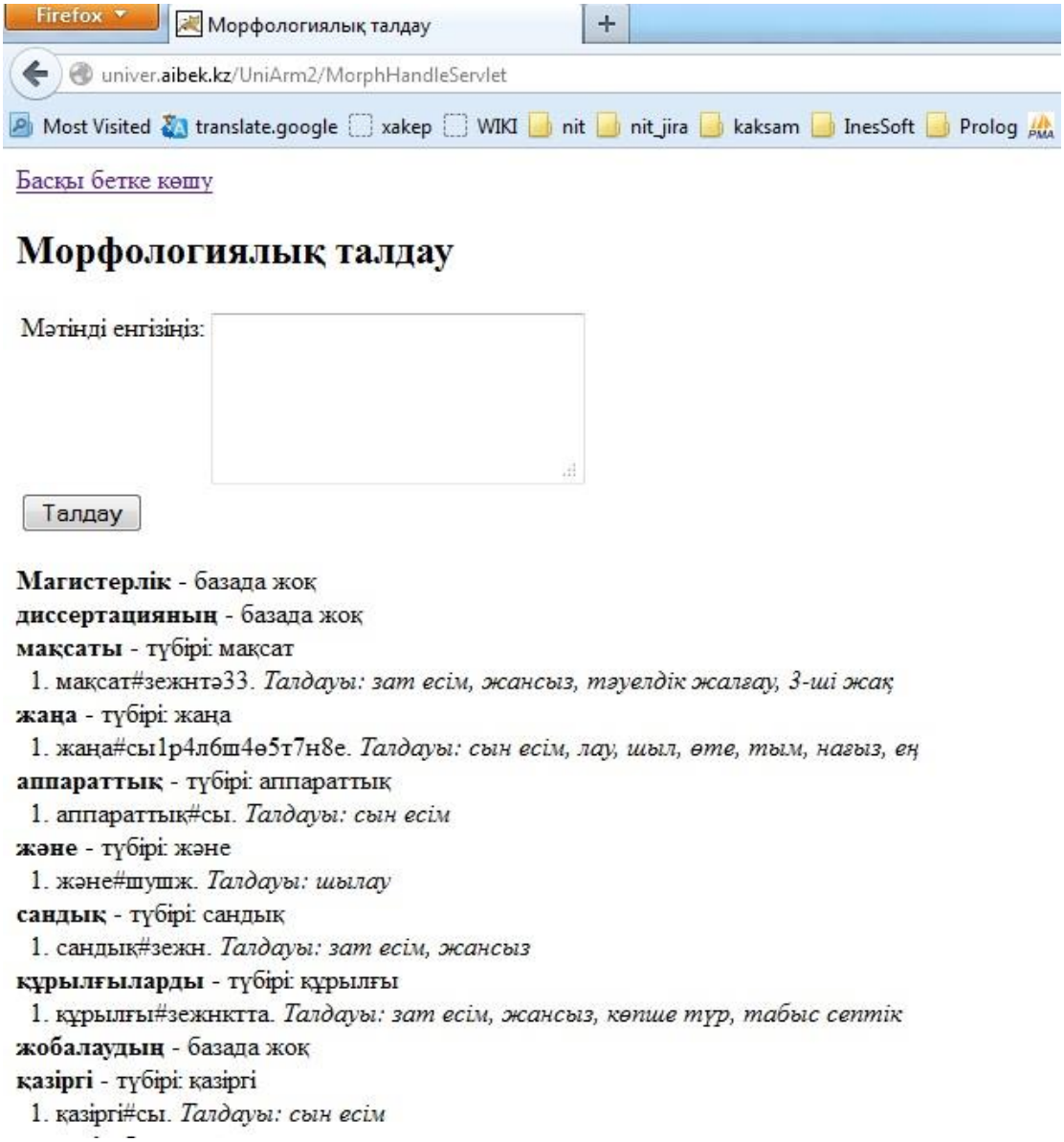


Рисунок 2. Окно морфологического анализатора

За несколько лет работы совместно со студентами и магистрантами над различными модификациями морфологического анализатора казахского языка, можно сделать вывод, что в основе наиболее эффективного анализатора должен лежать именно комбинированный метод его построения. Поскольку в базе данных начальных форм слов находится около 50 тысяч начальных форм слов, из которых сгенерировано более 3 миллионов словарных статей, то в случае работы декларативного метода такое количество является недостаточным для эффективной работы морфологического анализатора. Смею предположить, что для эффективной работы такого метода необходимо 150 тысяч начальных форм слов и около 9 миллионов словарных статей. В сложившейся ситуации комбинированный метод позволяет построить эффективный анализатор около 90%, остальные 10% ошибок можно устранить за счет пополнения базы данных начальных форм слов. Данные ошибки неизбежны из-за вхождений последовательностей символов окончаний и суффиксов друг в друга, когда их невозможно развести явно по семантическим признакам.

Литература

1. А.А. Дунаев. Исследовательская система для анализа текстов на естественном языке
2. Зализняк А.А. Грамматический словарь русского языка. Словоизменение − М.: Русский язык, 1980. − 880 с.
3. Добрушина Е.Р., Савина Г.Б., Гельбух А.Ф. Система точного морфологического анализа и синтеза // Программное обеспечение новой информационной технологии. − Калинин: НПО ЦПС, 1989.
4. Вороной С.М., Егошина А.А. Повышение эффективности интеллектуального поиска в полнотекстовых базах данных на основе автоматического аннотирования документов. VII Международная конференция «Интеллектуальный анализ информации ИАИ-2007».: Киев, 15-18 мая 2007 г.: Сб. тр.//Рос.ассоц.искусств.интеллекта и др.; Под ред. Т.А. Таран. – К.:Просвіта, 2007. –392 с.

А.С. МУКАНОВА, Б.Ж. ЕРГЕШ, РАЗАХОВА Б.Ш.

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

МОРФОЛОГИЯЛЫҚ ЕРЕЖЕЛЕРДІҢ ОНТОЛОГИЯЛЫҚ ЕРЕЖЕЛЕРІ

Бұл жұмыста қазақ тілінің морфологиялық ережелерінің онтологиялық модельдері көрметілген.

Кілттік сөздер: онтология, морфологиялық ережелер, семантикалық гиперграф.

Қазіргі кезде табиғи тілді өңдеу процестерінде қиындық тудыратын жағдай ол формалдау болып табылады. Сөзді тудыруды қандай да бір белгіленген траектория бойынша жүзеге асыру қиын, себебі, сөзден жаңа сөз тудыру оның бастапқы түбірінің мағынасына байлынысты болады. Сондықтан да білімді ұсынудың формалды құралын таңдау қажет. Білімді ұсынудың формалды құралының бірі ретінде семантикалық гиперграфты алуға болады. Семантикалық гиперграф семантикалық желінің кеңейтілген түрі, онда *n*-арлы қарым- қатынас көрсетілген, сол себепті де семантикалық гиперграф арқылы объектінің атрибуттарын ғана емес, сонымен қатар, олардың құрылымын, «бүтіндік» бейнеленуін көрсетуге болады[1,2].

H гиперграфы (***V, R***) жұбымен анықталады, мұндағы ***V***= $\{v_i\}$ – төбелер жиыны; ***R***= $\{r_j\}$ – қабырғалар жиыны;*i* = 1, 2, …, *n*;*j* = 1, 2, ..., *m*; әрбір қабырға келесі жиыннан тұрады ***V***, яғни, $r = \{(v_{j_s}, v_{j_t})\}$, $j_s \neq j_t$, s, t – натуралды сандар.

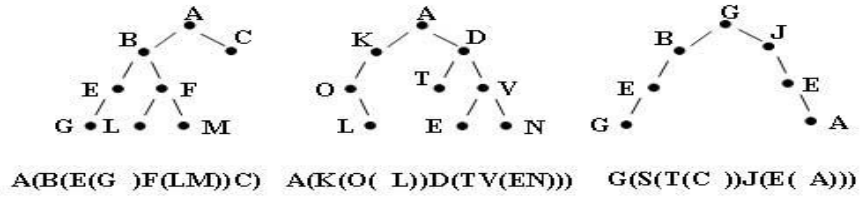
Онтология түсінігі грек тілінен аударғанда онтос – жан,мән,мағына, логос - білім дегенді білдіреді, яғни болмысты зерттейтін философиялық ілім.

Техникалық ғылымдарда онтологияның қолданылуының негізгі мәні белгілі бір білім облысы бойынша мәліметтер жиынының барлығын қамтитын және бөліктік формализацияны концептуальді сызбамен көрсетуі. Концептуальды сызба негізінде түсініктер жиыны + түсініктер жайлы мәліметтер (қасиет, қатынас, шектеу, аксиомалар және түсініктердің бекітілуі, бұл ақпараттардың барлығы таңдалынған пәндік облыс бойынша есептің шешілу процесін сипаттау үшін қажет) беріледі.

Компьютерлік лингвистикамен айналысатын мамандардың арасында онтологияның нақты тұрақталынған (классикалық) анықтамасын Губерт ұсынды: онтология – бұл концептуализацияның эксплицитті спецификациясы, мұнда концептуализация ретінде пәндік облыстың объекттерінің жиындары мен олардың арасындағы байланыстардың сипатталуы түсініледі.

Қазақ тілінің морфологиялық ережелерінің онтологиялық модельдері онтологияны құрудың негізгі ережелеріне сай құрылады, бірақ та білімді ұсыну тілі ретінде семантикалық гиперграф алынады. Осы формализм онтологияны семантикалық гиперграф түрінде бейнелеуге мүмкіндік береді: O (X, R, I), мұндағы X – пәндік облыс түсініктерінің жиыны (гиперграф төбесі), R – түсініктер арасындағы қарым- қатынастар жиыны (гиперграф доғалары мен қабырғалары), ал I – осы пәндік облыстағы түсініктер мен қарым- қатынастар аттарының жиыны.

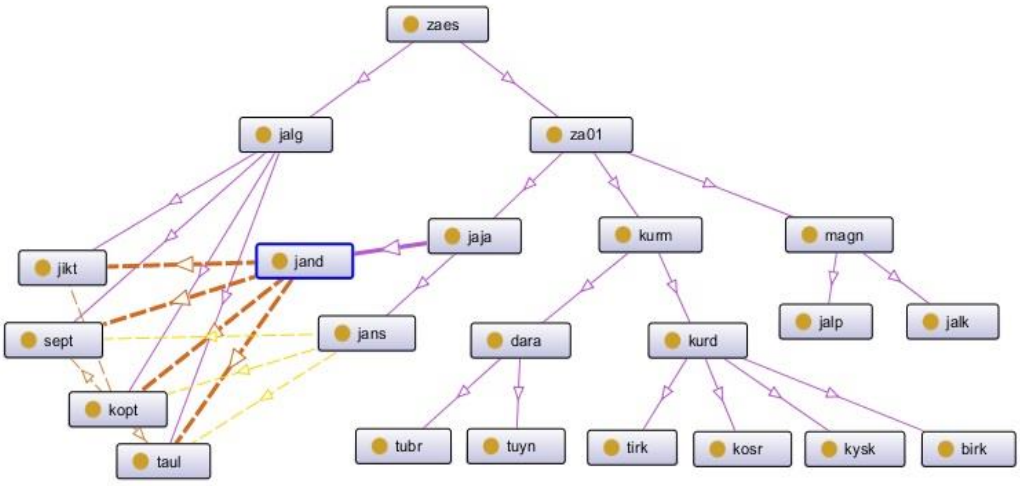
Семантикалық гиперграф арқылы түбір сөздерге жалғаулар мен жұрнақтарды қосу ережелері формалданады. Гиперграфтар негізінен алғанда көп өлшемді құрылым болып табылады. Бірақ көптеген жағдайда бір өлшемді құрылымдарды пайдаланған тиімді болып жатады. Сондықтан да көп өлшемді графтар құрылымындағы ақпараттың бір өлшемді құрылымға өткені біз үшін қажет. Графтарды бірөлшемді құрылым ретінде бейнелеу үшін сызықтық жақшалық жазбаларды қолдануға болады, яғни, гиперграфтарды жолдар түрінде бейнелеу [3]. Онда гиперграфтың төбелері ашылған және жабылған дөңгелек жақша түрінде бейнеленеді. Гиперграфтар мен олардың сызықтық жақшалық жазбалары арасында тығыз біртипті өзара байланыс бар. Сызықтық жақшалық жазба қандай да бір гиперграфты толықтай тексеріп, айналып өткеннен кейін жасалынады. Графтар мысалы және оларға сәйкес келетін сызықтық жақшалық жазбалардың мысалдары сурет 1-де көрсетілген.



Сурет 1. Сызықтық жақшалық жазба түрлері

Қазақ тілінде 9 сөз табы бар. Олар: зат есім, сын есім, сан есім, етістік, есімдік, үстеу, одағай, шылау, еліктеу сөздер. Төменде осы сөз таптарының онтологиялық модельдері, сызықтық жақшалық жазба арқылы алынған формалды ережелер көрсетіледі, ол ережелер қазақ тілінің сөздерін тудыру үшін қолданылады.

Сурет 2-де зат есімнің онтологиялық моделі көрсетілген.

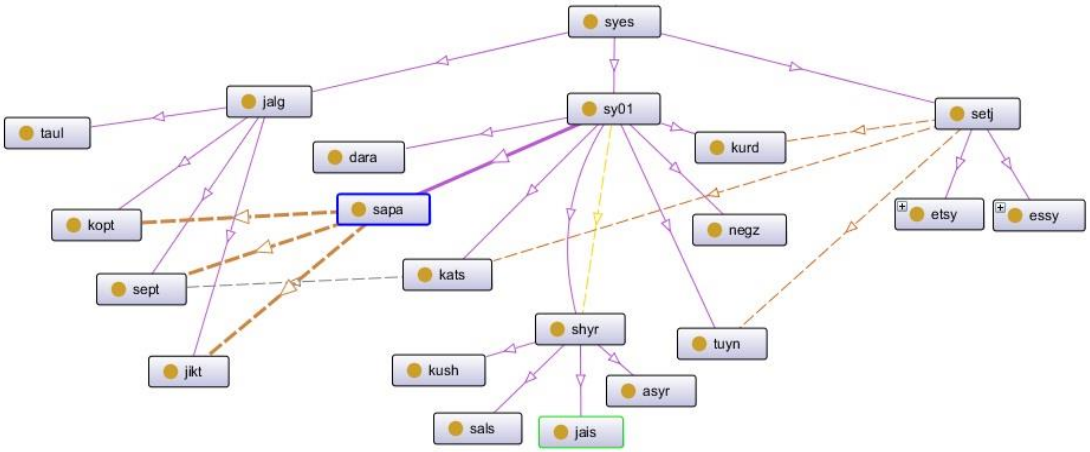


Сурет 2. Зат есімнің онтологиялық моделі

Формалды ережелер төмендегідей болады:

((зежа01))!ат0	(зат есім,жанды, 01 дауысты)! атау септік	((бала))
((зежа01)ның)!іл	((зат есім,жанды, 01 дауысты)ның)! ілік септік	((бала)ның)
((зежа01)ға)!ба	((зат есім,жанды, 01 дауысты)ға)! барыс септік	((бала)ға)
((зежа01)ны)!та	((зат есім,жанды, 01 дауысты)ны)! табыс септік	((бала)ны

Сурет 3- те сын есімнің онтологиялық моделі көрсетілген.

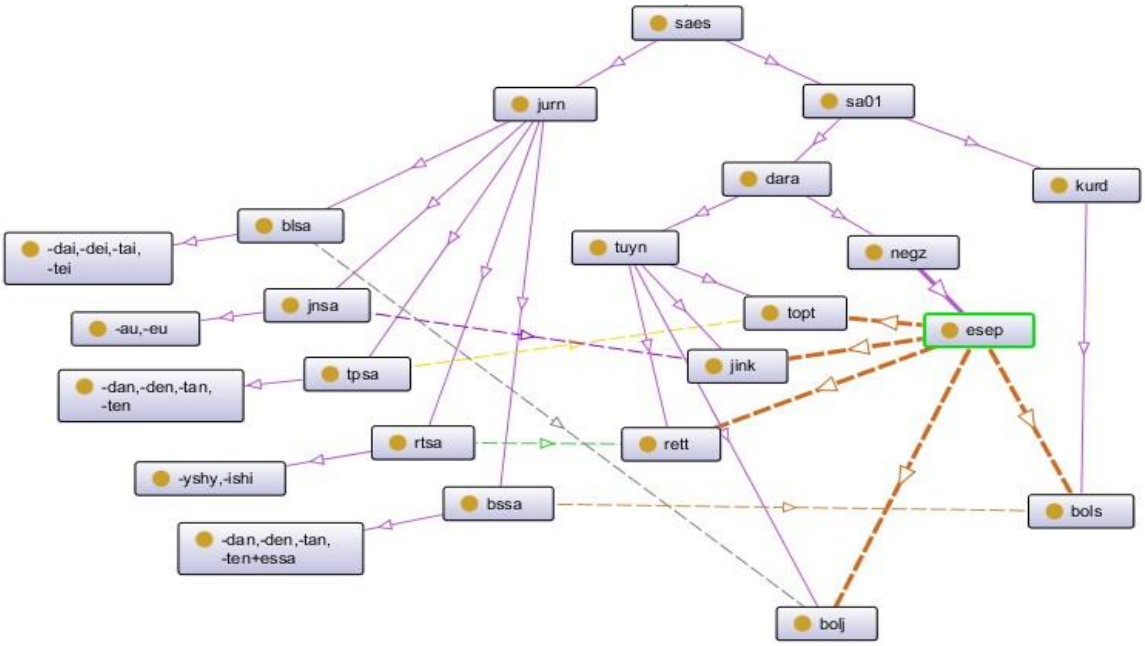


Сурет 3. Сын есімнің онтологиялық моделі

Сын есім үшін формалды ережелер:

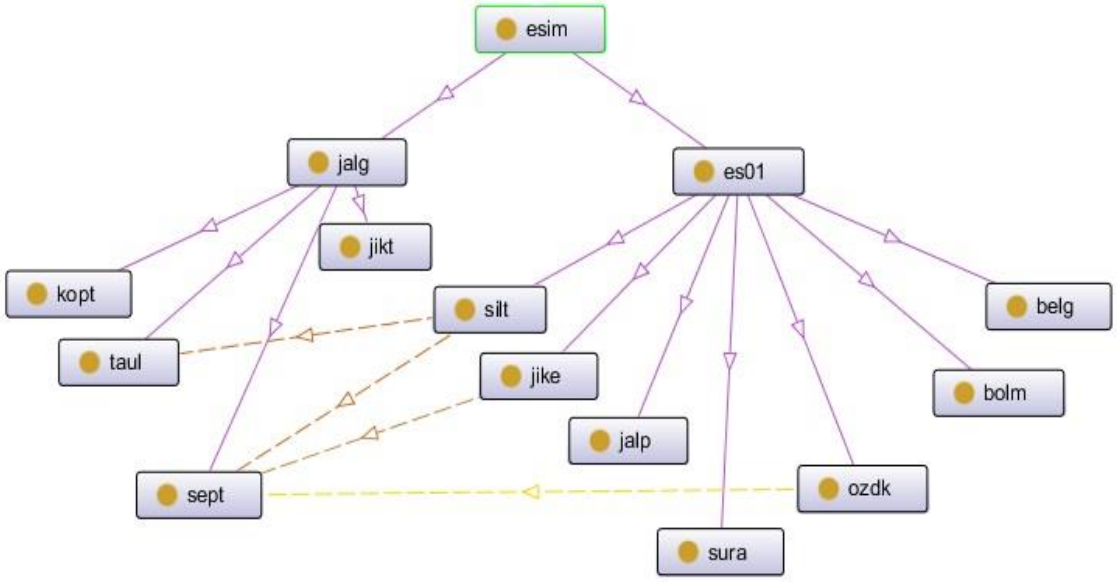
((сы011р)рақ)!сыса	сын есім, дауысты, рақ жұрнақ	(ақылды)рақ – «умнее»
((сы014л)лау)!сыса	салыстырмалы шырай	(ақылды)лау – «умнее» өте ақылды – «очень умный»
((сы01)ның)!іл	сын есім, ілік септік	(ақылды)ның – «умного»
((сы01)мын)!жк11	сын есім, I-жақ жіктік жалғау	(ақылды)мын – «я умный»
((сы01)сың)!жі22	сын есім, II-жақ жіктік жалғау	(ақылды)сың – «ты умный»

Сурет 4- те сан есімнің онтологиялық моделі көрсетілген.



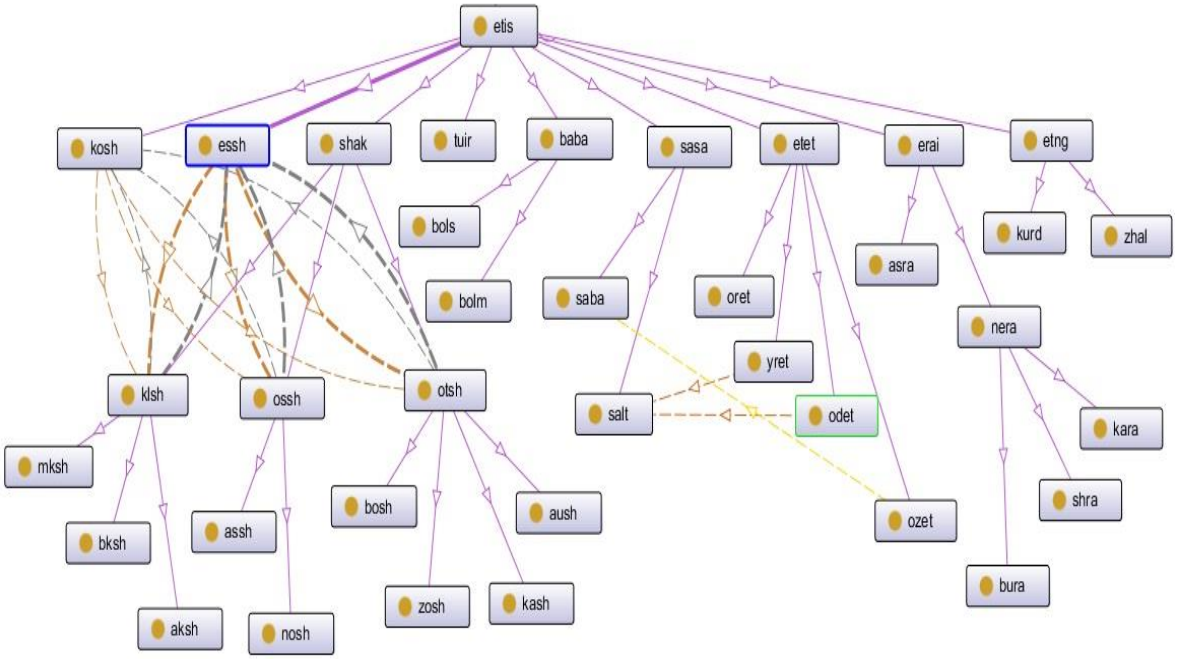
Сурет 4. Сан есімнің онтологиялық моделі

Сурет 5- те есімдіктің онтологиялық моделі көрсетілген.

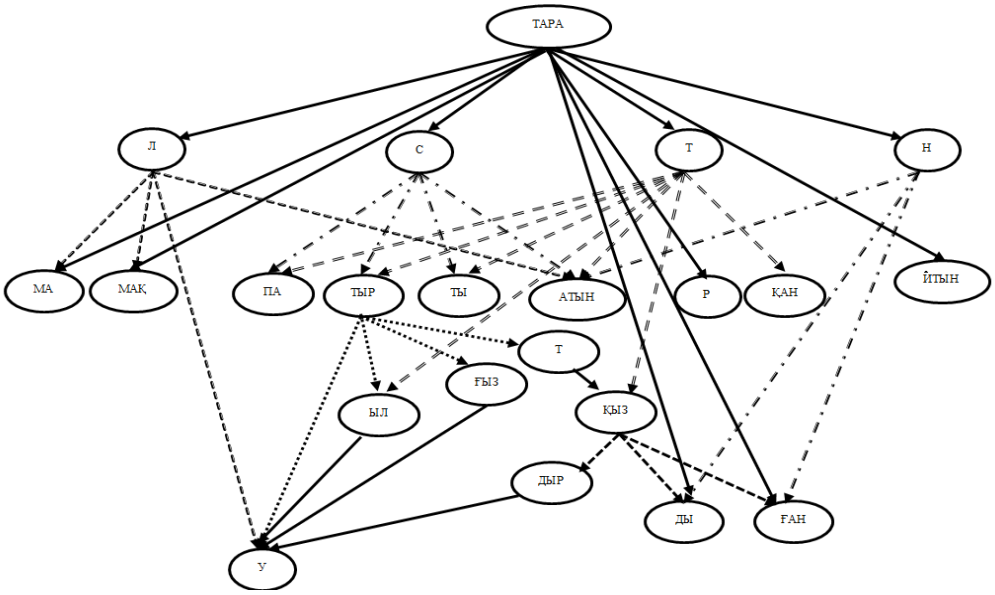


Сурет 5. Есімдіктің онтологиялық моделі

Сурет 6- да етістіктің онтологиялық моделі, сонымен қатар сурет 7- де қазақ тіліндегі «тара» етістігінің оның жаңа сөзжасамдарын жасауға мүмкіндік беретін семантикалық гиперграф түрінде бейнеленуі көрсетілген.



Сурет 6. Етістіктің онтологиялық моделі



Сурет 7. Семантикалық гиперграф түрінде бейнеленуі

Жоғарыда көрсетілген етістіктен етістік формаларын [4] тудыруды бейнелейтін семантикалық гиперграфты сызықтық жақшалық жазба арқылы бейнелейтін болсақ төмендегідей формалды ережелер шығады:

((ет01нг)с)!ое
(((ет01нг)с)тыр)!өг
((((ет01нг)с)тыр)ыл)!ые
((((ет01нг)с)тыр)у)!қа
((((ет01нг)с)тыр)т)!өг
(((((ет01нг)с)тыр)т)қыз)!өг
(((ет01нг)т)қалы)!өк
(((ет01нг)т)қанша)!өк
(((ет01нг)т)а)!ок
(((ет01нг)т)у)!қа
((ет01нг)л)!ые
(((ет01нг)л)ды)!өе

((((ет01нг)с)тыр)ғыз)!өг
(((ет01нг)с)ын)!өз
(((ет01нг)с)па)!бз
(((ет01нг)с)ты)!өе
(((ет01нг)с)қан)!өе
(((ет01нг)с)атын)!өе
(((ет01нг)с)ар)!ке
(((ет01нг)с)пақ)!ке
(((ет01нг)с)ып)!өк
(((ет01нг)с)қалы)!өк
(((ет01нг)с)қанша)!өк
(((ет01нг)с)а)!ок
((ет01нг)т)!өг
(((ет01нг)т)ыл)!ые
((((ет01нг)т)ыл)ма)!бз
(((ет01нг)т)па)!бз
((((ет01нг)т)па)у)!қа
(((ет01нг)т)тыр)!өг
(((ет01нг)т)қыз)!өг
((((ет01нг)т)тыр)т)!өг
((((ет01нг)т)тыр)ғыз)!өг
((((ет01нг)т)қыз)дыр)!өг
((((ет01нг)т)қыз)ғыз)!өг
(((((ет01нг)т)тыр)т)қыз)!өг
((((((ет01нг)т)тыр)т)қыз)дыр)!өг
(((((ет01нг)т)қыз)дыр)т)!өг
(((ет01нг)т)ты)!өе
(((ет01нг)т)қан)!өе
(((ет01нг)т)атын)!өе
(((ет01нг)т)ар)!ке
(((ет01нг)т)пақ)!ке
(((ет01нг)т)ып)!өк
(((ет01нг)л)ған)!өе
(((ет01нг)л)атын)!өе
(((ет01нг)л)ар)!ке
(((ет01нг)л)мақ)!ке
(((ет01нг)л)ып)!өк
(((ет01нг)л)ғалы)!өк
(((ет01нг)л)ғанша)!өк
(((ет01нг)л)а)!ок
(((ет01нг)л)ма)!бз
((((ет01нг)л)ма)с)!ое
((ет01нг)н)!өз
(((ет01нг)н)ды)!өе
(((ет01нг)н)ған)!өе
(((ет01нг)н)атын)!өе
(((ет01нг)н)ар)!ке
(((ет01нг)н)бақ)!ке
(((ет01нг)н)ып)!өк
(((ет01нг)н)ғалы)!өк
(((ет01нг)н)ғанша)!өк
(((ет01нг)н)ба)!бз
(((ет01нг)л)а)!ок
((ет01нг)ма)!бз
((ет01нг)у)!қа
((ет01нг)ды)!өе
((ет01нг)ған)!өе
((ет01нг)йтын)!өе
((ет01нг)р)!ке
((ет01нг)мақ)!ке
((ет01нг)п)!өк
((ет01нг)ғалы)!өк
((ет01нг)ғанша)!өк
((ет01нг)й)!ок

Осы формалды ережелердің көмегімен қазақ тіліндегі етістіктерден етістіктің 75-ке жуық жаңа сөзжасамдарын тудыруды автоматтандыруға болады.

Қорытынды

Қазақ тілінің морфологиялық ережелерінің онтологиялық ережелері құрылды, білімді ұсыну тілі ретінде семантикалық гиперграф қолданылды, соның нәтижесінде сөздерді түрлендіру мен тудыруды автоматты жүзеге асыруға мүмкіндік беретін формалды ережелер жасалды. 40000 бастапқы түбір сөзден тұратын базаны әрбір сөз табы үшін алынған формалды ережелер арқылы генерациялау барысында 3 200 000 жаңа сөзжасамдар алуға мүмкіндік берді.

Әдебиеттер

1. Berge C.C. Graphs and Hypergraphs, Elsevier Science Ltd. 1985
2. Vizing V.G.(). About a coloring of intsidentor in the hypergraph. Diskretn. Anal. Issled. Oper., Ser. 1, 14:3, 2007. p. 40–45.
3. Батищев, Д.И. Многоуровневая декомпозиция гиперграфовых структур. /Д.И. Батищев, Н.В. Старостин, А.В. Филимонов. //Прилож. К журналу «Информационные технологии» №5(141) 2008, С.1 - 32.
4. Ысқақов А. Қазіргі қазақ тілі. – 2-басылымы. Филология факультеттері студенттеріне арналған оқулық. – Алматы: Ана тілі, 1991, Б. 135-148.

Б.Ж.ЕРГЕШ, А.С.МУКАНОВА, Б.Ш.РАЗАХОВА

Л.Н.Гумилев атындығы Еуразия ұлттық университеті, Астана

ҚАЗАҚ ТІЛІНДЕГІ ЖАЙ СӨЙЛЕМДЕРДІҢ ОНТОЛОГИЯЛЫҚ МОДЕЛІ

Қазақ тілінде мынадай сөйлем мүшелері бар: *бастауыш, баяндауыш, толықтауыш, пысықтауыш, анықтауыш.*

Сөйлем құрамында сөйлем мүшелерінің белгілі бір орны бар. Қазақ тіліндегі сөйлемнің құрылымындағы басты ерекшелік – бастауыш сөйлемнің басында, баяндауыш көбінесе соңында қолданылады. Анықтауыш бастауыш пен толықтауыштың алдынан, ал толықтауыш көбінесе баяндауыштың алдынан; пысықтауыш - өзіне қатысты сөздің алдынан қолданылады. Бұл - сөйлемнің қазақ тіліне тән құрылымдық үлгісі. Дегенмен, сөйлем мүшелерінің орын тәртібі өзгеруі де мүмкін. Біз олардың өзгеруіне сәйкес екі, үш, төрт және бес мүшенің қатысуымен жасалатын топтарға бөліп қарастырамыз. Сөйлем мүшелері: бастауыш, баяндауыш, толықтауыш, анықтауыш және пысықтауышты сәйкесінше Бс, Бн, Т, А, П таңбаларымен таңбалайық.

1. <Бс> + <Бн>;
2. <Бс> + <Т> + <Бн>;
3. <Бс> + <П> + <Бн>;
4. <Бс> + <Т> + <П> + <Бн>;
5. <Бс> + <П> + <Т> + <Бн>;
6. <Бс> + <А> + <Т> + <Бн>;
7. <Т> + <А> + <Бс> + <Бн>;
8. <Т> + <П> + <Бс> + <Бн>;
9. <Т> + <Бс> + <П> + <Бн>;
10. <П> + <Бс> + <Т> + <Бн>;
11. <А> + <Бс> + <Т> + <Бн>;
12. <А> + <Бс> + <П> + <Бн>;
13. <Бс> + <П> + <А> + <Т> + <Бн>;
14. <Бс> + <А> + <Т> + <П> + <Бн>;
15. <П> + <Бс> + <А> + <Т> + <Бн>;
16. <П> + <Т> + <А> + <Бс> + <Бн>;
17. <А> + <Бс> + <П> + <Т> + <Бн>;
18. <А> + <Бс> + <Т> + <П> + <Бн>;
19. <А> + <Т> + <Бс> + <П> + <Бн>.

Лепті сөйлем мен өлең жолдарындағы сөздердің орын тәртібінде өзгеріс болуы мүмкін. Шындығында қазақ тілінде грамматикалық қатынастар сөз түрлендіруші формалар мен көмекші сөздер арқылы (оның ішінде әсіресе көмекші етістіктер арқылы) беріледі.

Қазақ тіліндегі жай сөйлемдердің синтаксистік ережелерінің формалды грамматика көмегімен математикалық моделдері [1] және семантикалық моделдері құрастырылған [2].

Бұл жұмыста жай сөйлемдердің жоғарыда келтірілген құрылымға сәйкес онтологиялық модель тұрғызылды.

Онтологиялық моделдің негізгі мәні белгілі бір білім облысы бойынша мәліметтер жиынының барлығын қамтитын және бөліктік формализацияны концептуальді сызбамен көрсетуі. Концептуальды сызбада түсініктер жиыны мен түсініктер жайлы мәліметтер (қасиет, қатынас, шектеу, аксиомалар және түсініктердің бекітілуі, бұл ақпараттардың барлығы таңдалынған пәндік облыс бойынша есептің шешілу процесін сипаттау үшін қажет) беріледі.

Онтологияның көптеген модельдері келесі компоненттерден тұрады:

- *концепттер*(түсінік, класстар),
- *концепттердің* қасиеттері (атрибуттары, ролдері),
- *қатынастар* концепттер арасында (тәуелсіздік, функциялар),
- қосымша *шектеулер*,олар аксиомалармен анықталады,
- қолданылу мысалдары.

Ұсынылатын онтологиялық модель сөйлем мүшелерінің семантикалық сипаттамаларымен құрастырылады, ал сөйлем мүшелерінің семантикалық сипаттамасы сөз таптарымен анықталады. Қазақ тілінің грамматикасынан белгілі сөйлем мүшелерінің қандай сөз таптары болатындығын қарастырайық [3].

Бастауыш болатын сөз таптары:

– атау түрдегі, тәуелдік жалғаулы және көптік жалғаулы зат есім, мысалы: Мына ***кітап*** тамаша жазылыпты. Айгүлдің ***үйі*** кеше қалаға көшті. ***Оқушылар*** еңбек ардагерлеріне көмектесті;

– заттанған сын есім (біріншіден, заттың орнына айтылуы керек, екіншіден, сөйлем ішінде басқа сын есім немесе сын есімнен шыққан сөз болуы керек), мысалы: ***Молшылық*** біздің *адал* еңбегімізбен жасалған;

– көптік жалғаулы заттанған сын есім, мысалы: Жақсылар елге еңбегімен танылады;

– сан есім (артынан айтылған зат есім жоқ болса), жинақты сан есім және оның тәуелденген түрі, есепті сан есімнің тәуелді түрі, ретті сан есім, бөлшек сандардың бөлшегі мен көрсеткішінің тәуелденген түрі, шақты, шамалы шылаулары бар сан есім, оның көптік түрі де тәуелдік жалғау жалғанған түрі де, мысалы: Бес - екіге қалдықсыз бөлінбейді. Олардың екеуі де өз мамандықтарын жақсы біледі. Жарысқа қатысушылардың бесіншісі бәрінен жүйрік. Оқушылардың екінің бірі үздік оқиды. Ауыл үйлерінің он шақтысы жайлауға көшіп үлгерді;

– жіктеу, сілтеу, сұрау, өзіндік (тәуелді жалғаулы өз), белгісіздік, болымсыздық (тәуелді жалғаулы ешбір) есімдіктері, мысалы; Олар кездесетін орынға межелі уақыттан ерте жетті. Ондай жалқауларға сол керек. Шәмшінің әндерін кім ұнатпайды дейсің. Өзі әнді тамылжытып ала жөнелді;

– зат есімнің тіркесінсіз етістіктен зат есімге айналған сөздер, мысалы: Білетіндер емтиханды тез тапсырып шығып жатыр;

– де етістігі арқылы объектке айналған етістіктер, мысалы: Кешіктім деген бір күнді жоғалтқанмен бірдей;

– объектке айналған немесе жұрнақ жалғанып басқа сөз табына айналған одағай, шылау және үстеу сөздер, мысалы: Әйт-шу дегендер малдың басын тез қайырды. Еріншектің ***ертеңі*** бітпес. .

Баяндауыш болатын сөз таптары:

– етістік, мысалы: Жұмысшылар сегіз сағат жұмыс ***жасады***.

Толықтауыш болатын сөз таптары:

– ілік септігінен басқа септік жалғауларында тұрған зат есім және затқа айналған сын есім, сан есім, есімше, есімдік, мысалы: Айжан жаңа жылды үлкен жетістіктермен қарсы алды. Мен кеше онымен сөйлесіп қалдым. Тәжірибелі ұстаздардың еңбегі жастарға өнеге болады. Төрт екіге қалдықсыз бөлінеді. Маржан өз ***айтқанынан*** қайтпады.

Пысықтауыш болатын сөз таптары:

– үстеу, мысалы: Бүгін күн жылынды. Айгүлдер асықпай шығып кетті.

– сын есім, мысалы: Айман жылы сөйлеп, оқушылардың тілін тапты. Бұл дұрыс айтылған екен.

– көсемше, мысалы: Ақын өз өлеңін мәнерлеп оқыды.

– барыс, жатыс, шығыс және көмектес септіктерінде тұрған сөздер, мысалы: Балалар мектепке жиналды. Қалада зәулім ғимараттар көп. Автобус ауылдан ұзап кетті. Сені дауысыңнан таныдым.

– шейін, дейін бола, қарай, таман шылаулары тіркескен барыс септіктегі сөздер, мысалы: Кешке таман ауыл қарттарына концерт ұйымдастырылды. Түнге қарай күн суытты.

– кейін, ары, соң, бұрын шылаулары тіркескен кейбір шығыс септіктегі сөздер, мысалы: Программаны жазбастан бұрын математикалық модель құру қажет. Біз бір айдан соң мектеппен қоштасамыз.

– бірге, қабат, қатар, шылаулары тіркескен кейбір көмектес септіктегі сөздер, мысалы: Айгүл өзімен бірге құрбысын ертіп келді. Ауыл тұрмысы қаламен қатар өсіп келеді.

– арқылы, арқасында, үшін, сайын шылаулары тіркескен сөздер, мысалы: Диқан еңбегінің жемісін ***күн сайын*** бақылады.

Анықтауыш болатын сөз таптары:

– сын есімнің атау түрі, мысалы: Арман қызыл түсті жақсы көреді;

– сан есімнің атау, туынды түрі, мысалы: Егістікте он комбайын жүр. Он екінің бөлгіштері: бір, екі, үш, төрт, алты және өзі;

– зат есім (қатар тұрған екі зат есімнің бұрын айтылғаны) , мысалы: Атай қыш құмыра жасағанды ұнатады;

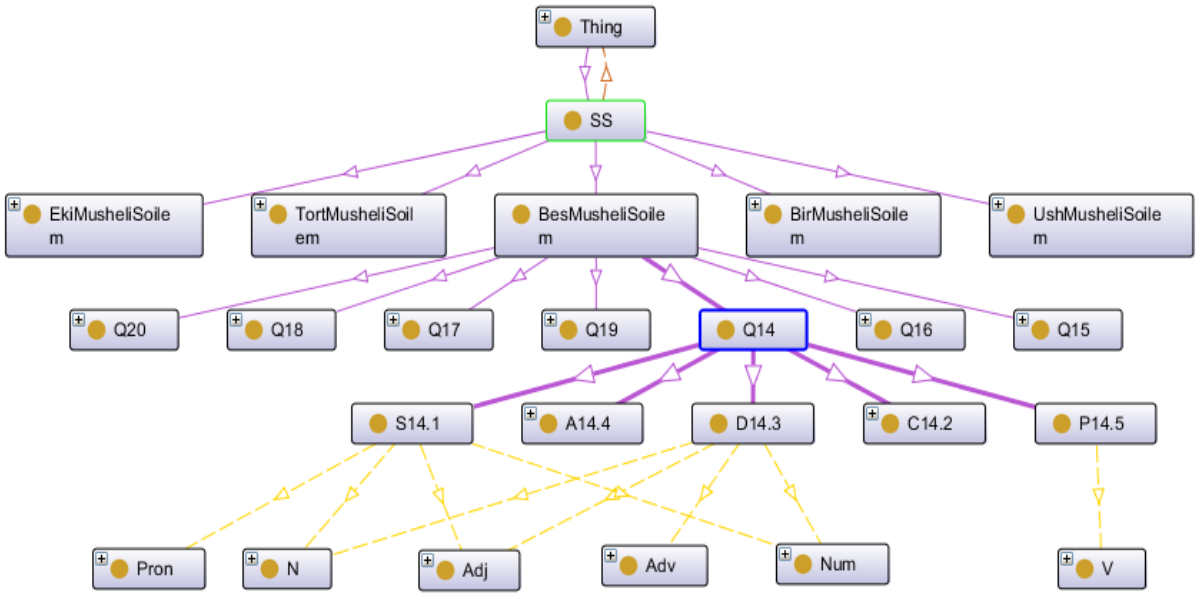
– сілтеу, өзіндік, жалпылау, белгісіздік есімдіктердің атау түрі, мысалы: Мына сурет түрлі түсті бояумен салыныпты. Мынадай табиғатты өз көзіңмен көргенге не жетсін! Барлық халық бейбітшілікті қалайды. Жолаушылардың бірнеше күні бар;

– есімше мен етістік есімдер, мысалы: Алынған сыйлықтар сәбилерге тапсырылды;

– еліктеуіш сөздер, мысалы: Гуу-гу әңгімемен ауылға да жеттік;

– ілік септігіндегі барлық сөз таптары, мысалы: ***Майраның*** апасы мектепке келді. Мұны ***айтқан кісінің*** атын білесің бе? ***Үлкеннің*** айтқанын тыңдау қажет.

Қазақ тілінің жай сөйлемінің онтологиялық моделінің фрагменті 1-суретте көрсетілген, ал 1-кестеде онтолгиялық моделді құруда қолданылған атаулар мен белгілер көрсетілді.



Сурет 1 – Қазақ тілінің жай сөйлемінің онтологиялық моделінің фрагменті

Кесте 1 – Онтологиялық моделді құруда енгізілген атаулар

Қысқаша белгіленуі	**Атауы**
SS(Simple Sentence)	Жай сөйлем
Q	Құрылымы
Q_1	Бірінші индексті құрылым

S (Subject)	Бастауыш
A (Addition)	Толықтауыш
D (Determination)	Анықтауыш
C (Condition)	Пысықтауыш
P (Predicate)	Баяндауыш
N (Noun)	Зат есім
Adj (Adjective)	Сын есім
Num (Numeral)	Сан есім
Adv (Adverb)	Үстеу
Pron (Pronoun)	Есімдік
V (Verb)	Етістік

Құрастырылған онтологиялық моделдерді синтаксистік талдауға қолдануға болады. Синтаксистік талдаудың басты мақсаты – сөйлемнің құрылымын талдау. Құрылымды тілдің контексті бос грамматикасын талдауға сәйкес ағаш ретінде қабылдауға болады. Синтаксистік талдау нәтижесі сөздердің симантикалық базасына сілтеу жасайтын синтаксистік шығарылым бұтағы болып табылады. Синтаксистік талдау барысында сонымен бірге сөйлем құрылымымен байланысты қателер де табылады.

Әдебиеттер

1. Уталина Б. Ш.,А.Ә. Шәріпбаев. Контексті бос грамматика арқылы қазақ тілі сөйлемдер жиынының анықталуы. //Қазақстан Республикасының Ұлттық Ғылым Академиясының Баяндамалары. - Алматы, 2005. -№5. - Б 123-128.

2. Б.Ш. Разахова, Ф.М. Туледиярова. Семантика желі көмегімен қазақ тілінің жай сөйлемдерін формалдау // Вестник. Астана: Евразийский национальный университет им. Л.Н.Гумилева, 2012. – Специальный выпуск.– С.403-409.

3. Балақаев М. Б. Қазіргі қазақ тілі: Сөз тіркесі мен жай сөйлем синтаксисі. – Астана: Л.Н.Гумилев атындағы ЕҰУ, 2006. -237б.

Г.К. ЕЛИБАЕВА, Б.З. АНДАСОВА

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

МӘТІНДІК ҚҰЖАТТАРДЫ КЛАССИФИКАЦИЯЛАУДА ОНТОЛОГИЯНЫ ҚОЛДАНУ

Бүгінгі таңда әр түрлі саладағы электронды құжаттар санының қарқынды өсуі және олардың Интернет желісінде қолжетімді болуы ақпараттың басым бөлігінің компьютерде электронды мәтіндік құжаттар түрінде сақталуына әкеп соғады. Көптеген ұйымдарда қажетті білімдердің едәуір бөлігі құжаттық деректер қорында болады. Осындай жағдай мәтінді талдау (Text Mining) саласына, яғни мәтіндік құжаттардан білімдерді автоматты түрде алу мен өңдеу әдістеріне жоғары қызығушылық танытып отыр. Табиғи тіл мәтіндерінің құрылымымен автоматты түрде білімдерді алу қиын. Мұндай білімдер сараптамашы арқылы жеңіл алынады, бірақ электронды құжаттар санының аса көп болатынын ескерсек, олардың адам арқылы тиімді өңделуі жұмсалатын уақыт пен ресурстар тұрғысынан шығынды көп қажет етеді.

Білімдерді алудың түпкі мақсаты – жобалық шешім қабылдау барысындағы сарапшының немесе автоматтандырылған жүйенің ақпараттық қолдауы болып табылады. Мамандармен құрастырылған құжаттарда әр түрлі мәселелерді шешуге арналған әдістер, параметрлерді

таңдауға арналған ұсыныстар және ұйым қызметінің түрлі саласында пайдалы болатын т.б. білімдер сипатталуы мүмкін. Осылайша, білімдерді алу жүйесінің негізгі қызметі құжаттар қорындағы пайдалы мәліметтерді ақпараттық іздестіру болып табылады. Алайда, осы есеппен қатар автоматты классификациялаудың, кластерлеудің және құжаттарды аннотациялаудың аралық есептері де шешілуі тиіс. Осы уақытқа дейін осындай табиғи тіл мәтіндерінің есептерін шешудің көптеген жаңа әдістері жасалынды, сонымен қатар қолданыстағы бар әдістердің тиімділігі де жоғарылап жатыр.

Көптеген ұйымдардың білімдерді басқару (Knowledge Management) жүйелерінде осы ұйымдардың мамандандырылуына сәйкес пәндік саланы сипаттайтын пәндік онтологияларды сарапшылар құрастырады. Онтологияны қолданып сипатталатын білімдер моделі концепттер (түйінді ұғымдар) жиынтығы мен олардың арасындағы байланысты көрсетеді. Мәтіндік талдау есептерінде онтологияны қолдану мәтіндік құжаттардан білімдерді алу мен өңдеу есептерін шешудің тиімділігін жоғарылатады.

Онтология – интеллектуальды жүйелерде білімдерді ұсынудың бір түрі. Онтология деп қарастырылатын пәндік саланың ұғымдар жүйесін, олардың арасындағы қатынастар мен амалдарды түсінеміз, басқаша айтқанда, онтология – пәндік сала мазмұнының анықталуы (спецификациясы). Мысалы, «Интеллектуальды жүйелер» онтологиясы мынадай түрде болуы мүмкін: *Интеллектуальды жүйелер = {интеллект; нейрон; нейрондық желі; кері байланыс; логика; білім; ...}*, сонымен қатар концепттер арасындағы байланыстардан тұрады, «*Кері байланыс – нейрондық желінің* қасиеті».

Онтологияны рольдік кластерлеуде қолдану. Онтология концепттерін «нысан», «құрал», «қасиет» және «іс-әрекет» кластары (рольдері) бойынша рольдік кластерлеуде осы концепттерді (түйіндік ұғымдарды) әр түрлі мағыналық категорияларға орналастыру қарастырылады. Мұндай категория концепттері «қарапайым» деп аталады. Рольдердің мүмкін болатын комбинацияларын «күрделі» концепттерді құруда қолдану болады. Айталық, «компьютер жылдамдығын талдау әдісі» сияқты сөз тіркесін жоғарыда келтірілген рольдерге тән 4 қарапайым концепттен тұратын күрделі концепт ретінде қарастыруға болады. Бұл мысалда «әдіс» – құрал, «талдау» – іс-әрекет, «жылдамдық» – қасиет, «компьютер» – нысан ретінде болып тұр[1].

Құжаттар бойынша ақпараттық іздестіру есептерінде осыған ұқсас әдістерді қолдану біршама дәрежеде сұранысты құрайтын ұғымдар семантикасын есепке алуға мүмкіндік береді және іздеу дәлдігін жоғарылатады. «Әдісті талдау» деген сұратуда іздестіру жүйесі «әдіс»-ті – нысан ретінде, және «талдау»-ды – іс-әрекет ретінде түсінеді. Бұл ретте қажет емес құжаттар қарастырылмайды. Рольдік кластерлеуді қолдану іздестіру толықтығын жоғарылатуға қабілетті.

Онтологияны классификациялау мен кластерлеу есептерінде қолдану. Құжаттарды классификациялау мен кластерлеу есептерінде онтологияны қолдану табысты нәтижелерге жеткізіп отыр. Келтірілген жағдайлардың барлығында онтологияны қолдану пәндік саланы айқындайтын маңызды концепттер жинағын ұсынады. Олардың қолданылуы пәндік салаға жатпайтын ұғымдарды талдауға кететін машиналық уақытты жоғалтпауға мүмкіндік береді, ал классификациялауда – классификатордың аса шығындық оқытуларын жүргізбейді, себебі классификатор құрастырылған онтологиямен беріледі. Аталған есептерді шешу сапасы құрастырылған онтологияның сапасы мен толықтығына тікелей байланысты болады.

Құжаттық деректер қорын пайдаланатын мамандарға жұмыс барысында құжаттардың барлық жиынтығы емес, тек оны қызықтыратын пәндік салаға сәйкес құжаттар ғана қажет болуы мүмкін. Мұндайда, деректер қорындағы құжаттарды категориялар бойынша классификациялау есебі өзекті болып табылады. Мысал келтірсек, құжаттарды классификациялау спамдарды фильтрлеу есептерінде, хаттарды тақырыптар бойынша таратуда, электронды сауда жүйелерінде және де басқа көптеген интеллектуальды жүйелерде ерекше орын алады. Сонымен қатар, ақпараттық іздестіру есептерінде құжатты белгілі бір класқа алдын-ала топтастыру, сұраныс тақырыбына жатпайтын құжаттарды алып тастауға мүмкіндік береді, әрі уақыт пен есептеу қорларын үнемдейді.

Құжаттарды тақырыптар бойынша қолмен орналастыру классификациялаудың алғашқы әдісі болып табылады. Бірақ, бүгінгі таңда өңдеуге болатын құжаттардың саны өте көп, ал бұл сарапшылар жұмысы барысында, пайдамен салыстыруға келмейтін құралдар мен уақыт шығындарына әкеп соғады. Сондықтан, 1960 жылдардан бастап мәтіндік құжаттарды автоматты түрде классификациялау мәселелері үлкен қызығушылыққа ие болып келеді. Бұл саладағы сарапшы жұмысын автоматтандыруға арналған бастапқы тәсілдер, мәтінді өңдеу жүйелеріне *«егер – онда»* түріндегі ережелерді жазудан тұрды, яғни сарапшы берген шарт орындалған жағдайда құжат нақты тақырыпқа бөлініп отырды. Классификациялау шарты мынадай түрде болды: *Егер (ДНФ) → Онда (категория)*, мұндағы, *ДНФ* – дизъюнктивті нормальды формада өрнектелген шарт, ал *категория* – бұл ДНФ ақиқат болғанда құжатты орналастыратын тақырыптар. Бұл әдістің қарапайым және тиімді екендігі көрініп тұр, бірақ ережелерді жазу және олардың өзектілігін негіздеу үшін сарапшы жұмысы талап етіледі.

Өткен ғасырдың 90-жылдарының басында мұндай ережелер машиналық оқыту әдістерімен ығыстырылды. Бұл әдістердің артықшылығы, көрініп тұрғандай, жүйелер сарапшының қатысуын талап етпейді және классификациялау ережесін жазуға мұқтаж емес. Ережелерді оқытатын таңдамалар негізіндегі жүйелер құрастырады. Қазіргі кезде, классификациялау есептерін шешуде «қарапайым» байес классифкаторы, Роккио әдісі, «k жақын көршілестер» әдісі, тіректік вектор әдісі және осы әдістердің түрлі нұсқалары (модификациялары) аса танымал болып отыр. Ықтималды байес классификаторынан басқа әдістердің барлығы құжаттың векторлық бейнеленуін қолданады, оның мазмұны құжат ішіне кіретін терминдердің векторы түрінде ұсынылады. Классификатор – бұл маңызды құжат, оның векторы оқыту кезеңінде құрылады және термин салмақтарының мәнін орташа мәнге келтіруден тұрады. Жоғарыда келтірілген әдістердің барлығы ортақ қасиеттерге ие, тек вектор-классификаторды оқыту және құрастыру әдісімен ерекшеленеді. Екі вектор арасындағы бұрышты олардың ұқсас дәрежелері ретінде есептеу классификацияның өзі болып табылады: егер құжат векторы классификатор векторына жақындау болса, онда құжат сол берілген категорияға жатқызылады.

Егер классификациялау үшін пәндік саланың онтологиясы қолданылатын болса, онда құжат векторын онтология векторының өзімен салыстыруға болады. Мұнда машиналық оқытудың классикалық әдістерінен екі маңызды айырмашылық байқалады. Бірінші айырмашылық: онтологияны қолдану классификаторды оқыту кезеңінен бас тартуға мүмкіндік береді. Пәндік саланы онтология түрінде сипаттаудың өзі классификатор болып табылады, сондықтан оқытатын таңдамалардан орташа мәнге ие («орташаланған») құжатты құрастыру үшін уақыт пен есептеу қорлары жұмсалмайды. Екінші айырмашылық: қарастырылатын онтологияға кірген терминдер ғана құжат векторына кірістіріледі. Бұл дегеніміз, онтология концепттерінің жинағына кірмейтін ұғымдар, терминдердің салмақтарын есептеу процесінен алынып тасталады. Сонымен қатар, онтология түріндегі классификатор «орташаланған» құжат түріндегі классификатордан ерекшеленеді. Екі жағдайда да классификатор пәндік салаға сәйкес «эталонды» құжат моделі болып табылады. Егер ол «орташаланған» құжат болып табылса, онда оның құрамына құжаттарда қолданылған, бірақ сипаттайтын бөлімге қатысы жоқ терминдер кіруі мүмкін. Онтология жағдайында, керісінше, ешқандай артық ұғымдарсыз, классификатор пәндік саланың сипаттамасы болып табылады. Жалпы айтқанда, бұл классификатор – түрлі жүйелер мен түрлі есептер құрамында қолдану тұрғысынан қарастырғанда анағұрлым әмбебап.

Құжаттың класқа (онтологияға) сәйкес келу дәрежесі құжатта табылған, берілген онтологиядағы барлық терминдер салмақтарының қосындысы ретінде есептеледі [2]:

$$R_{dC} = \sum_{t \in C} w_{td} \ ,$$

мұндағы, R_{dC} – d құжатының C кластерге сәйкестік дәрежесі, w_{td} – d құжатындағы t терминінің салмағы.

Онтологиядағы концепттердің рольдер бойынша бөлінгенін ескере отырып, рольдер үшін әр түрлі салмақтар енгізу қажет, сонымен бірге күрделі концепттерді бөлек өңдеу керек. Өлшеудің келесі әдісі біршама жақсы:

$$w_{td} = \begin{cases} tf, & \textit{егер концепт қарапайым және оның ролі «нысан»,} \\ 0.1 \cdot tf, & \textit{егер концепт қарапайым,} \\ (1+k)*tf, & \textit{егер концепт күрделі,} \end{cases} \qquad (1)$$

мұндағы, tf – концепттің құжатқа ену саны, k – концепттің күрделілігін ескеретін коэффициент.

Пәндік сала «нысан» типіндегі концептпен анағұрлым толық сипатталады деген болжам бар, ал қалған типтегі концепттер әр түрлі пәндік салаға қатысты инвариантты болуы мүмкін. Бұл тұжырым классификаторды тексеру барысында эксперименталды түрде расталған – егер құжатта кездескен барлық қарапайым концепттерге олардың мәтінге ену санына сәйкес келетін тең салмақтары меншіктелсе, онда құжат бірден тең ықтималдылықпен барлық кластарға жатқызылатын болады. Бұдан шығатын қорытынды, нысан – ұғымын анағұрлым маңызды ету керек, ал қалғандарына салмақтарды басқа сызбанұсқа бойынша меншіктеген дұрыс.

(1)-ші формуладан көрініп тұрғандай өлшеудің ұсынылған сызбанұсқасы бұл пікірді ескереді. Сонымен қатар, (1)-формулада күрделі терминдердің салмақтарын жай терминдердің санына байланысты жоғарылату ескерілген.

Қорытындылай келсек, онтологияны қолданатын классификаторлар тиімді болып табылады. Онтологияны қолдану арқылы дұрыс классификацияланған құжаттар пайызы едәуір жоғары болып отыр. Мұндай жағдайда классификациялау сапасы құрылған онтологияға тікелей қатысты болады. Онтологиялар білімдерді басқарудың басқа да есептерінде қолданылуы мүмкін, оның тиімділігі оларды құру мен қолдау шығындары тұрғысынан дәлелденіп отыр. Құжатты тақырыпқа жатқызу ондағы нақты терминдердің болуымен ғана емес, сонымен қатар, құжат пен онтологияның жақындық шамасын есептеу негізінде жүргізіледі. Сондықтан да, қарастырылып отырған тәсіл құжаттарды классификациялауда бұрыннан бар әдістердің тиімділігі мен жаңадан шыққан әдістердің әмбепаптығын өзіне біріктіреді.

Әдебиеттер

1. Норенков И.П. Задачи управления знаниями, извлекаемыми из текстовых документов. // Электронное научно-техническое издание «Наука и образование», 2011, 9.

2. Bevainyte A., Butenas L. Document classification using weighted ontology// Materials Physics and Mechanics, 2010, №9

Г. ШЫНАТАЙ

Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана

ҚАЗАҚ ТІЛІНДЕГІ СӨЗ ТІРКЕСТЕРДІ ӨҢДЕУ

Қазақ тіліндегі мәтінді,сөйлемді, сөз тіркестерін математикалық лингвистика әдісімен зерттеу және сөйлемдерді талдау мен құруды автоматтандыру проблемасы қазіргі кезде лингвистика және информатика саласында өте өзекті болып табылады.Соның ішінде сөйлемдерді семантикалық талдау мәселесі компьютерлік лингвистика бағытында маңызды.Өйткені бұл мәселенің шешімі авторлық құқықты қорғаумен тікелей байланысты,

мысалы антиплагиат.Бұл мәселені шешуге арналған әдістер мен программалар бар, алайда мағынасы бойынша қазақ тіліндегі мәтіндерді салыстыратын жетілдірілген программа мен оның теориялық негіздемесі жоқ.

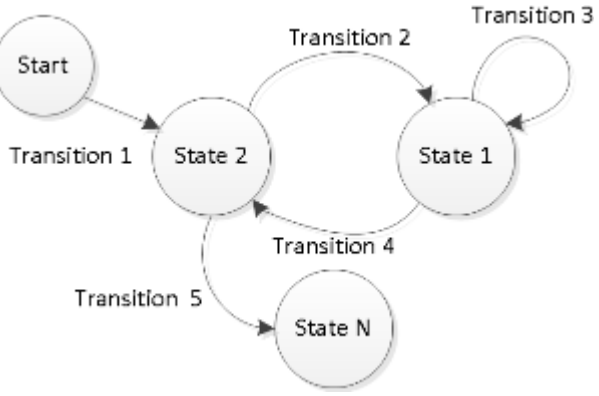
Қазіргі таңда Хэмминг және Хопфилд нейрожелілерін қолдану арқылы лингвистика саласының көптеген өзекті мәселелерін шешуге болады.Семантикалық талдау барысында сөз тіркестерін өңдеу қарастырылады. Сөз тіркесі сөздердің бір бірімен тіркесуімен жасалады. Сөздер зат пен құбылысты, сапа мен белгіні немесе іс-әрекетті атайды. Сөз тіркестерінде зат пен құбылыс, сапа мен белгі немесе іс-әрекеттер жеке сөздердегідей дара күйінде емес, өзара бір-бірімен байланысты болады. Сөз тіркестерінің синтаксисі сөздердің өзара тіркесу қабілеттілігі, тіркесу тәсілдері мен сұлбаларын (формаларын) және сөз тіркестерінің құрамы мен түрлерін морфологиямен тығыз байланыста қарастырады. Онда сөздерді сөз тіркесі мен сөйлемнің бөлшектері ретінде, ал жалғауларды сөздердің бір бірімен қиюластырып тұратын морфологиялық-синтаксистік категория ретінде зерттейді.

Сөз тіркестері грамматикалық тәсілдер арқылы байланысқан кем дегенде екі толық мағыналы сөзден құралады. Құрастырушы сыңарлар бірі *бағыныңқы,* екіншісі *басыңқы* мүше ретінде қызмет атқарады. Сыңарлар бір бірімен өзара сабақтасып, белгілі бір мағыналық және синтаксистік заңдылықтар негізінде байланысады. Сөз тіркесі атау құралы ретінде негізгі сөз (басыңқы) арқылы затты, құбылысты, үрдісті (процесті), сапаны белгілейді.

Сөздер жалғаулар арқылы байланысуы мүмкін. Сөздер байланысуының бес түрі бар: *қабысу*, *матасу*, *меңгеру, қиысу*, *жанасу*.Стратегияларда ақырлы автоматпен деңгей бойынша бөлу жүріледі.

Жұмысымызда қазақ тіліндегі сөз тіркестерін өңдеу үшін ақырлы автоматты қолданудың сипаттамалары анықталады Осы мақсатты жүзеге асыру үшін келесі міндеттер қойылды:

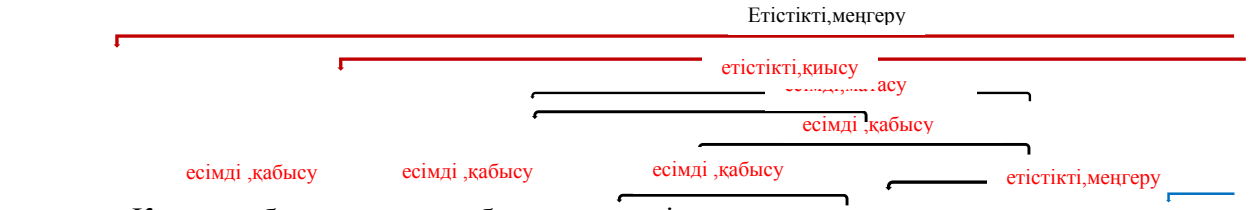
- мәтінді сөз, сөйлем, азат жолдарға бөлу;
- ішкі құрылымдарын білместен бұл деңгейлерге біробразды қарау;
- бұл деңгейлерді бастапқы күйінде қалтырмау керек, себебі мәтіндердің көп мөлшерде көшірілімі жасалынады;



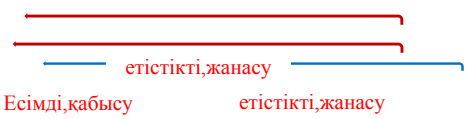
Деңгейлер индекстермен белгіленген тізім болып сақталынады.Логикалық түрде мұнда диапазондар концепциясы кірістіріледі (Range, [1], [2], [3]); Диапазондарда бастапқы және ақырғы индекстері бар: RangeItem(BeginIndex, EndIndex), немесе бастапқы индекс және жылжу саны: Range(BeginIndex, Count). Индекстер мәтіндегі символдардың позициясымен нақты сәйкес келеді. Егер толық мәтінді белгілейтін болсақ, онда келесі түрде болады RangeItem(1, Length(Text)).Ал егер бізге мәтін ортасындагы азат жол керек болса RangeItem(312031, 312355) белгіленуі қолданылады. Сөз тіркестерінде қолданылуы мысалы: есімді (1.1), етістікті (1.2), меңгеру (2.1.1), қабысу(2.1.2), матасу(2.1.3),қиысу (2.1.4),меңгеру (2.2.1), қабысу(2.2.2), қиысу(2.2.2) .Қабыса байланысқан есімді сөз тіркестері зат есім мен зат есімнің тіркесі түрінде (3.1.2.1.1,3.1.2.2.1), сын есім мен зат есімнің тіркесі түрінде (3.1.2.1.2, 3.1.2.2.2.1),есімше мен зат есім тіркесі түрінде (3.1.2.1.4, 3.1.2.2.4.1).Тізімдер диапазоны мәтіннің тура көшірілімі жасалынбаған мәтін бөлімдері болып көрсетіледі.

1	Сөз тіркестері
1.1	Есімді
1.2	Етістікті
2.1.1	Меңгеру
2.1.2	Қабысу
2.1.3	Матасу

2.1.4		Қиысу
2.2.1		Меңгеру
2.2.2		Қабысу
2.2.3		Қиысу
2.2.4		Жанасу
3.1.2.1.1	Зат есім	3.1.2.2.1.1 Зат есім
3.1.2.1.2	Сын есім	3.1.2.2.2.1 Зат есім 3.1.2.2.2.3 Сан есім 3.1.2.2.2.4 Есімдік
3.1.2.1.3	Сан есім	3.1.2.2.3.1 Зат есім 3.1.2.2.3.2 Сын есім 3.1.2.2.3.4 Есімдік
3.1.2.1.4	Есімше	3.1.2.2.4.1 Зат есім 3.1.2.2.4.2 Сын есім 3.1.2.2.4.3 Сан есім 3.1.2.2.4.4 Есімдік
3.1.2.1.5	Үстеу	3.1.2.2.5.1 Зат есім 3.1.2.2.5.2 Сын есім 3.1.2.2.5.3 Сан есім 3.1.2.2.5.4 Есімдік



Қиысуда бағыныңқы сөз басыңқы сөздің грамматикалық мағынасына, тұлғасына бейімделе



тиісті жалғауда айтылып байланысады.

1. Бағыныңқы сөз – есімді, орын тартібі арқылы, қабысу, сын есім + зат есім.
2. Басыңқы сөздің –есімді, орын тартібі арқылы, қабысу, сын есім + зат есім+ілік септік жалғауы.
3. Грамматикалық мағынасына – есімді, қабысу, сын есім+ зат есім.
4. Грамматикалық тұлғасына – есімді, қабысу, сын есім+ зат есім.
5. Мағынасына бейімделе – етістікті, меңгеру, жалғау арқылы, зат есім + барыс септік жалғауы + етістік.
6. Тұлғасына бейімделе – етістікті, меңгеру, жалғау арқылы, зат есім + барыс септік жалғауы + етістік.
7. Сөздің мағынасына – есімді, жалғау арқылы, матасу, зат есім + ілік септік + зат есім
8. Сөздің тұлғасына – есімді, жалғау арқылы, матасу, зат есім + ілік септік + зат есім
9. Тиісті жалғауда - есімді, қабысу, сын есім + зат есім.
10. Бейімделе байланысады - етістікті, жанасу, үстеу + етістік.
11. Қиысуда байланысады – етістікті, жалғау арқылы, меңгеру, зат есім+ етістік.
12. Сөз байланысады – етістікті, қиысу, зат есім + етістік+ тәуелдік жалғауының III-жағы, жекеше

Жұмыс барысында диапазондардан абстракцияны алға қою мәселесі шығып отыр. Жалпы мәтіннің бірінші және соңғы жолдарын көрсететін диапазондар тізіміне қарағанда, біз тек соның негізіндегі мәтінді көре алуымыз керек. Сөздерді анықтау басты мәселе емес. Сөйлемдерді анықтау кезінде кейбір қиындықтармен кездесеміз. Мысалы сөйлемнің соңын анықтайтын легальді белгілердің,яғни нүкте, сұрақ, леп белгілер мәселесі.Міне осындай легальді белгілердің қайталанып,қатар келуі немесе бір сөйлемнің ішінде бірнеше рет кездесуі сөйлемдер санының өз санынан артық болып кетуіне әкеледі. Бұл мәселені шешудің бірден бір жол – ақырлы автомат. Ақырлы автоматтың қолданылуының бірегей себебі,келіп түскен символдар тізімінің құрылымын анықтайды: сөз, тыныс белгілер, функциялар, құрылымдар, әдісі және өрісі бар толық кластар. Осындай тәртіппен код анализаторлары, компиляторлар, компьютерлер жұмыс істейді.Қорытындылай келсек, алынған ғылыми нәтижелердің ғылыми-практикалық құндылығы қазақ тілінің теориясын жетілдіруге және оның қолданыс аясын кеңейтуге ықпал жасайтындығы. Жұмыс нәтижесі қазақ тілінде мәтіндік процессорлар, ақпараттық технологиялар мен жүйелер және басқа да программалық дестелер жасауда өз үлесін қосары сөзсіз.

Әдебиеттер

1. «Мәтіндік анализатор» жобасының негіздемелері (Borland C++ Builder 6.0)

2. Ақырлы автоматтарған арналған мәтінді деңгейлерге бөлетін өтілім кестелері

3. Серғалиев М.С. Сөйлем және сөз тіркесі. Материлы учебно-теор. конф.Уақыт ұштастырған желі: көшпелілер әлемі XX ғасырдағы тарих ғылымында. -Астана, 2004, -Б.20-25

4. Балақаев М. Б. Қазіргі қазақ тілі: Сөз тіркесі мен жай сөйлем синтаксисі. – Астана: Л.Н.Гумилев атындағы ЕҰУ, 2006. -237б.

СӨЙЛЕУЛЕРДІ СИНТЕЗДЕУ ЖӘНЕ ТАНУ ЖҮЙЕЛЕРІ
СИСТЕМЫ РАСПОЗНАВАНИЯ И СИНТЕЗА РЕЧИ
SPEECH RECOGNITION AND SYNTHESIS SYSTEMS

Т.И. ИБРАГИМОВ[1], Ф.И. САЛИМОВ[2]

[1]Казанский федеральный университет, г.Казань, Россия
[2]Казанский федеральный университет,
Институт прикладной семиотики АН РТ, г.Казань, Россия

ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ СИНТЕЗА ТАТАРСКОЙ РЕЧИ ПО ОРФОГРАФИЧЕСКОМУ ТЕКСТУ

В настоящее время проблема синтеза речи считается, в основном, решенной. Существующие синтезаторы довольно качественно произносят тексты, написанные на многих европейских языках, в том числе и на русском языке[1-4]. Насколько нам известно, синтезаторов речи для тюркских языков, нашедших широкое распространение, сегодня не существует. В данной публикации мы остановимся на трудностях, с котрыми столкнулись при разработке синтезатора татарской речи.

Первая версия системы конкатенативного синтеза татарской речи по произвольному орфографическому тексту была создана еще в 2009 году. И хотя качество синтезированной речи по мнению большинства экспертов было удовлетворительным, по многим причинам эта разработка осталась на уровне опытного лабораторного образца. В настоящее время в рамках совместного Татаро-Швейцарского проекта “Ана-теле” завершается разработка новой версии синтезатора, предназначенного для работы в системе дистанционного обучения татарскому языку. Поскольку лингвистическое обеспечение при создании качественных систем синтеза является определяющим[3], в данной статье обобщается опыт, приобретенный при разработке татарского синтезатора.

Синтезатор татарской речи предназначен для озвучивания произвольных татарских текстов, относится к числу конкатенативных. Процесс синтеза речи в рамках этой модели можно представить как склеивание по правилам фонетики заранее озвученных фрагментов языка в слова и затем - в предложения.

На рисунке ниже приведена общая схема синтезатора, в которой указывается какие действия при наличии каких ресурсов и в какой последовательности должны выполняться для преобразования произвольного предложения в звучащую фразу.

[3] Попытки создания универсальных синтезаторов, покрывающих одновременно несколько различных языков, не увенчались успехом. Как известно, языки отличаются не только грамматическими, но и фонетико-фонологическими и ритмико-просодическими системами. По этой причине синтезаторы речи, разработанные на базе систем индоевропейских языков, оказались неприменимыми в качестве преобразователей татарских текстов в речь.

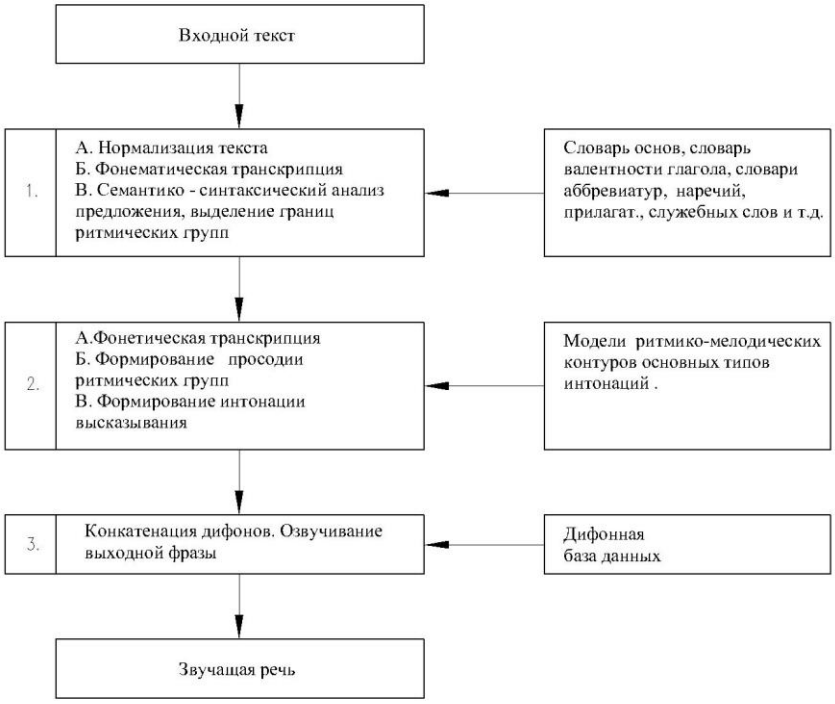


Рис1. Блок-схема функционирования синтезатора татарской речи

1**.** Нормализация включает предварительную обработку входного текста: расшифровку аббревиатур, перевод чисел в числительные, удаление или расшифровку различного рода символов (%, №, $) и пр. Нормализация производится по определенным правилам с помощью заранее подготовленных словарей. Выяснилось, что многие аббревиатуры в письменном татарском языке употребляются в русской версии (НИИ, УВД, США). При доминирующем положении в обществе русского языка перевод их на татарский язык может инициировать образование языкового барьера. При создании словаря аббревиатур рассматривались в основном именно такие варианты, при этом расшифровка таких терминов переводилось на татарский язык. Определенная сложность возникает при переводе числа в числительное. Например, цифра 10 в зависимости от контекста может быть расшифрована как «ун» или «унынчы». Для расшифровки таких фрагментов необходим анализ контекста ближайшего окружения соответствующего числа. Кроме того, полноценный синтезатор должен уметь воспроизводить также математические формулы. Эта задача относится к одновременной реализации в рамках одной системы синтеза нескольких языков, кроме того система кодирования формул математического языка может быть различной, что выводит за рамки простого перевода формул в текстовое представление. По этой причине такая задача является достаточно сложной и в данной версии не рассматривалась.

2. Большие трудности в создании системы синтеза речи по орфографическому тексту возникают при фонемной транскрипции входной фразы. Под фонемной транскрипцией понимается перевод единиц орфографического текста в звуковые единицы языка. Сложность преобразования «графема–фонема» для разных языков различна. Что касается татарского языка, то трудности обусловлены, прежде всего, ненаучным характером его орфографии. Помимо множества незначительных двусмысленностей в правилах типа «… некатегоричное будущее время глаголов образуется суффиксом -ыр/ер, а в *некоторых случаях* суффиксом –ар/әр (ит+әр)», в орфографии значится такое правило: «слова, заимствованные письменным путем из русского языка или через него, пишутся так, как принято в русской орфографии» [6]. Использование данного ïðàâèëà ïðèâåëî к тому, что в ñëîâàðÿõ òàòàðñêîãî языка около 10-15% слов ñîõðàíÿþò ÷åðòû àðàáñêîãî письма, а 35- 40% слов пишутся ïî-ðóññêè. Ðàñòîæäåíèÿ

ìåæäó ãðàôèêîé и звуковой системой ÿçûêà îêàçàëиñü çíà÷èòåëüíûì è òðóäíî ïîääàþùèìñÿ ñèñòåìàòèçàöèè. В связи с таким положением дел при создании транскрибиратора пришлось выбирать между следующими альтернативами:

- íàó÷èòü ñèíòåçàòîð произносить àðàáñêèå ñëîâà ñ àðàáñêèì àêöåíòîì, ðóññêèå ñëîâà ïîðóññêè, à òàòàðñêèå – ïî-òàòàðñêè
- îçâó÷èâàòü çàèìñòâîâàíèÿ в соответствии с ôîíåòè÷åñêîé системой òàòàðñêîãî ÿçûêà.

Современная орфография придерживается первого варианта. Для того, чтобы ïðîèçíîñèòü ðóññêèå заимствования на языке îðèãèíàëà, фонетическая система òàòàðñêîãî языка была дополнена гласными фонемами /ō/, /û/, /ē/, согласными - /ö/, /ù/, /q/, /g/, /в/, а òàêæå специальными символами "ь" и "ъ". В результате современная графика татарского языка представляет эклектику трех языковых систем - собственно татарской, а также русской и арабской. Но это еще не все. Дополнение фонетической системы татарского языка тремя гласными и пятью согласными оказалось недостаточным для произношения всех заимствований так, как в оригинале. Потребуется, по крайней мере, дополнительное включение в фонетическую систему татарского языка фонем /а/ и /ль/, а в оптимальном случае - всех русских гласных и согласных. Так, в словах [*кабинетка*] и [*гаражга*] фонемы /а/ в начальном и конечным слогах оказались невзаимозаменяемыми, а слова [*тол*], [*толь*] и [*тӧл*] контрастируют не только по гласным фонемам. Данное обстоятельство вынудило разработчиков отказаться от предложенных лингвистами орфоэпических норм и выбрать второй путь - озвучивать заимствования в согласии с фонетической системой òàòàðñêîãî языка.

Данный подход предполагает осуществление фонематической транскрипции в два этапа. На первом этапе необходимо определить - является ли входное слово заимствованным или исконно татарским, на втором – с помощью соответствующего алгоритма выполнить саму транскрипцию. Такой способ (выполнение транскрипции в два этапа) вполне естественен. Действительно, чтобы правильно транскрибировать, к примеру, слова *Казан и Казань, капкан (охотничья принадлежность)* и *капкан (проглотил)* необходимо знать – к каким языкам они принадлежат, а также обладать знаниями о фонетико-фонологических системах татарского и русского языков.

Вторая часть задачи, т.е. собственно фонематическая транскрипция, имеет непосредственное отношение к фонетике и фонологии языка и требует теоретического осмысления основных положений соответствующих разделов грамматики.

Основной единицей фонологии является фонема. В фонетике утверждается наличие в татарском языке 12 гласных и 28 согласных. В татарском языкознании фонема определяется как «*неделимые звуковые единицы языка, которые служат для построения словоформ и для различения звуковых видов*»[6]. Выделение фонем производится на основании правила: «*Если два звука, находящиеся в одинаковых фонетических позициях (одинаковость фонетической позиции включает и одинаковость контекста), позволяют различить смысл слов, то они являются различными фонемами*». При этом «одинаковость» или «неодинаковость» фонетической позиции звука устанавливается на основе графического представления слов. Предполагается, что смысл слова в речевой коммуникации устанавливается в результате опознавания составляющих его фонем. Так, считается, что в паре "*тӧл (вдовый)*" - "*тōл (взрывчатое вещество)*" гласные ӧ и ō, имеют одинаковую фонетическую позицию, а значения слов различаются за счёт фонемной идентификации звуков. На основе данных воззрений в татарском языке выделяются широкие (произносимые при более широком растворе рта) гласные /ō/, /ы/, /ēе/, а также согласные фонемы /ц/, /щ/, /в/, /к’/, /г’/.

Между тем трудно говорить о фонетическом равенстве слов, принадлежащим разным звуковым системам. Как пишет Л.Р. Зиндер: "*Совокупность фонем данного языка представляет не простой набор разрозненных единиц. Фонемы находятся в определенных отношениях друг к другу, определенным образом связаны между собой, составляя известную систему*" [7].

Что касается различения смысла слова, то «достаточно часто фонемное решение принимается не на основе фонетической информации, а в результате опознавания морфологической структуры слова, синтаксических связей в словосочетании, наконец, смысла всего высказывания» [8]

Для наших целей[4] фонемный состав татарского языка определялся на основе теоретических положений Л.В. Щербы, В.А. Богородицкого и их последователей. В [9] Л.В. Щерба определяет фонему как *"... кратчайшее общее фонетическое представление данного языка, способное ассоциироваться со смысловыми представлениями и дифференцировать слова и могущее быть выделяемо в речи без искажения фонетического состава слова..."*.

Опираясь на данное понимание сущности фонемы и учитывая роль речевых звуков в передаче значений, при создании транскриптора в звуковой системе татарского языка было выделено 9 гласных (/а/, /ә/, /о/, /ө/, /у/, /ү/, /ы/, /е/, /и/) и 23 согласных фонем (/б/, /в/, /г/, /д/, /ж/, /з/, /й/, /к/, /л/, /м/, /н/, /п/, /р/, /с/, /т/, /ф/, /х/, /ш/, /ч/, /ж/, /ң/, /h/, /w/. Очевидно, что состав фонем любого языка должен удовлетворять условиям необходимости и достаточности для различения лексических и грамматических значений. Проверка этих положений на множестве грамматических морфем показала, что данный состав фонем удовлетворяет этим условиям.

Вторая проблема, связанная с транскрипцией орфографических текстов татарского языка, заключается в решении вопроса – как транскрибировать заимствования. Другими словами, "научить" ли синтезатор произносить арабские слова с арабским акцентом, русские слова - по-русски, а татарские - по-татарски либо озвучивать заимствования согласно фонетике спонтанной, не укладывающейся в рамки нормативной граматики, татарской речи.

Если бы заимствования в лексике татарского языка были единичными, то вопрос - как писать заимствования – потерял бы свою актуальность. Но они многочисленны. В лексике современного татарского языка русские заимствования составляют более 40% (в составляющемся электронном словаре из 26000 основ слов 12000 оказались неадаптированными русскими заимствованиями), арабские заимствования (неадаптированные или частично адаптированные) составляют около 10%[11]. Тексты деловых бумаг, научного и публицистического стилей приблизительно на 25% покрываются заимствованиями из русского языка. Произносить эти слова по-русски при помощи включенных в фонологическую систему татарского языка семи – восьми русских фонем не представляется возможным. В звуковой картине слова корабль [карàбъл’] по-русски звучит не только /к/, но и ударная и безударная /а/, а также согласные /б/ и /л/. По-видимому, «пересаживать» часть фонем одной системы звуков в другую систему невозможно. Как пишет В.А.Богородицкий: *"... звуковая система каждого языка представляет не простое собрание звуков, но гармонически связанное целое, ...полагаю, что только что указанная особенность физиологическая и акустическая гласного /а/ (более задняя артикуляция по сравнению с русской /а/.) может находиться в связи с более глубоким произношением татарского /к/ и /г/*" [10].

При написании алгоритма фонематической транскрибирации, было принято решение, что действие закона сингармонизма, лежащее в основе фонетической системы татарского языка, распространяется в заимствованиях не на целое слово, а на слог. Слог как произносительная единица может быть только целиком палатилизованным или целиком непалатилизованным, а в слове могут встречаться как палатилизованные, так и непалатилизованные слоги. С учетом вышесказанного, предложение "*Комендантка слесарьларны эзлəп ике тапкыр гаражга барырга туры килде*. (В поисках слесарей коменданту пришлось дважды сходить в гараж)" после транскрипции приобрело вид [*Кəминдантка эслисарларны эзлəп ике тапкыр гəражга барырга туры килде*].

Значительное затруднение в произнесении заимствований так, как они звучат в языках, из которых заимствуются, вносит и то обстоятельство, что образование грамматических форм и

[4] изучение звуковой системы языка в целях внедрения его в новые информационные технологии

сложных слов от основ-заимствований происходит путем присоединения аффиксов к основам из различных языков (*колхоз+чыларга, аэро+чана+лар*). Согласование основ и аффиксов на сегментном и супрасегментном уровнях, как правило, не представляется возможным. Алгоритм фонематической транскрипции разбивает словоформу на три части *основа+основа+аффиксное окончание* (*газ+үлчәгеч+ләрне*), причем первая основа может отсутствовать[5]. Для каждой основы определяется язык происхождения, и далее транскрибирование каждой основы осуществляется по соответствующему набору правил. Что же касается окончания, то мягкость/твердость гласных определяется анализом последней гласной предваряющей основы. Заметим, что если в основе словоформы встречается единственная гласная /и/, то вид окончания определяется принадлежностью языку-основы(*спирт [спирт+ны], грипп [грип+ны], кил [кил+ә+ләр]*).

С учетом изложенного алгоритм фонематической транскрипции заимствований разрабатывался на базе установленного состава фонем и в соответствии с фонетикой спонтанной татарской речи. Фонетика спонтанной речи (под термином «татарский народный говор») достаточно полно описано В.А.Богородицким в [10].

3. Фонетическая транскрипция состоит в преобразовании фонематически транскрибированного слова входного предложения к форме его звучания (с учетом контекстного окружения) на выходной фразе. Алгоритм транскрипции состоит из трех частей, включающих формирование звуковой оболочки слогов, стыков слогов и стыков слов, объединенных в ритмические группы[12]. В формировании звуковой оболочки слогов основное внимание уделялось изменениям длительностей гласных в речевом потоке в зависимости от их характеристик (краткий/долгий) и позиции в слове, а также в РГ.

Наиболее значимыми в этом плане являются изменения в длительностях кратких гласных, таких как [ы], [е], [о], [ө]. Так, гласные сокращаются полностью (*тешем – тшем, белән – блән*) в тех случаях, когда[13]:

а) гласные [ы] и [е] входят в состав слогов типа СГ,

б) согласный (С) является взрывным либо шипящим,

в) данный слог находится в начале слова,

г) за слогом СГ следует слог, начинающийся на сонорный либо на шипящий.

Если не выполняется условие г), то гласный слога СГ сокращается почти до полного исчезновения (*чыдам [чдам], кыяр [кйар]*). Если не выполняется условие в), то возможны два варианта: 1) в случае, когда СГ употребляется в середине слова, гласные [ы] и [е] сокращаются лишь наполовину, 2) в случае, когда СГ употребляется в конце слова (ритмической группы), то сокращение вовсе и не происходит. Указанные явления могут быть объяснены с позиции действия фактора удобопроизнесения.

Алгоритм формирования звучания стыков слов, входящих в одну РГ, предназначен для описания фонетических явлений, наблюдаемых при объединении слов в одно фонетическое целое. В разработанном алгоритме рассматриваются случаи объединения слов, когда предшествующее слово заканчивается, а следующее начинается на одинаковые гласные, разные гласные, согласный и гласный, а также - на согласные. Алгоритм учитывает изменение качества речевого звука либо замену одного звука другим в составе пограничных фонем. Основное внимание при этом было уделено ассимиляциям пограничных фонем по признаку "звонкость - глухость". Примеры: *башсыз - башшыз* (безголовый), *тозсыз - тоссыз* (несоленый), *бүдка – буткы*, *аяк асты - аягасты* и т.д. В противоположность ассимиляциям по месту образования (*урман+лар=урман+нар*), ассимиляции по признаку "звонкость - глухость" являются более значимыми в обеспечении качественного синтеза.

4. Ритмико-интонационное оформление фразы является одной из основных и трудно решаемых задач в разработке любого синтезатора. В татарском языке в отсутствие словесного ударения основную роль в формировании ритмико-мелодической структуры

[5] Вообще говоря, основ может быть и более двух, но в этом случае транскрипция строится через словари

высказывания играет ритмическая группа (речевой такт, ритмическая структура, фонетическое слово), точнее - разбиение фразы на ритмические группы.

Исследования показали[14], что распределение, как интенсивности, так и длительности слогов в речевом такте подчиняется определенным закономерностям. Произнесение ритмической группы имеет более энергичное начало и несколько расслабленное завершение, вследствие чего длительности слогов к концу речевого такта несколько увеличиваются.

Так, анализ конечных ритмических групп во фразах с интонацией завершенности выявил следующие закономерности:

1. Длительность гласного конечного открытого слога в два раза больше длительности гласного начального слога данной ритмической группы. Интенсивность конечного слога, независимо от того является ли он открытым или закрытым, примерно на 25% меньше интенсивности начального слога. Интенсивность конечного открытого слога речевого такта, находящегося в конце фразы, в отличие от интенсивности аналогичного слога внутреннего речевого такта постепенно падает и сводится к нулю.
2. Когда конечной слог является закрытым, то различия в длительностях начального и конечного слогов ритмической группы не наблюдаются. Отличия по интенсивности сохраняются. Интенсивность последнего слога примерно на 25% ниже интенсивности первого слога.

Согласно нашему предположению, интонация общего вопроса также сосредоточена в конечной ритмической группе фразы. Отчасти это обусловлено особенностями татарского языка. В отличие от русского в татарском языке процедура организации общего вопроса, кроме интонации, предполагает участие грамматических форм – частицы *-мы/-ме* и *-мыни/-мени*. Данные частицы присоединяются к любой части речи, выполняющей в предложении функцию сказуемого. Согласно синтаксису татарского языка, сказуемое находится в конце предложения. Эти особенности позволили изучать в сопоставительном плане такие фразы как, например, "*Бу камыш исеме*? (Это запах камыша?)" и "*Бу килеш исеме*. (Это имя падежа.)", "*Ул арыш саламы*? (Он кладет рожь?)" и "*Бу арыш саламы*. (Это ржаная солома?)"

Выявлено, что вопросительная частица *-мы/-ме* отличается от неморфемных сочетаний "*мы*" и "*ме*" несколько большей длительностью и интенсивностью, а также конечным участком гласных "ы" и "е". У неморфемных сочетаний звучание данных гласных плавно переходит на нет, в то время как у частиц звучание заканчивается на том месте, в котором интенсивность звука является значительной.

В настоящее время просодический анализ производится на основе знаков препинания в предложении. Выделение границ всех ритмических групп предложения требует полного его синтаксического анализа.

5. В разработке синтезаторов речи конкатенативного типа важное значение имеет выбор речевой единицы в качестве исходного элемента конкатенации. Первоначально была предпринята попытка сформировать элементную базу из слогов. Расчет был таков, что слоги как наименьшие единицы артикуляции представляют более целостную структуру и количество их в татарском языке существенно меньше, чем в русском или английском языке[15]. Вскоре, однако выяснилось, количество слогов, выявленных на основе анализа изолированных слов значительно меньше того объема, которого мы получили бы на базе анализа предложений и фраз. Кроме того выяснилось, что слоги как просодические единицы несут в себе различного рода дополнительную окраску, которых трудно учесть при озвучивании текста. Современная версия синтезатора использует дифонную базу.

В процессе создания дифонной базы были выполнены следующие виды работ:

1. Построена таблица сочетаемостей фонем с учетом пробела между словами, состоящая из 33 строк (32 фонемы и плюс «пробел») и такого же количества столбцов.
2. Выделены фонемные сочетания, способные следовать в начале слова, середине слова и в его конечной позиции (на основе табличных данных). Общее количество данных трех классов фонемосочетаний составляет порядка 3000 единиц.

3. Выявлено, что помимо представительства указанных классов фонологических структур, элементная база синтезатора включает репрезентанты фонемосочетаний, встречающихся на стыках слов, в конечной позиции предложений, а также слова длиной в одну и две фонемы. (ә, ат, ит, өч, ки и т.д.).

4. Создание дифоной базы на основе выделенных фонемосочетаний производилось по следующей схеме:

а) В составе псевдофраз нужное фонемосочетание было озвучено диктором и оцифровано. Псевдофраза состояла из трех ритмических групп, одна из которых представляла 3–х или 4–х сложное квазислово (слово, лишенное смысла, но имеющее характерное для татарского языка звучание)

б) Программным путем устанавливались границы входящих в дифон фонем, а затем границы дифона.

в) Аудированием в различных контекстах проводилась оценка качества звучания дифона.

г) Выполнялась разметка периодов основного тона содержащихся в базе дифонов.

д) Производилась конвертация дифона в базу данных синтезатора.

На настоящее время элементная база синтезатора содержит 2370 дифонов. Увеличение элементной базы, замена отдельных дифонов более качественными по мере необходимости будут продолжаться.

Алгоритмы озвучивания построены на известной технологии TD-PSOLA. Этот подход позволяет производить модификацию просодических характеристик базовых элементов, размеченных по периодам основного тона.

Литература

1. Б.М.Лобанов, Л.И.Цирульник «Компьютерный синтез и клонирование речи». - Минск, «Белорусская Наука», 2008. - 316 с.
2. Зиновьева Н.В., Кривнова О.Ф., Захаров Л. М. Программный синтез русской речи (синтезатор «Агафон»). *Труды международного семинара Диалог'95 по компьютерной лингвистике и ее приложениям*. Казань, 1995.
3. Кривнова О.Ф., Зиновьева Н.В., Захаров Л.М., Строкин Г.С., Бабкин А.В. TTS Synthesis For Russian Language // *Web Journal of Formal, Computational & Cognitive Linguistics. N1*. 1997.
4. Чистиков П.Г., Хомицевич О.Г. Автоматическое определение границ предложений в потоковом режиме в системе распознавания русской речи // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. — 2011. — Спец. вып. Биометрические технологии. – С. 115-123.
5. Т.И. Ибрагимов, Ф.И. Салимов Из опыта построения синтезатора татарской речи, Тезисы докладов Международного симпозиума "Типология аргументной структуры и синтаксических отношений", Казань, "Отечество", 2004, с.334-336
6. Татарская грамматика. Казань, Татарское книжное издательство, 1993, т. 1, 581с.
7. Л.Р.Зиндер. Общая фонетика. - М.: Высшая школа, 1979.
8. Проблемы и методы экспериментально-фонетического анализа речи..Изд-во Ленинградского университета, 1980, 148с.
9. Л.В. Щерба Русские гласные в качественном и количественном отношении, СПб, 1912, 155с.
10. Богородицкий В.А. Введение в татарское языкознание. Казань, Татгосиздат, 1953.
11. Ибрагимов Т.И., Салимов Ф.И., Хусаинов Р.Р. Синтезатор татарской речи: вопросы транскрипции заимствований и планирование языка./ «Компьютерная лингвистика и интеллектуальные технологии». Тр. Международн. семинара Диалог-2002.т.2., М.:2002. с. 228-234
12. T.I. Ibragimov, F.I. Salimov, D.S. Suleymanov, R.R. Khusauinov The Exprerimental Version of the Tatar Speech Synthesizer, Interactive Systems: The Problems of Human - Computer Interaction, Ulyanovsk State Technical University, 2003, c.204-206

13. Ибрагимов Т.И., Салимов Ф.И., Сулейменов Д.Ш., Хусаинов Р.Р Синтезатор татарской речи: фонетический эллипсис и изменения речевых звуков на границе слов и слогов ./ Компьютерная лингвистика и интеллектуальные технологии». Тр. Международн. семинара Диалог-2003., М.:2003

14. Т.И. Ибрагимов, Ф.И.Салимов, М.Р.Сайхунов. Вариативность чтения текстов на татарском языке: паузирование. // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL – 2006. Казань, декабрь, 2006, Казань, 2007. С. 117 -121

15. Ибрагимов Т. И. Изучение образования слогов и структуры их сочетаний в татарском литературном языке. Автореф. дис. канд. филол. наук. - Казань, 1970

А.Ф. ХУСАИНОВ

Казанский (Приволжский) федеральный университет,
Институт прикладной семиотики Академии наук Республики Татарстан, г. Казань,
Российская Федерация

СИСТЕМА АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ФОНЕМ ТАТАРСКОГО ЯЗЫКА

Аннотация. В данной работе описывается система автоматического распознавания фонем татарского языка, процесс создания которой включает в себя проектирование и запись корпуса звучащей татарской речи и выделение значимых акустических особенностей татарской речи. Кроме того, предложен и реализован подход к построению системы автоматического транскрибирования татарских текстов. В качестве базовых единиц распознавания были использованы 57 фонем татарского языка, для каждой из которых была построена математическая модель, основанная на аппарате скрытых Марковских моделей. В конечном итоге, предлагаемая система показала 61%-ное качество распознавания фонем тестового корпуса.

Ключевые слова: фонемный распознаватель, корпус звучащей речи, татарский язык, фонетическая транскрипция.

Введение

Область речевых технологий представляет собой значимое и активно развивающееся направление научных исследований, которое, в целом, представляет собой процесс анализа речи, как звуковой волны, выделения из неё значимых параметров, и дальнейшего их использования для конкретного приложения. При этом направления использования речевых технологий могут существенно различаться. Так, принято выделять следующие основные направления:

- автоматическое распознавание речи;
- синтез речи;
- идентификация и верификация языка;
- идентификация и верификация диктора;
- распознавание эмоций диктора;
- распознавание тематики разговора.

Схематично структура области речевых технологий может быть представлена следующим образом (Рис. 1):

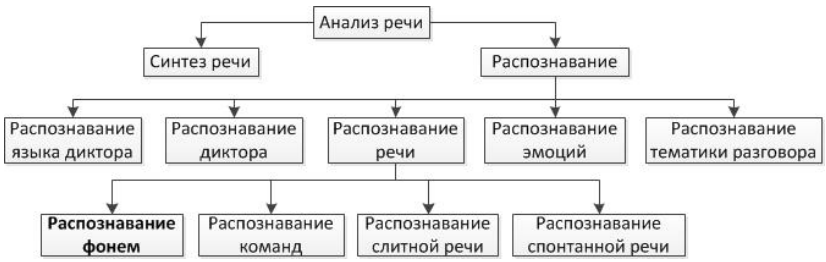


Рисунок 1. Структура области речевых технологий.

В данной работе рассматривается подход к решению одной из подзадач автоматического распознавания речи, а именно, распознавания фонем в контексте татарского языка. Система автоматического распознавания фонем может являться как самостоятельным элементом, например, при исследованиях в области экспериментальной фонетики, так и вспомогательным модулем при решении других задач распознавания речи.

Решение поставленной задачи распознавания татарских фонем осуществляется в четыре основных этапа:

1. проектирование и создание корпуса звучащей татарской речи одного диктора;
2. разработка и реализация правил транскрибирования татарских текстов;
3. создание акустических моделей фонем татарского языка;
4. программная реализация системы автоматического распознавания фонем.

Звуковой корпус

В качестве исходного материала при создании моделей фонем языка используется корпус звучащей речи. При этом необходимо наличие аннотации корпуса, включающей в себя текстовую и/или фонетическую разметку всех речевых фрагментов. Однако ручное фонетическое аннотирование является очень дорогостоящим и длительным процессом, а также требует наличие множества квалифицированных фонетистов, что делает затруднительным создание данного типа разметки для корпуса татарского речи. Альтернативным решением служит подход под название phoneme alignment, который позволяет в параллельном режиме осуществлять как фонетическую разметку корпуса, так и обучение моделей фонем. Данный подход был использован в работе, а для его применения была создана текстовая аннотация записанных голосовых файлов.

Первым этапом создания корпуса звучащей речи является формирование текста для озвучивания. При создании фонетического анализатора в качестве искомой характеристики текста оправдан выбор показателя частотности употребления входящих в него слов. Для этого были проанализированы 5061 текст татарской литературы и публицистики. Основные характеристики использованных произведений представлены в Таблице 1.

Таблица 1. Характеристика исходных текстов на татарском языке.

Параметр	Значение
Количество текстов	5061
Общий объем текстов	337 МБ
Общее количество слов	25 584 505
Количество различных слов	1 418 909

На основе данных текстов была построена статистика частотности слов. Первые 10788 самых часто употребимых слова были выбраны для озвучивания в речевом корпусе. Запись звуковых фрагментов осуществлялась со следующими параметрами:

- формат файла: WAV PCM;
- частота дискретизации: 22 kHz;
- количество бит на отсчет: 16 бит.

Созданный корпус имеет параметры, представленные в Таблице 2.

Таблица 2. Основные характеристики звукового корпуса.

Параметр	Значение
Общее количество файлов	10788
Общая продолжительность записей	4:56:45
Количество файлов в обучающем корпусе	9631
Продолжительность обучающем корпусе	4:26:42
Количество файлов в тестовом корпусе	1157
Продолжительность тестовом корпусе	0:30:03

Акустические особенности татарского языка

Для дальнейшего анализа необходимо осуществить переход от текстового представления озвученных слов к их фонемной транскрипции. Для этого решаются следующие подзадачи:

- выделение значимых особенностей татарской речи;
- определение фонемного алфавита;
- построение правил транскрибирования, основанной на фонемном алфавите.

В качестве основных базовых элементов языка, отличающихся в акустическом плане, а также способных оказывать влияние на смысл слова, были выбраны фонемы, представленные в Таблице 3.

Таблица 3. Набор фонем татарского языка.

Фонема	Описание	Пример	Фонема	Описание	Пример
A	открытый а	арасында- A2RA1SYNDA	M	твердый м	моны - MONY
A1	умеренно огубленный а	татар - TA2TA1R	M1	мягкий м	һәм - HH1AAM1
A2	сильно огубленный а	да - DA2	N	твердый н	аның - A2NYNN
AA	ә	дә - D1AA	N1	мягкий н	мин - M1IN1
U	огубленный у	ул - UL	P	твердый п	алып - A2LYP
UU	нейтрального образования ү	күрү - K1UU1R1UU	P1	мягкий п	итеп - IT1EP1
UU1	переднего образования ү	сүтеп - S1UU1T1EP1	R	твердый р	бар - BA2R
O	о	тора - TORA2	R1	мягкий р	бер - B1ER1
OO	ө	өчен - OOTch1EN1	S	твердый с	соң - SONN
I	и	иде - ID1E	S1	мягкий с	үс - UU1S1
Y	ы	аның - A2NYNN	T	твердый т	тора - TORA2
E	е	бер - B1ER1	T1	мягкий т	бит - B1IT1
E1	русское э	кеше - K1ESS1E1	F	твердый ф	туфан - TUFA2N
B	твердый б	бу - BU	F1	мягкий ф	фикер - F1IK1ER1
B1	мягкий б	бер - B1ER1	X	твердый х	халык - XA2LYK
W	губно-губной в	авыл - A2WYL	X1	мягкий х	хәзер - X1AAZ1ER1
V	в заимствованных словах	трамвай - TRA2MVA1J	HH	твердый h	һаман - HHA2MA1N

Фонема	Описание	Пример	Фонема	Описание	Пример
G	твердый г	гына - GYNA2	HH1	мягкий h	һәм - HH1AAM1
G1	мягкий г	генә - G1EN1AA	Ts	твердый ц	немец - N1EM1E1Ts
ZZ	ж	жанр - ZZA2NR	Tch	твердый ч	чыгып - TchYGYP
J	й	шулай - SSULA2J	Tch1	мягкий ч	өчен - OOTch1EN1
C	твердый җ	җавап - CA2WA1P	SS	твердый ш	шул - SSUL
C1	мягкий җ	иҗат - IC1A2T1	SS1	мягкий ш	кеше - K1ESS1E1
Z	твердый з	зур - ZUR	Tsh	щ	училище - UTchILITshE
Z1	мягкий з	үзе - UU1Z1E	NN	ң	аның - A2NYNN
K	твердый к	юк - JUK	D	твердый д	да - DA2
K1	мягкий к	бик - B1IK1	D1	мягкий д	дә - D1AA
L	твердый л	ул - UL	Sil	пауза	
L1	мягкий л	әле - AAL1E			

На основе определенного инвентаря фонем были выявлены акустические закономерности в татарском языке, приведём некоторые из них:

- аккомодация (в слове, в зависимости от первой гласной, используются либо все твердые, либо все мягкие согласные), например, бар – BA2R, бер – B1ER1;
- уменьшение огубленности фонемы А от начала к концу слова, например, балалар – BA2LA1LAR;
- замена некоторых звонких согласных, идущих рядом с другим глухим согласным, на свои глухие пары, например, тозсыз – TOSSYS;
- замена звонких согласных в конце слова на свои глухие пары, например, тоз – TOS;
- представление буквы Я в качестве пары J (й) и AA (ә) в случае, если перед ней идет буква И, например, иясе – IJAAS1E.

Для создания автоматической системы транскрибирования было разработано автоматизированное рабочее место фонетиста, которое предоставляет возможность создания формализованной записи правил. Форма создания и редактирования правил транскрибирования представлена на Рис. 2.

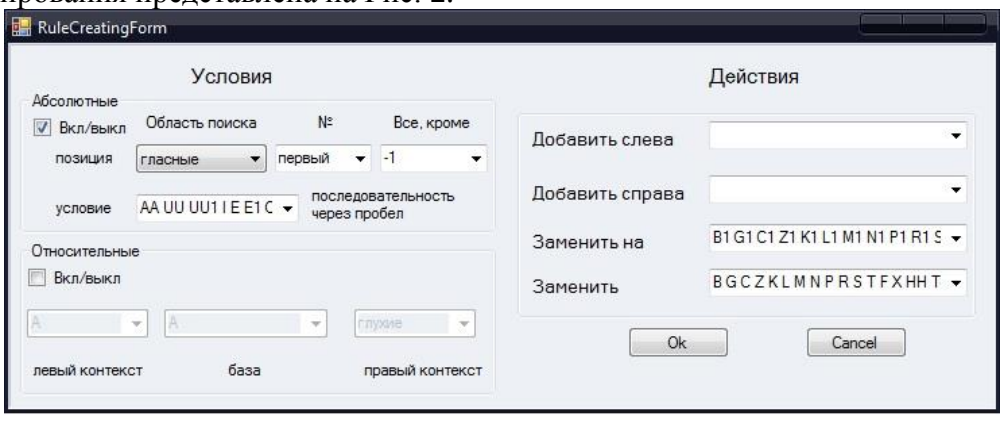


Рисунок 2. Форма создания правил транскрибирования.

Правила могут быть двух различных типов: абсолютные и относительные. Абсолютные правила оперируют конкретным расположением той или иной фонемы в слове и позволяют заменять их другими. Примером данного типа правил может служить изображенное на Рис. 2 правило аккомодации, в котором в зависимости от того, является ли первая главная в слове гласной переднего ряда, согласные заменяются на свою мягкую или твердую пару.

Вторым типом правил служат относительные правила, которые позволяют обрабатывать различные контексты следования тех или иных фонем, например, сочетание фонем Z - S заменяется на сочетание S – S, как, например, в слове тозсыз (T-O-S-S-I-Z). Общее количество созданных правил равняется 37.

Акустические модели фонем

Созданные на подготовительном этапе обучающий корпус речи и система транскрибирования позволяют реализовать алгоритм обучения акустических моделей фонем. Данный алгоритм носит название forced alignment и не требует наличия вручную фонетически аннотированного корпуса. Для реализации алгоритма будет использован инструмент HTK Toolkit, первоначально созданный в университете Кэмбриджа, а в настоящее время принадлежащий компании Microsoft.

Каждая фонема была смоделирована скрытой Марковской моделью, состоящей из трех состояний, с ограничениями на переход на более ранние состояния. Каждое из трех состояний моделировалось, в свою очередь, смесью Гауссовских распределений. Структура модели фонемы представлена на Рис. 3.

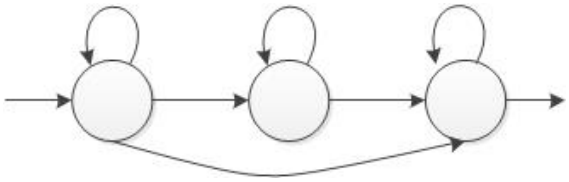


Рисунок 3. Структура модели фонемы.

Было проведено две серии экспериментов. В первой серии изучалась сходимость процесса обучения на корпусе, совпадающем с корпусом обучения. Вторая серия экспериментов проводилась на тестовом корпусе, не участвовавшем в обучении моделей. В обоих экспериментах количество Гауссовских распределений в смесях постепенно наращивалось, после каждого увеличения происходило два цикла переобучения всех моделей. Зависимость качества распознавания от количества итераций, т.е. количества распределений в Гауссовских смесях, представлена на Рис. 4 и 5. Качество распознавания анализировалось по двум критериям: Corr и Acc, которые вычисляются по следующим формулам:

$$Corr = \frac{N - D - S}{N} * 100\%, \text{ где}$$

N – число фонем,

D – число пропущенных при распознавании фонем,

S – число неправильно распознанных фонем.

$$Acc = \frac{N - D - S - I}{N} * 100\%, \text{ где}$$

I – число "лишних" фонем.

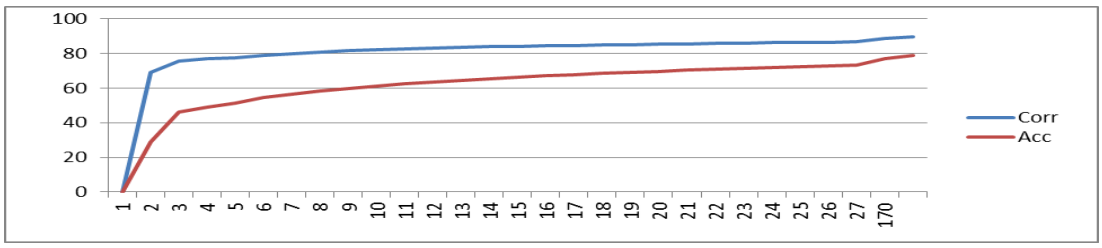


Рисунок 4. Зависимость качества распознавания от количества итераций на обучающем корпусе.

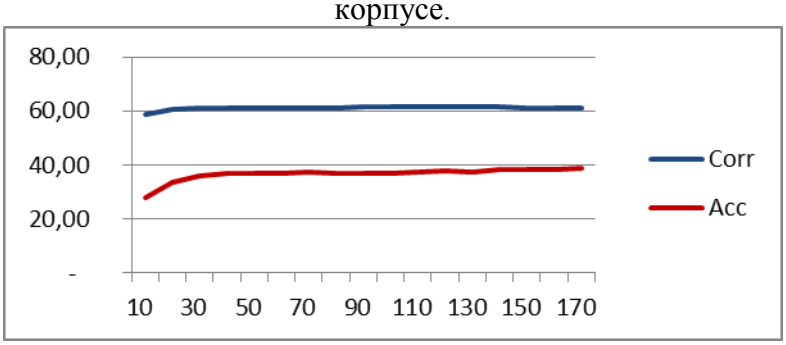


Рисунок 5. Зависимость качества распознавания от количества итераций на тестовом корпусе.

Анализ полученных результатов показывает, что рост числа распределений в Гауссовских смесях и увеличение числа циклов обучения моделей с определенного момента не оказывает существенного влияния на качество распознавания фонем на тестовом корпусе. Это связано с тем, что для качественного обучения большего числа распределений необходим всё больший объем исходной обучающей информации.

Таким образом, при построении системы фонетического распознавателя было решено выбрать модели фонем, полученные на 40 итерации. В этих моделях число распределений в Гауссовских смесях равняется 29.

Созданный программный модуль предоставляет возможности записи речевого фрагмента с помощью микрофона, а также загрузки необходимого звукового файла. Общий вид формы представлен на Рис. 6.

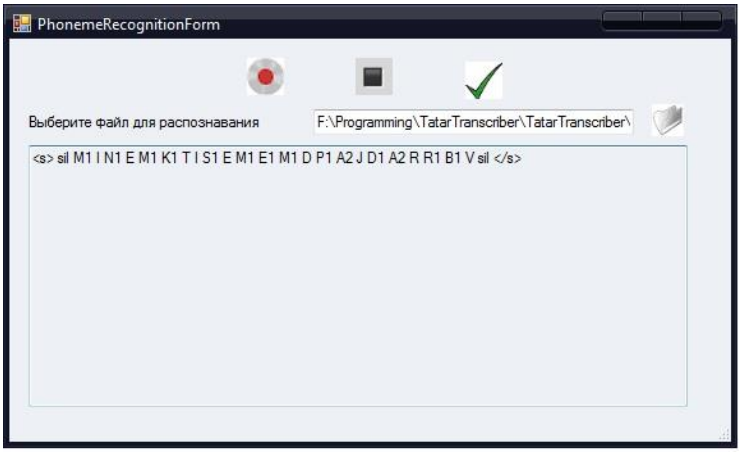


Рисунок 6. Форма распознавания фонем татарского языка.

Заключение

Построение и реализация алгоритма автоматического транскрибирования татарских текстов, а также создание аннотированного корпуса звучащей татарской речи позволило

реализовать программный модуль автоматического распознавания фонем татарского языка. Применяемый при создании моделей фонем аппарат скрытых Марковских моделей показал хорошее качество обучения. Проведенные серии экспериментов позволили выявить оптимальные характеристики моделей для их дальнейшего использования в системе. Дальнейшее улучшение качества работы распознавателя возможно засчет увеличения размера обучающего корпуса татарской речи, что позволит обучить модели на основе Гауссовских смесей большей размерности.

Литература

1. Lopes, C, Perdigao, F. Phone recognition on TIMIT database. Speech technologies. 285-302 (2011).
2. Young, S: The HTK book (for HTK version 3.4). (2009).
3. Gales, M, Young, S. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing. C. 113 (2007).

ZHANDOS YESSENBAYEV[1,2], MUSLIMA KARABALAYEVA[2], FIRUZA SHAMAYEVA [3]

[1] Nazarbayev University Research and Innovation System, Astana, Kazakhstan
[2] L.N. Gumilev Eurasian National University, Astana, Kazakhstan
[3] The Korkyt-Ata Kyzylorda State University, Kyzylorda, Kazakhstan

A BASELINE LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION FOR KAZAKH

Abstract

In this paper we present our experiments on large vocabulary continuous speech recognition task for Kazakh. This includes an acoustic database collection, acoustic and language modeling experiments. The overall performance of a system is 6.9% WER.

Keywords: speech recognition, acoustic database, acoustic and language modeling for Kazakh

Introduction

Speech recognition is a process of automatic conversion of human speech into corresponding text. Modern automatic speech recognition systems (ASR) advanced significantly from simple speaker-dependent word recognition to speaker-independent large vocabulary continuous speech recognition for broadcast news and telephone conversation transcriptions. Despite of widespread use of such systems in daily life, most of them are concerned with the languages like English, German, Japan, Russian, etc. As for Kazakh language, it is still underrepresented in speech recognition research. Thus, the primary goal of this work is to build a baseline large vocabulary continuous speech recognition system.

Fig. 1 outlines a standard architecture of a modern ASR system, which includes feature extraction and pre-processing, acoustic and language modeling, system combination and decoding. First step to build such a system for Kazakh would be collecting enough audio data, and creating the acoustic and language models. This is exactly the way we approach the problem.

This paper presents an acoustic database of Kazakh speech in Section 2, the experiments and conclusions are given in Sections 3 and 4, respectively.

Figure 1. The architecture of a standard ASR system.

Acoustic Database

Most of the modern speech processing systems requires large amount of audio and text data for training the acoustic and language models. Depending on the type of an application data needed varies from high quality microphone read speech (WSJ0 [1]) to conversational telephone speech (Switchboard [2] or CALLHOME [3]), from continuous speech (TIMIT [4]) to connected (TIDIGITS [5]) and isolated words (PhoneBook [6]). In our current work, we collected a corpus of 28 hours high quality microphone read Kazakh speech of 169 native speakers for the large vocabulary continuous speech recognition tasks. The acoustic database is initiated as a part of the Kazakh Language Corpus compiled in [7].

Text materials

The text materials to be uttered were carefully selected from the primary text corpus and divided into two parts: short sentences and stories.

The "sentences" part has more than 12K different sentences randomly and equally extracted from the five stylistic genres mentioned above. The sentences are chosen so that they have more than 120K words contained in the list of the most frequent words covering the 95% of all the texts in the primary corpus. Additionally, the sentences were grouped according to their length in words. Thus, we have 10 groups of sentences having the lengths from 6 to 15 words in each.

The "stories" part contains the short online news extracted from massmedia section of the primary text corpus. Each story has not more than 300 words.

All the text materials were subdivided into numbered small and nonintersecting sets to be uttered by the speakers. A standard set for one speaker has exactly 75 sentences (by 10 sentences from five shorter groups and by 5 sentences from five longer groups) and 1 story.

Speakers

The speakers that took part in the recordings are volunteers recruited by advertisements in the local newspapers and personal referral. The main criteria of speaker selection were a region where he/she learned Kazakh or spent most of his/her life, age, gender and the ability to read Kazakh.

The first criterion helped to capture variability present in speech due to the speakers' settlement both local and external. Totally there are 15 region groups: 14 official regions ("oblast") of Kazakhstan and one group for those who lived outside of the country.

The speakers are divided into four age groups not including children and school students:

- I group – 18-27 years;
- II group – 28-37 years;
- III group – 38-47 years;
- IV group – 48 years and above.

We did not strictly balance the speakers by their gender due to the difficulties in finding the volunteers but still tried to keep the number of speakers of one gender per profile not more than 3. The female and male distributions are 57% and 43%, respectively.

The other important criterion was the ability to read Kazakh since not all the interviewees could read in Kazakh sufficiently fluent, what is a common issue in a bilingual country such as Kazakhstan. Additionally we kept the records of the speakers' education whether they graduated last from school, college or university.

Totally, we recorded 169 speakers. The following Table 1 presents the distribution of the speakers across the regions, gender and age groups. The blank spots show the speaker profiles that we could not recruit. Mostly, these correspond to the distant regions and elder male groups.

Table 1. The distribution of the speakers.

Age group	I		II		III		IV		
Region	F1	M1	F2	M2	F3	M3	F4	M4	Sum
1	3	3	2	1	2	1	2	1	15
2	2	3	2	1			2	1	11
3	1	1	2	3	2	1	1		11
4	3	2		1		1			7
5	2	2	2	1	2	2	2	1	14
6	2	2	2	2	2		1	2	13
7	2	2	1	2	2		2	1	12
8	2	1	1	2	1	1	2	1	11
9	3	2	2	1	3	1	1	1	14
10	1	1	2	2	1	1	2	1	11
11	2	1	2	1	1		2		9
12	2	2	2		2	1	2	1	12
13	2	2	2	1	1	1	1	1	11
14	2	1	1	1	1	2	1	2	11
15	1	3		1	2				7
Total	**30**	**28**	**23**	**20**	**22**	**12**	**21**	**13**	**169**
	34%		25%		20%		20%		

Recording setup

The actual recording sessions took place in a sound-proof studio of the university with the assistance of a sound operator. Before the recordings, the speakers were instructed, documented and given some time to prepare as well as asked to fill in the copyright transfer form for the audio data with their voice. They were not constrained on the manner, speed or time except for the correctness of reading. The average time for a recording session per speaker was about 40-45 minutes, though there were cases that last up to 2 hours.

Audio data were captured using the professional vocal microphone Neumann TLM 49 and digitized by LEXICON I-ONIX U82S sound card. The format of the recorded audio files is 44.1 kHz 16-bit PCM-encoded mono WAVE file format. All the recorded audio files were manually post-processed to have each utterance (sentences and stories) in a separate file and in the corresponding directories. The size of the speech corpus is about 8.5 GB on disk. The total duration of the audio files is about 28 hours with 23 hours of “sentences” and 5 hours of “stories” parts, respectively.

Transcripts and annotation

- Each audio file is provided with its corresponding orthographic transcription and TIMIT-style word-level segmentation as well as morpho-syntactic annotation files. All the data processing to obtain these files were performed manually by the trained linguists.

The transcriptions files contain the exact orthographic transcription of the utterances, which may differ from the original text. For example, the numbers, abbreviation, foreign words and dates are expanded depending on how they were uttered by the speakers. In addition, the transcription of the stories have the sentence boundaries labeled with <s> and </s>.

The segmentation was performed using WaveSurfer [8], an open-source tool for sound visualization and manipulation, which supports TIMIT word-level transcription format. Although it supports Unicode, it does not support Kazakh symbols well. Therefore, we used an ASCII version of the Kazakh letters. Also, we used # symbol for the pauses and silence, and ^ symbol for other non-speech events.

The morpho-syntactic annotation is includes part-of-speech and morpheme segmentation of each word as well as the information on syntax for each sentence.

Experiments

Acoustic Modeling

An acoustic model was trained using CMU Sphinxtrain-1.0.8 [9]. The front-end module was set to output default parameters such as 13 mel-frequency cepstral coefficients with their first and second derivatives. Additionally, speaker adaptation techniques such as cepstral mean normalization [10], LDA [11] and MLLT [12] are performed on feature vectors. We used a context-dependent tied-state continuous Hidden Markov Model with 8 Gaussian mixtures per state [13].

The dictionary is compiled from the transcriptions and contains about 30000 words with their spellings as a phonetic transcription. It should be noted that there is still no consensus regarding the

Kazakh phonetic alphabet among the linguists [14]. Therefore, since the orthographic transcription of Kazakh roughly corresponds to a broad phonetic transcription, for the phoneme set a reduced form of the Kazakh alphabet is used, i.e. it includes those letters used in writing of Kazakh words. Also, for some letters there are variations in pronunciation depending on letter's position or context in a word. Thus, letters, Е, О and Ө are pronounced as diphthongs in the beginning of a word. Letters Ю, Я are generally diphthongs except when used in the contexts CV and CVC, in such cases they obey vowel harmony and pronounced as their soft counterparts. Additionally, there is a SIL phone for silence.

Language Modeling

As for the language model, here we used our text materials to create a standard tri-grams based model with Good-Turing smoothing [15] compiled into ARPA format by CMU-Cambridge Language Model Toolkit 0.7 [16]. The format of language model file is as follows:

```
\data\
ngram 1=nr          # number of 1-grams
ngram 2=nr          # number of 2-grams
ngram 3=nr          # number of 3-grams

\1-grams:
p_1     wd_1 bo_wt_1
\2-grams:
p_2     wd_1 wd_2 bo_wt_2
\3-grams:
p_3     wd_1 wd_2 wd_3
\end\
```

where *ngram k* – is the number of the corresponding n-grams, *p_k* - the logarithm (base 10) of conditional probability *p* of an n-gram, *wd_k* – a word in n-gram, and *bo_wt_k* - the logarithm (base 10) of the backoff weight for the n-gram.

For our experiments we have totally over 12500 sentences, which produce 29586 unigrams, 100354 bi-grams and 120755 tri-grams.

Recognition Results

All the audio data was separated into training and test sets. The test set is balanced based on gender and includes one representative from each region. The quantitative information about both sets is given in Table 2. The overall performance of recognition on test data is 6.9% WER.

Table 2. Distribution of data in training and test sets.

	Train set	Test set
# of speakers	153	16
# of audio files	11367	1176

Conclusions and Future Work

In the current work we have conducted the experiments on large vocabulary continuous speech recognition task for Kazakh. First we build the first acoustic database of Kazakh speech, which is balanced with respect to gender, region and age group. Next, we build the acoustic and language models using CMU Sphinx toolkits. Finally, we evaluate our system on test data obtaining a word error rate of 6.9%.

While we build a state-of-the-art speech recognition system, it is assumed to be a baseline for our future work on speech recognition research. Thus, our next step will be to improve WER by exploiting class-based language models with morphological cues. This kind of approach seems more effective for inflectional languages such as Kazakh, Turkish and Russian.

Acknowledgements

The work is supported by the Ministry of Education and Science of the Republic of Kazakhstan.

References

1. John Garofalo, et al., “CSR-I (WSJ0) Complete,” Linguistic Data Consortium, Philadelphia, 2007.
2. Godfrey J. J., “Holliman E. Switchboard-1 Release 2,” Linguistic Data Consortium, Philadelphia, 1997.
3. Canavan A., Zipperlen G., “CALLHOME Japanese Speech,” Linguistic Data Consortium, Philadelphia, 1996.
4. John S. Garofolo, et al., “TIMIT – Acoustic-Phonetic Continuous Speech Corpus,” Linguistic Data Consortium, Philadelphia, 1993.
5. Leonard R. G., Doddington G., “TIDIGITS.” Linguistic Data Consortium, Philadelphia, 1993.
6. Pitrelli, J.; Fong, C.; Wong, S.H.; Spitz, J.R.; Leung, H.C., "PhoneBook: a phonetically-rich isolated-word telephone-speech database," Int. Conf. on Acoustics, Speech, and Signal Processing, 1995, vol.1, pp.101-104.
7. Z. Yessenbayev, O. Makhambetov, and M. Karabalayeva, “Kazakh Text Corpus: Description, Tools and Statistics,” Int. scientific-theoretical conference “Modern Kazakh Linguistics: Actual Problems of Applied Linguistics”, 2012, pp. 61-65.
8. Wavesurfer. URL: http://www.speech.kth.se/wavesurfer/
9. CMU Sphinxtrain. Online: http://sourceforge.net/projects/cmusphinx/files/sphinxtrain/1.0.
10. Liu, F.-h., Stern, R.M., Huang, X., Acero, R., “Efficient Cepstral Normalization for Robust Speech Recognition,” In Proceedings of the workshop on Human Language Technology, 1993, pp. 69–74.
11. Haeb-Umbach, R., Ney, H., “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” IEEE Int. Conf. on Acoustics, Speech, and Signal Process, 1992, vol. 1, pp. 13–16.
12. Gopinath, R.A., "Maximum likelihood modeling with Gaussian distributions for classification," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1998, vol.2, pp. 661-664.
13. S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modeling,” In Proceedings of the workshop on Human Language Technology (HLT '94). Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 307-312.
14. Torekhanov S. Sharipbayev A. On the current state of the Kazakh phonetics, L.N. Gumilyov Eurasian National University Bulletin, Astana, 2001, Vol. 3-4, pp. 6-9.
15. Good, I.J., "The population frequencies of species and the estimation of population parameters", Biometrika, 1953, 40 (3–4): 237–264.
16. Cmuclmtk. Online: http://sourceforge.net/projects/cmusphinx/files/cmuclmtk/0.7/

А.К.БУРИБАЕВА

Евразийский национальный университет им. Л.Н. Гумилева, НИИ «Искусственный интеллект», Астана, Казахстан

РАСПОЗНАВАНИЕ КАЗАХСКИХ СЛОВ НА ОСНОВЕ ДИФОННОЙ БАЗЫ

Абстракт: В работе предложен метод распознавания слов на основе дифонной базы и принципы создания дифонной базы казахского языка. Система распознает не отдельные дифоны, а целые слова по эталонам, синтезированным из дифонов. Автоматическое

генерирование эталонов слов из дифонов позволит сделать шаг в сторону сверхбольших словарей.

Ключевые слова: распознавание слов, дифонная база, алгоритм DTW

Введение

Автоматическое распознавание устной речи естественного языка является одним из актуальных направлений развития искусственного интеллекта. Результаты в этом направлении позволят решить проблему создания средств эффективного речевого взаимодействия человека с компьютером. Речевой ввод обладает рядом преимуществ, таких, как естественность, оперативность, смысловая точность ввода, освобождение рук и зрения пользователя, возможность управления и обработки в экстремальных условиях.

Исследованием проблемы распознавания речи уже более 50 лет занимаются специалисты нескольких научных областей. Методы и алгоритмы, которые используются, могут быть разделены на четыре больших класса:

- Методы дискриминантного анализа, основанные на Байесовской дискриминации [1];
- Скрытые модели Маркова [2];
- Искусственные нейронные сети [3];
- Динамическое программирование – временные динамические алгоритмы (DTW) [4];

Следует отметить ряд преимуществ, к которым стремятся при разработке систем распознавания речи:

1. Непрерывная речь — возможность, позволяющая пользователям говорить естественно (непрерывно), не делая паузы между словами (дискретный ввод речи).
2. Большие словари — способность обрабатывать большое количество слов как общей, так и специальной категории из технических и предметных областей знаний с целью увеличения мощности и эффективности систем распознавания голоса.
3. Независимость от диктора — способность системы распознавать слова без персональной настройки компьютера путем повторения одного и того же речевого сигнала.

Наиболее часто и успешно при распознавании слитной речи используется скрытая марковская модель (HMM) [5, 6] либо ИНС [6, 7]. Для распознования выбираются различные базовые единицы: фонемы, аллофоны, дифоны и трифоны и т.д. Для распознавания отдельных слов все же эффективны временные динамические алгоритмы (DTW) [8].

В связи с тем, что распознавание целых слов надежнее, мы выбрали технологию распознавания слов на основе наговоренной дифонной базы [9]. Суть в том, что система не распознает дифоны по отдельности, а сначала синтезирует из них эталоны слов, и затем распознает целые слова по алгоритму DTW.

Преимущество системы в том, что для добавления нового слова нет необходимости обучать систему наговаривая слово, а достаточно ввести слово в текстовом виде. Автоматическое генерирование эталонов слов из дифонов позволит сделать шаг в сторону сверхбольших словарей, а дикторонезависимости системы можно добиться усреднением эталонов.

1. Создание дифонной базы казахского языка

Дифон — звуковая единица, имеющая протяженность от середины одного звука до середины последующего. Дифонная модель основана на предположении, что существуют стационарные участки звуков, и они не зависят от влияния соседних звуков (коартикуляции): в середине этого стационарного участка и проводится граница.

Акустическая база системы распознавания речи включает три типа дифонов – начальный, серединный и конечный.

Начальные и конечные дифоны, как правило, представляют половинки первой и последней фонем слова с включением переходных участков от пробела к фонеме, а также от фонемы к пробелу соответственно. Они определеяются согласно позиционным правилам казахских звуков:

- звуки **а, ә, е, ө, ұ, ү, ы, і** встречаются во всех позициях;
- гласный звук **о** встречается только в началном слоге;
- звуки **л, р, й, ң, у (w)** не встречаются в начале слова;
- согласные **б, д, ғ, г** не встречаются в конце слова.

Однако звуки «о», «ө» и «е» являются дифтонгами, и согласно казахской орфоэпии перед звуками «о», «ө» идущими в начале слова, есть маленькая согласная вставка «у», а перед звуком «е» - согласная вставка «й». Исходя из этого, звуки «о», «ө» и «е» были удалены из списка начальных полудифонов и вместо них введены полудифоны «у» и «й».

Учитывая все вышеуказанные правила, был составлен список начальных и конечных дифонов (Таблица 1):

Таблица 1. Начальные и конечные дифоны казахского языка

Начальные		**Конечные**	
а0	м0	а2	ө2
ә0	н0	ә2	п2
б0	п0	е2	р2
г0	с0	ж2	с2
ғ0	т0	з2	т2
д0	у0	й2	у2
ж0	ұ0	к2	ұ2
з0	ү0	қ2	ү2
й0	ш0	л2	ш2
к0	ы0	м2	ы2
қ0	і0	н2	і2
		ң2	

Для составления матриц серединных дифонов сначала был автоматически сгенерирован список всевозможных сочетаний звуков казахского языка. Затем были удалены из списка сочетания, противоречащие следующим казахским позиционным правилам:

- звуки **а, ә, о, ө, ұ, ү** сочетаются со всеми согласными;
- звуки **е, ы, і** не сочетаются с согласным **у (w).**
- в казахском языке не встречаются подряд идущих 2 гласных;
- глухие и звонкие согласные не сочетаются;
- **согласный у** не встречается после согласных звуков;

Некоторые сочетания были удалены в связи с тем, что они по статистике не встречаются вообще [10].

В итоге мы получили около 500 звукосочетаний казахского языка.

Но для качественного распознавания их не достаточно, так как казахский язык является сингармоническим языком.

Рассмотрим дифоны из тех звукосочетаний, в которых один из звуков гласный (таблицы 3, 4). Их количество остается без изменений, так как гласные определяют огубленность/неогубленность и мягкость/твердость согласного.

Таблица 2. Дифоны с гласно-согласным звукосочетанием

	б	**г**	**ғ**	**д**	**ж**	**з**	**й**	**к**	**қ**	**л**	**м**	**н**	**ң**	**п**	**р**	**с**	**т**	**у**	**ш**
	аб		ағ	ад	аж	аз	ай		ақ	ал	ам	ан	аң	ап	ар	ас	ат	ау	аш
	әб'	әг'		әд'	әж'	әз'	әй'	әк'		әл'	әм'	ән'	әң'	әп'	әр'	әс'	әт'	әу'	әш'
	еб'	ег'		ед'	еж'	ез'	ей'	ек'		ел'	ем'	ен'	ең'	еп'	ер'	ес'	ет'	еу'	еш'
	об о		оғ о	од о	ож о	оз о	ой о		оқ о	ол о	ом о	он о	оң о	оп о	ор о	ос о	от о	оу о	ош о

	өб ˈо	өг' о		өд ˈо	өж ˈо	өз' о	өй ˈо	өк ˈо		өл ˈо	өм ˈо	өн ˈо	өң ˈо	өп ˈо	өр ˈо	өс ˈо	өт ˈо	өу ˈо	өш ˈо
	ұб о		ұғ о	ұд о	ұж о	ұз о	ұй о		ұқ о	ұл о	ұм о	ұн о	ұң о	ұп о	ұр о	ұс о	ұт о	ұу о	ұш о
	үб ˈо	үг' о		үд ˈо	үж ˈо	үз' о	үй ˈо	үк ˈо		үл ˈо	үм ˈо	үн ˈо	үң ˈо	үп ˈо	үр ˈо	үс ˈо	үт ˈо	үу ˈо	үш ˈо
	ы б		ы ғ	ы д	ы ж	ыз	ы й		ы қ	ы л	ы м	ы н	ы ң	ы п	ы р	ыс	ыт		ы ш
	і б'	і г'		і д'	і ж'	і з'	і й'	і к'		і л'	і м'	і н'	і ң'	і п'	і р'	і с'	і т'		і ш'

Таблица 3. Дифоны с согласно-гласным звукосочетанием

	а	**ә**	**е**	**о**	**ө**	**ұ**	**ү**	**ы**	**і**
б	ба	б'ә	б'е	б ˚о	б'ˈөө	б ˚ұ	б'ˈ ү	бы	б'і
г		г'ә	г'е		г'ˈөө		г'ˈ ү		г'і
ғ	ға			ғ ˚о		ғ ˚ұ		ғы	
д	да	д'ә	д'е	д ˚о	д'ˈөө	д ˚ұ	д'ˈ ү	ды	д'і
ж	жа	ж'ә	ж'е	ж ˚о	ж'ˈөө	ж ˚ұ	ж'ˈ ү	жы	ж'і
з	за	з'ә	з'е	з ˚о	з'ˈөө	з ˚ұ	з'ˈ ү	зы	з'і
й	йа	й'ә	й'е		й'ˈөө	й ˚ұ	й'ˈ ү	йы	й'і
к		к'ә	к'е		к'ˈөө		к'ˈ ү		к'і
қ	қа			қ ˚о		қ ˚ұ		қы	
л	ла	л'ә	л'е		л'ˈөө	л ˚ұ	л'ˈ ү	лы	л'і
м	ма	м'ә	м'е	м ˚о	м'ˈөө	м ˚ұ	м'ˈ ү	мы	м'і
н	на	н'ә	н'е	н ˚о	н'ˈөө	н ˚ұ	н'ˈ ү	ны	н'і
ң	ңа	ң'ә	ң'е		ң'ˈөө	ң ˚ұ	ң'ˈ ү	ңы	ң'і
п	па	п'ә	п'е		п'ˈөө	п ˚ұ	п'ˈ ү	пы	п'і
р	ра	р'ә	р'е		р'ˈөө	р ˚ұ	р'ˈ ү	ры	р'і
с	са	с'ә	с'е	с ˚о	с'ˈөө	с ˚ұ	с'ˈ ү	сы	с'і
т	та	т'ә	т'е	т ˚о	т'ˈөө	т ˚ұ	т'ˈ ү	ты	т'і
у	уа	у'ә		у ˚о	у'ˈөө	у ˚ұ	ұ'ˈ ү		
ш	ша	ш'ә	ш'е	ш ˚о	ш'ˈөө	ш ˚ұ	ш'ˈү	шы	ш'і

Возникает вопрос, «а как быть с дифонами с согласно-согласным звукосочетанием?» Можно ли, например, один и тот же дифон «ст» применить для синтеза эталонов слов «астау» и «үстөу»? Или же применить каждому свой: твердый, негубной «ст» для «астау», а мягкий, губной «ст» для «үстөу»? И сколько вариантов «ст» может быть вообще?

Мы решили разбить казахские дифоны, состоящие из исключительно согласных звуков на следующие группы по сингармоническим тембрам:

1) Твердый негубной/ Твердый негубной;
2) Твердый негубной /Мягкий негубной;

3) Твердый негубной/ Твердый губной;
4) Твердый негубной/ Мягкий губной;
5) Мягкий негубной/ Твердый негубной;
6) Мягкий негубной/ Мягкий негубной;
7) Мягкий негубной/ Твердый губной;
8) Мягкий негубной/ Мягкий губной;
9) Твердый губной / Твердый негубной;
10) Твердый губной / Мягкий негубной;
11) Твердый губной / Твердый губной;
12) Твердый губной / Мягкий губной;
13) Мягкий губной / Твердый негубной;
14) Мягкий губной / Мягкий негубной;
15) Мягкий губной / Твердый губной;
16) Мягкий губной / Мягкий губной.

Таким образом, один и тот же дифон может иметь 16 вариантов (таблица 5).

Таблица 4. Дифоны с согласно-согласным звукосочетанием с учетом всех признаков

	тн/тн	**тн/мн**	**тн/тг**	**тн/мг**	**мн/тн**	**мн/мн**	**мн/тг**	**мн/мг**	**тг/тн**	**тг/мн**	**тг/тг**	**тг/мг**	**мг/тн**	**мг/мн**	**мг/тг**	**мг/мг**
б д	бд	бд'	бд^{o}	бд'^{o}	б'д	б'д'	б'д^{o}	б'д'^{o}	б^{o}д	б^{o}д'	б^{o}д$_{o}$	б^{o}д'$_{o}$	б'^{o}д	б'^{o}д'	б'^{o}д$_{o}$	б'^{o}д'$_{o}$
б ж	бж	бж'	бж^{o}	бж'^{o}	б'ж	б'ж'	б'ж^{o}	б'ж'$_{o}$	б^{o}ж	б^{o}ж'	б^{o}ж$_{o}$	б^{o}ж'$_{o}$	б'^{o}ж	б'^{o}ж'	б'^{o}ж^{o}	б'^{o}ж'$_{o}$
б з	бз	бз'	бз^{o}	бз'^{o}	б'з	б'з'	б'з^{o}	б'з'^{o}	б^{o}з	б^{o}з'	б^{o}з$_{o}$	б^{o}з'$_{o}$	б'^{o}з	б'^{o}з'	б'^{o}з$_{o}$	б'^{o}з'^{o}
г б	гб	гб'	гб^{o}	гб'^{o}	г'б	г'б'	г'б^{o}	г'б'^{o}	г^{o}б	г^{o}б'	г^{o}б$_{o}$	г^{o}б'$_{o}$	г'^{o}б	г'^{o}б'	г'^{o}б$_{o}$	г'^{o}б'$_{o}$
г д	гд	гд'	гд^{o}	гд'^{o}	г'д	г'д'	г'д^{o}	г'д'^{o}	г^{o}д	г^{o}д'	г^{o}д$_{o}$	г^{o}д'$_{o}$	г'^{o}д	г'^{o}д'	г'^{o}д$_{o}$	г'^{o}д'$_{o}$
г ж	гж	гж'	гж^{o}	гж'^{o}	г'ж	г'ж'	г'ж^{o}	г'ж'^{o}	г^{o}ж	г^{o}ж'	г^{o}ж$_{o}$	г^{o}ж'$_{o}$	г'^{o}ж	г'^{o}ж'	г'^{o}ж$_{o}$	г'^{o}ж'$_{o}$
гз	гз	гз'	гз^{o}	гз'^{o}	г'з	г'з'	г'з^{o}	г'з'^{o}	г^{o}з	г^{o}з'	г^{o}з^{o}	г^{o}з'^{o}	г'^{o}з	г'^{o}з'	г'^{o}з^{o}	г'^{o}з'$_{o}$
ғ б	ғб	ғб'	ғб^{o}	ғб'^{o}	ғ'б	ғ'б'	ғ'б^{o}	ғ'б'^{o}	ғ^{o}б	ғ^{o}б'	ғ^{o}б$_{o}$	ғ^{o}б'$_{o}$	ғ'^{o}б	ғ'^{o}б'	ғ'^{o}б$_{o}$	ғ'^{o}б'$_{o}$
…																
ш т	шт	шт'	шт^{o}	шт'^{o}	ш't	ш't'	ш't^{o}	ш't'$_{o}$	ш^{o}т	ш^{o}т'	ш^{o}т$_{o}$	ш^{o}т'$_{o}$	ш'^{o}т	ш'^{o}т'	ш'^{o}т^{o}	ш'^{o}т'$_{o}$
ш ш	шш	шш'	шш$_{o}$	шш'$_{o}$	ш'ш	ш'ш'	ш'ш$_{o}$	ш'ш'$_{o}$	ш^{o}ш	ш^{o}ш'	ш^{o}ш^{o}	ш^{o}ш'^{o}	ш'^{o}ш	ш'^{o}ш'	ш'^{o}ш^{o}	ш'^{o}ш'^{o}

Но в ходе эксперимента выяснилось, что учитывание всех этих признаков при распознавании не обязательно. Для качественного распознавания достаточно учесть мягкость/твердость дифонов. В результате в базе для каждого дифона с согласно-согласным звукосочетанием оставили всего 4-х варианта.

Таблица 5. Дифоны с согласно-согласным звукосочетанием с учетом мягкости и твердости составляющих звуков

	тн /тн	**тн/ мн**	**мн /тн**	**мн/ мн**
б	бд	бд'	б'д	б'д'

д				
бж	бж	бж'	б'ж	б'ж'
бз	бз	бз'	б'з	б'з'
гб	гб	гб'	г'б	г'б'
гд	гд	гд'	г'д	г'д'
гж	гж	гж'	г'ж	г'ж'
гз	гз	гз'	г'з	г'з'
ғб	ғб	ғб'	ғ'б	ғ'б'
...				
шт	шт	шт'	ш'т	ш'т'
шш	ш ш	ш ш'	ш' ш	ш' ш'

В итоге, дифонная база казахского языка составила около 1000 дифонов.

2. Фонетический транскриптор

Для разработки фонетического транскриптора были исследованы орфоэпические правила казахского языка. Для удобства читателя в данном тексте правила разбиты на группы, которые занумерованы:

1) В казахском языке, если слово начинается на гласное «е», то при произношении перед ней слышится «й», если слово начинается на гласные «о», «ө», то при произношении перед ними образуется краткая вставка «у», например, «ет» – «йет», «он» – «уон», «өнер» – «уөнер».

2) Если слово начинается на согласные «р» или «л», то при произношении перед этими звуками слышится гласные «ы», «і», в зависимости от твердости или мягкости согласных, здесь «r», «l» означает мягкие аналоги «р» и «л», например, «рас» – «ырас», «рет» – «ірет», «лас» – «ылас», «лезде» – «ілезде».

3) При произношении заимствованного звука «ю» в составе слова слышится «йұу», «йұу», в зависимости от твердости или мягкости гласных в остальных слогах. Например: «қою» – «қойұу», «түю» - «түйұу»;

4) При произношении заимствованного звука «я» в составе слова слышится «йа», «йә», в зависимости от твердости или мягкости гласных в остальных слогах. Например: «аян» – «айан», «әлия» – «әлійә»;

5) При произношении заимствованного звука «и» в составе слова слышится «ый», «ій», в зависимости от твердости или мягкости гласных в остальных слогах. Например, «ине» – «ійне», «жина» – «жыйна». Если перед или после «и» идут согласные «қ», «ғ», то при произношении звука «и» всегда слышится «ый». Например, «қиын» – «қыйын», «қиғаш» – «қыйғаш».

6) При произношении дифтонга «у» в составе слова слышится «ұу», «үу», в зависимости от твердости или мягкости гласных в остальных слогах. Например, «туыс» – «тұуыс», «күту» – «күтүу».

7) Гласные звуки «ұ», «ү», «о», «ө» в начале или в первом слоге слова при произношении изменяют в следующих слогах гласные звуки «ы», «і» на гласные звуки «ұ», «ү»

соответственно. Например, «қолтық» – «қолтұқ», «құлын» – «құлұн», «күлкі» – «күлкү», «көлік» – «көлүк»;

8) Гласные звуки «ү», «ө» в начале или в первом слоге слова при произношении изменяют в следующих слогах гласный звук «е» на гласный «ө», например, «үлкен» – «үлкөн», «өнер» - «өнөр».

9) Гласные звуки «ә», «ү», «і» в начале или в первом слоге слова при произношении изменяют в следующих слогах гласный звук «а» на его аллофон «ә», например, «ләззат» – «ләззәт», «діндар» – «діндәр».

10) Если в слове звуки «с» и «ш», «с» и «ж» или «з» и «ш» встречаются подряд, вместо них произносится двойной звук «шш». Также вместо заимствованного звука «щ» произносится «шш». Например: «досжан» - «дошшан», «басшы - башшы», «сөзшең – сөшшөң», «көшсең»-«көшшөң», «ащы»-«ашшы».

11) Если в слове звуки «з» и «ж» встречаются подряд, то вместо них произносится двойной звук «жж», а а если звуки «з» и «с» встречаются подряд, то вместо них произносится двойной звук «сс». Например: «бозжорға» - «божжорға», «азсыну - ассынұу».

12) Если в слове после звука «н» встречается «б» или «п», то при произношении звук «н» заменяется на «м». Например: «мінбер – мімбер», «ойынпаз» – «ойымпаз».

13) Если в слове после звука «н» встречается «г», «ғ», «к» или «қ» то при произношении звук «н» заменяется на «ң». Например: «түнгі» – «түңгү», «қашанғы» - «қашаңғы», «зиянкес» – «зыйаңкес», «сәнқой» – «сәңқой».

14) При произношении в составе слова звукосочетний мл, ғн, ғл между двумя звуками образовавется краткая вставка гласных «ы», «і», в зависимости от твердости и мягкости слога соответственно. Например, «мемлекет» – «мемілекет», «бағлан» – «бағылан», «яғни» – «йағыный».

15) Не сочетаемые звуки, встречающиеся во многих сложных словах заменяется звуком по произношению, например, «шашбау» - «шашпау», «атбегі»-«атпегі», «атжалман» - «атшалман», «Көпбосын» - «Көппосұн», «түпдерек» - «түбдөрөк», «көпжиын» - «көбжыйын», «көпмүше – «көбмүшө», «түпнегіз» - «түбнегіз», «тасбауыр» - «таспаұуұр» и т.д.

Транскриптор реализован как программа, заменяющая одни символы другими в соответствии с правилами, содержащимися в управляющем файле. Правила написаны в сотвтетствии с вышеуказанными орфоэпическими правилами казахского языка по каждому пункту:

1) #е=йе, #о=уо, #ө=үө;

2) #л^а=ыл^а, #л^о=ыл^о, #л^ұ=ыл^ұ, #л^ә=іл^ә, #л^ү=іл^ү, #л^е=іл^е, #л^і=іл^і, р^а=ыр^а, #р^о=ыр^о, #р^ұ=ыр^ұ, #р^ә=ір^ә, #р^ү=ір^ү, #р^е=ір^е, #р^і=ір^і;

3) аю=айұу, ою=ойұу, ұю=ұйұу, ыю=ыйұу, үю=үйүу, ею=ейүу, қию=қыйұу, #тию#=тійүу, кию=кійүу, #сию#=сыйұу, #жию#=жыйұу, а^ию=а^ыйұу, о^ию=о^ұйұу, ұ^ию=ұ^ұйұу, ы^ию=ы^ыйұу, ә^ию=ә^ійүу, ө^ию=ө^үйүу, ү^ию=ү^үйүу, і^ию=і^ійүу, е^ию=е^ійүу;

4) ая=айа, оя=ойа, ұя=ұйа, ыя=ыйа, қия=қыйа, #сия=сыйа, #жия=жыйа, #мия=мыйа, #зия=зыйа, а^ия=а^ыйа, о^ия=о^ұйа, ұ^ия=ұ^ұйа, ы^ия=ы^ыйа, ә^ия=ә^ійә, ү^ия=ү^үйә, ия^а=ыйа^а;

5) #ми=мый, #жи=жый, а^и=а^ый, о^и=о^ый, ұ^и=ұ^ый, ы^и=ы^ый, ә^и=ә^ій, ө^и=ө^ій, ү^и=ү^ій, і^и=і^ій, е^и=е^ій, и^а=ый^а, и^о=ый^о, и^ұ=ый^ұ, и^ы=ый^ы, и^ә=ій^ә, и^ө=ій^ө, и^ү=ій^ү, и^і=ій^і, и^е=ій^е, қи=қый, ғи=ғый, иқ=ыйқ, иғ=ыйғ;

6) а^у=а^ұу, о^у=о^ұу, ұ^у=ұ^ұу, ы^у=ы^ұу, ә^у=ә^үу, ө^у=ө^үу, ү^у=ү^үу, і^у=і^үу, е^у=е^үу, у^а=ұу^а, у^о=ұу^о, у^ұ=ұу^ұ, у^ы=ұу^ы, у^ә=үу^ә, у^ө=үу^ө, у^ү=үу^ү, у^і=үу^і, у^е=үу^е;

7) о^ы=о^ұ, ұ^ы=ұ^ұ, ө^і=ө^ү, ү^і=ү^ү;

8) ө^е=ө^ө, ү^е=ү^ө;

9) і^а=і^ә, ү^а=ү^ә, ө^а=ө^ә;

10) сш=шш, сж=шш, зш=шш, шс=шш, щ=шш;

11) зж=жж, зс=сс;

12) нб=мб, нп=мп;

13) нг=ңг, нғ=ңғ, нк=ңк нқ=ңқ;

14) мл = мiл, ғн=ғын, ғл=ғыл;

15) шб=шп, тб=тп, тж=тш, пб=пп, пд=бд, пж=бж, пм=бм, пн=бн, сб=сп, сд=ст, қб=қп, қг=кг, қд=ғд, қж=ғж, қз=ғз, қм=ғым, қн=ғын, кб=кп, кг=кк, кд=кт, кж=гж, кз=гз, км=гм, кн=гн, зк=зг, зп=зб, зт=ст, қл=ғыл.

Каждое правило подстановки состоит из двух частей, разделенных между собой знаком «=». Слева от этого знака стоят исходные символы буквенной записи слова, справа – символы которыми они должны замениться в транскрипции.

Для транскрибирования заданного слова последовательно ищется в нем вхождение левой части очередного правила, и если таковое обнаруживается, то вместе неё подставляется правая часть этого правила.

В качестве транскрипционных знаков для гласных звуков использованы в основном соответствующие казахские буквы. Твердые казахские согласные транскрибируются также казахскими буквами, а соответствующие мягкие согласные аналогичными латинскими буквами.

Знак «#» означает начало или конец слова в зависимости от местоположения: если «#» стоит перед символами, то это начало слова; если «#» стоит после символов, то это конец.

Знак «^» означает любые символы в любом количестве между двумя звуками.

Каждое правило подстановки состоит из двух частей, разделенных между собой знаком «=». Слева от этого знака стоят исходные символы буквенной записи слова, справа – символы которыми они должны замениться в транскрипции.

Для транскрибирования заданного слова последовательно ищется в нем вхождение левой части очередного правила, и если таковое обнаруживается, то вместе неё подставляется правая часть этого правила.

Знак «#» означает начало или конец слова в зависимости от местоположения: если «#» стоит перед символами, то это начало слова; если «#» стоит после символов, то это конец.

Знак «^» означает любые символы в любом количестве между двумя звуками.

Рекомендуется внести в управляющий файл эти группы в порядке номеров, не меняя порядка правил в группах, поскольку порядок замен, очевидно, важен.

Кроме орфоэпических правил, в транскриптор были включены правила, определяющие мягкость и огубленность согласных:

16) әб=әb, әг=әg, әд=әd, әж=әv, әз=әz, әй=әj, әк=әk, әл=әl, әм=әm, ән=әn, әң=әq, әп=әf, әр=әr, әс=әs, әт=әt, әу=әu, әш=әw, еб=eb, ег=eg, ед=ed, еж=ev, ез=ez, ей=ej, ек=ek, ел=el, ем=em, ен=en, ең=eq, еп=ef, ер=er, ес=es, ет=et, еу=eu, еш=ew, іб=ib, іг=ig, ід=id, іж=iv, із=iz, ій=ij, ік=ik, іл=il, ім=im, ін=in, ің=iq, іп=if, ір=ir, іс=is, іт=it, іш=iw, бә=bə, гә=gə, дә=də, жә=və, зә=zə, йә=jə, кә=kə, лә=lə, мә=mə, нә=nə, ңә=qə, пә=fə, рә=rə, сә=sə, тә=tə, уә=uə, шә=wə, бе=be, ге=ge, де=de, же=ve, зе=ze, йе=je, ке=ke, ле=le, ме=me, не=ne, ңе=qe, пе=fe, ре=re, се=se, те=te, ше=we, бі=bi, гі=gi, ді=di, жі=vi, зі=zi, йі=ji, кі=ki, лі=li, мі=mi, ні=ni, ңі=qi, пі=fi, рі=ri, сі=si, ті=ti, ші=wi.

17) өб=өb, өг=өg, өд=өd, өж=өv, өз=өz, өй=өj, өк=өk, өл=өl, өм=өm, өн=өn, өң=өq, өп=өf, өр=өr, өс=өs, өт=өt, өу=өu, өш=өw, үб=үb, үг=үg, үд=үd, үж=үv, үз=үz, үй=үj, үк=үk, үл=үl, үм=үm, үн=үn, үң=үq, үп=үf, үр=үr, үс=үs, үт=үt, үу=үu, үш=үw, бө=bө, гө=gө, дө=dө, жө=vө, зө=zө, йө=jө, кө=kө, лө=lө, мө=mө, нө=nө, ңө=qө, пө=fө, рө=rө, сө=sө, тө=tө, уө=uө, шө=wө, бү=bү, гү=gү, дү=dү, жү=vү, зү=zү, йү=jү, кү=kү, лү=lү, мү=mү, нү=nү, ңү=qү, пү=fү, рү=rү, сү=sү, тү=tү, уү=uү шү=wү;

Поясним знаки, использованное в правилах замены. Латинские знаки в группе 16 означают, что звук является мягким и негубным, в группе 17 цифра «2» после согласного означает, что этот согласный – твердый губной, а в группе 18 цифра «3» после согласного означает, что этот согласный – мягкий губной.

В общем счете фонетический транскриптор составил около 400 правил.

3. Синтез эталонов слов

Эталоны слов распознаваемого словаря формируются из эталонов дифонов, полная база которых в объеме приблизительно трех тысяч создается для каждого диктора заранее [9]. Отметим, что создание такой базы в дальнейшем избавляет пользователя от необходимости создавать какие-либо эталоны голосом.

Под дифоном, соответствующим межфонемному переходу внутри слова, будем понимать участок стандартной длины: 3 окна в 368 отсчетов слева от метки между звуками и 3 таких же окна справа от той же метки. Эталон дифона – набор 6-ти соответствующих векторов. Кроме того, мы используем участок в 3 окна в начале слова и участок в 3 окна в конце слова, условно называя их соответственно начальным и конечным полудифоном слова (переход от молчания к речи и наоборот). Все вектора, входящие в эталоны дифонов, играют роль кодовых векторов и образуют кодовую книгу *B*. Все эталоны дифонов нумеруются, нумеруются также все кодовые вектора.

Каждое слово словаря автоматически транскрибируется, по транскрипции строится цепочка имен дифонов. Каждое из них заменяется эталоном соответствующего дифона. Полученная цепочка векторов образует эталон слова [9].

Тестирование

В результате работы была построена система, распознающая слова по эталонам, синтезированных из дифонов.

В тестировании данной системы участвовали 5 дикторов. Для каждого из них были созданы собственные дифонные базы двух видов: база состоящая из 500 дифонов, в которой каждое звукосочетание имеет только один аналог, и полная база, состоящая из 1000 дифонов, в которой согласные звукосочетания имеют по 4 варианта. После создания дифонной базы, дикторы произносили по 50 слов по два раза: для распознавания слов на основе неполной дифонной базы и для распознавания слов на основе полной дифонной базы. В результате распознавание слов на основе полной дифонной базы оказалось надежнее примерно на 15%.

Полный результат эксперимента представлен в таблице 6.

Таблица 6. Результаты распознавания слов

Звук	**Полная база**	**Неполная база**
Диктор 1	95,4%	80,1%
Диктор 2	94,8%	78,8%
Диктор 3	95,5%	79,4%
Диктор 4	93,5%	77,5%
Диктор 5	94,2%	75,5%

Таким образом, выяснилось, что использование расширенной базы дифонов эффективнее и надежнее.

Заключение

Что означают полученные результаты? Во-первых, мы получили возможность распознавать сверхбольшие словари, так как автоматическое генерирование эталонов облегчает обучение системы. Алгоритм DTW вполне надежен для этого. Полагаем, что дикторонезависимости можно добиться через усреднение эталонов. Но даже пока она дикторозависима, создание дифонной базы займет максимум 2-3 часа.

Наиболее сложной в этой технологии является переход к слитной речи, так как сложно определить границы слов в непрерывной речи. Затем вместо обычного словаря нужен текстовый корпус со всевозможными предложениями и словосочетаниями. Можно распознавать сочетании фраз как целые слова, но таких сочетаний будет много. Поэтому, эффективно использование такой системы для определенной предметной области.

Но раз уж мы сделали шаг в сторону больших словарей, то возможно при кропотливой работе эту проблему можно решить.

Литература

1. Raut, C.K., Bayesian discriminative adaptation for speech recognition, Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on Eng. Dept., Cambridge Univ., Cambridge , 19-24 April 2009, Page(s): 4361 – 4364
2. Lawrence R. Rabiner (February 1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition". Proceedings of the IEEE **77** (2): 257–286. doi:10.1109/5.18626.
3. Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub, Speech Recognition using Artificial Neural Networks and Hidden Markov Models, IEEE MULTIDISCIPLINARY ENGINEERING EDUCATION MAGAZINE, VOL. 3, NO. 3, SEPTEMBER 2008
4. Винцюк, Т.К. *Анализ, распознавание и интерпретация речевых сигналов*. Киев, Наук. думка, **1987.**
5. Negin Najkar , Farbod Razzazi, Hossein Sameti An evolutionary decoding method for HMM-based continuous speech recognition systems using particle swarm optimizationб Pattern Anal Applic, DOI 10.1007/s10044-012-0313-7
6. Mondher Frikha*, Ahmed Ben Hamida A Comparitive Survey of ANN and Hybrid HMM/ANN Architectures for Robust Speech Recognition American Journal of Intelligent Systems 2012 2(1): 1-8 DOI: 10.5923/j.ajis.20120201.01
7. J.P. Hosom, R. Cole, and M. Fanty. Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding. //Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, July 1999.
8. Isolated Digit Recognition Using MFCC AND DTW, International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Volume-1, Issue-1, 2012, pp 59-64
9. Шелепов, В.Ю., Ниценко А., Дорохина, Г.В., Карабалаева, М.Х., Бурибаева, А.К. (2012) О распознавании речи на основе межфонемных переходов. Вестник. Астана: Евразийский национальный университет им. Л.Н.Гумилева, 2012. – Специальный выпуск.– С.436-440
10. Ж. Есенбаев, О. Махамбетов, М. Карабалаева , Текстовый корпус казахского языка, материалы международной научно-практической конференции «Современное казахское языкознание: актуальные вопросы прикладной лингвистики», Алматы, 2012, стр.61-66

С.А. АЛТЫНБЕК, М.М. МУРАТБЕКОВ, А.М. АБЫЛАЕВА, А.С. ТУРГИНБАЕВА

Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан

ЛОГИКА ПОСТРОЕНИЯ АЛГОРИТМОВ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАСПОЗНАВАНИЯ РУКОПИСНОГО КАЗАХСКОГО ТЕКСТА

Целью данной работу является исследование распознавания рукописного казахского текста и проверка распознанного текста на орфографию и морфологию казахского языка. Реализации данной проблемы актуально для Казахстана, которая бы переносила накопленную веками информацию в книгах и рукописных текстах в цифровую информацию. Реализация методов распознавания необходима также в автоматизированных системах, предназначенных для использования в криминалистике, медицине, военном деле. Такие

применения теории распознавания, как кластерный анализ (таксономия), выявление закономерностей в множестве экспериментальных данных, прогнозирование различных процессов или явлений широко используются в научных исследованиях. Большую роль методы распознавания (классификации) играют в активно развивающихся интеллектуальных информационных системах.

В данной работе предлагается подход, состоящий в построении систем распознавания рукописного текста, объединяющих модули выделения признаков и классификатор с применением нейронных сетей, имеющих меньшее количество весов относительно многослойных полносвязных нейронных сетей. Системы должны сами выделять признаки и обладать инвариантностью к искажению входных символьных изображений. Предлагается метод построения таких систем на основе сверточных нейронных сетей.

Исследуем алгоритм пошаговой реализации модулей программ, для мультиагентной нейронной сети, которая одновременно, разными методами распознает текст.

Логика последовательности модулей программы:

1) Последовательное применение графических фильтров:
- удаление шумовых помех
- скелетизация изображения
- округление углов и выравнивание линий

2) Выделение букв:
- разбивка изображения на строки
- разбивка строк на слова
- разбивка слов на буквы методом предсказания длин [1-2,4]
- вычисление угла наклона букв

3) Обучение нейронов Кохонена-Гроссберга:
- векторизация символов с учетом угла наклона
- сравнение выявленных символов с эталонами из базы данных [3,5]

4) Орфографический и грамматический анализ:
- поиск в базе данных слов по найденным буквам
- выявление наиболее вероятных слов в предложении по найденным словам
- предсказание слов по количеству букв и по смыслу предложения

5) Сохранение результатов:
- обучение нейронов по принятому изображению и результатов распознавания
- при отсутствии в базе данных подобного подчерка - создается новый профиль для сохранения эталонных векторов

Помимо вышеуказанных пунктов нейронная сеть, должна будет применить уже известные методы и алгоритмы распознавания для повышения качества. Модель мультиагентной сети позволит методом голосования между всеми методами распознавания выделять лучшие результаты. Кроме того, в мультиагентную сеть можно постоянно добавлять новые подпрограммы, которые в итоге станут способны решать широкий спектр задач машинного зрения.

Предлагаемый для разработки исследование позволит создать систему автоматизации документооборота, которая обязательно присутствуют в средствах ввода бумажных документов, естественно, путем сканирования. Задача распознавания произвольного рукописного текста является актуальной сегодня, и проблема не будет закрыта в ближайшие десятилетия. Задача распознавания рукописных текстов как научная проблема и как информационная технология находится на подъеме, благодаря большому интересу к этой области в коммерческих кругах, среди компьютерных компаний, в научном сообществе.

Литература

1 Бусленко, Н.П. Метод статистических испытаний / Н.П. Бусленко, Ю.А. Шрейдер. М.: Государственное издательство физико-математической литературы, 1961. - 228с.

2 Вапник, В.Н. Восстановление зависимостей по эмпирическим данным / В.Н. Вапник М.: Наука, 1979. - 447с.
3 Вапник, В.Н. Теория распознавания образов / В.Н. Вапник, А .Я. Черновенкис. М.: Наука, 1974. - 414 с.
4 Гмурман, В.Е. Теория вероятностей и математическая статистика: Учеб. пособие / В.Е. Гмурман. М.: Высш. шк., 1999. — 479с.
5 Горелик, А.А. Методы распознавания / А.А. Горелик, В.А. Скрипник.1. М: Высш. шк, 1977. 222 с.

А.А.ШАРИПБАЕВ[1], Г.Ж.ЖЕТИМЕКОВА[2]

[1]Л.Н. Гумилев атындағы ЕҰУ, [2]Е.А.Бөкетов атындағы ҚарМУ

БЕЙНЕНІ ТАНУ ЕСЕПТЕРІНДЕ НАҚТЫ ЕМЕС ЛОГИКАНЫҢ ҚОЛДАНЫЛУЫ ЖӘНЕ ЕРЕКШЕЛІКТЕРІ

Бейнені танудың қазіргі кездегі жетістіктерінің бірі нейрожелілік әдістер арқылы Интернет жүйесі бойынша адамдардың бетінің бейнесін тану болып отыр.

Жасанды нейрондық желі негізінде интеллектуальды жүйелер, бейнелерді тануды, бақылаудың орындалуын, тиімділікті, ассоциативті жады және басқарудың мәселелерін орындап келе жатыр.

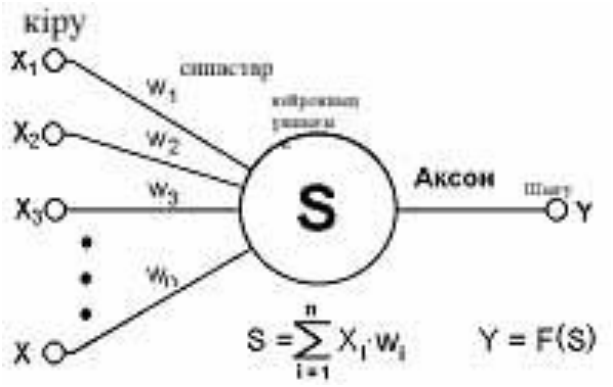
Нейрожелілік әдіс - әртүрлі нейрондық желілердің типінде қолданылатын негізгі әдістердің бірі болып табылады. Бейнені тану үшін әртүрлі нейрондық желілерді қолданудың негізгі бағыттары қолданылады.

Олар:

- бейненің белгіленген белгісі бойынша немесе кілттік сипаттамалардан арылу үшін;
- бейнелер үшін жіктелу;
- тиімді тапсырмаларды шешу.

Бейнені танудың нейрожелілік әдіс арқылы жұмыс жасауы қазіргі кездегі бірнеше саланы қамтамасыз етеді.

Жасанды нейрондық желінің құрылымы төменде көрсетілген:



Сурет 1. Жасанды нейрондық жүйенің құрылымы

Автоматты басқару есептерінде анықталмағандықтың сипаттамасын беру үшін үш тәсіл қолданылады:

- ықтималдық (стохастикалық);
- нақты емес логиканы қолдану (fuzzy logic);
- хаостикалық жүйелер.

Нақты емес логиканың алгоритмін енгізу мүмкін болатын барлық экспертті жүйенің жұмысын соның ішінде келесілерді қарастырады:

- процессті сызықтық емес түрде тексеру (өндіріс);
- өз-өзінен оқылатын жүйелер (классификаторлар) тәуекелді және төтенше жағдайларды зерттеу;
- бейнені тану;
- қаржылық анализ (құнды қағаздар нарығы);
- мәліметтерді зерттеу (корпоративті сақтау).

Нақты емес жүйенің кемшілігі болып келесілер табылады:

- нақты емес жүйені құрастырудың стандартты тәсілінің жоқтығы;
- сәйкес тәсілдермен нақты емес жүйенің математикалық анализіне талдау жасай алмауы;
- ықтималдықпен салыстыру тәсілін нақты емес жүйеге алып қолдану.

Нақты емес логика қазіргі басқару теориясының жедел түрде дамып келе жатқан бағыты. Нақты емес жүйенің негізінде жинақтың теориясы жатыр, онда болжау функциясының элементі жинақтың бинарлы еместігі (иә/жоқ) қарастырылады. Бұл өмірдегі нақты емес логиканың “жақсы”, “жоғары”, “баяу” және т.б. түсініктерін анықтайды. Нақты емес логика логикалық операциялардың бірнеше түрлерімен жұмыс жасайды: біріктіру, қиылысу, терістеу және т.б.

Нақты емес логика мәліметтер қорын және жаңа дәуірдің, кезеңнің эксперттік жүйесін құруға, нақты емес ақпаратты сақтау және өңдеу тәсілдеріне мүмкіндік береді.

Сонымен қатар нақты емес логиканың қолданылу аймағы - әртүрлі сипаттаудағы – электронды жүйелер, технологиялық процесстер және т.б. түрлері жатады. Дәстүрлі анализ тәсілдерімен және ықтималдықтардың тәсілін салыстыра отырып, нақты емес басқарудың нақты, дәл нәтиженің алынатындығын айтуға болады.

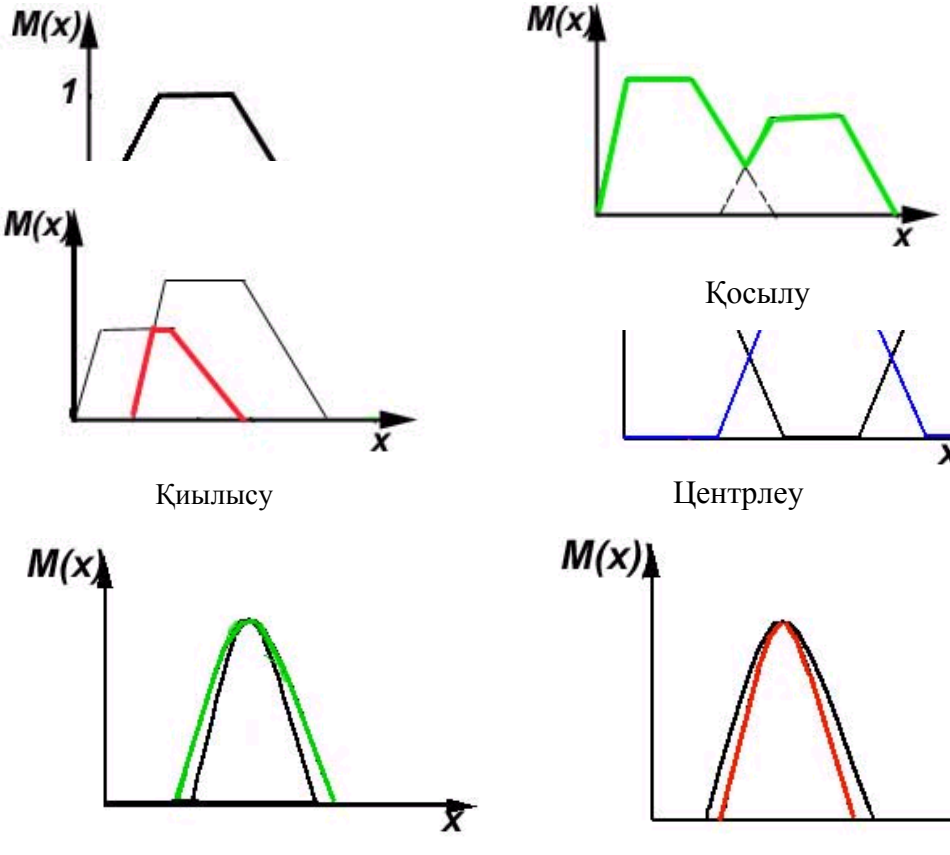
Нақты емес логиканың тәсілімен тапсырманы шешудің алгоритмінің негізгі мінездемесі болып, кейбір нақтылылар жиынтығының, ережелер жиынтығының болуын, әр ереже оқиғалар (шарттар) қасиетінің жиынтығынан және нәтижелерден тұратындығы табылады.

Тапсырма есеп болып қойылғаннан соң, шарттар мен нәтижелерден тұратын арнайы алгоритмдермен өңдеу ережелері іске қосылады. Өңдеу идеясы түрленуден (фаззификация - fz) нақты емес мәні өңдеуден және санды формиада нәтижені шығару жатады. Функцияның типін таңдау шешіліп жатқан есептерге тәуелді болады. Fz – операциясы интегралды Лаплас, Фурье түрлендірулеріне ұқсас және ол бір кеңістіктен екінші кеңістікте өту мүмкіндігіне ие бола алады. Жаңа кеңістікте нақты емес айнымалыны логикалық операциямен қолдану арқылы өңдеуге болады. Алынған нәтиже логикалық өңдеу кері тартуды қолдана отырып (дефаззификация - dfz) – бастапқы санды айнымалы кеңістігіне өтеді.

Нақты емес логикалық негізгі қолданылу ерекшелігі дәстүрлі тәсілді автоматты басқарумен салыстырғандағы айырмашылықты шешу есебі үшін келесі басқарулардан тұрады:

- басқару процесінің жылдамдығын едәуір өсіру, ол нақты емес контроллерді қолдану кезінде орындалады;
- объектілер үшін басқару жүйесін құру мүмкіндігі, дәстүрлі математикамен қиын құрылатын алгоритмдерді құрастыру, қызмет етуі;
- классикалық регулятор базасында адаптивті регуляторды синтездеу мүмкіндігі;
- есептеуіштен ақпаратты өңдеу кезінде кездейсоқ оқиғалардың алгоритмдерінің дәлдігін көтеру;
- басқарушы алгоритимдермен жұмыс жасау кезінде қателік шешімдерді қабылдаудың ықтималдығын азайту.

Нақты емес логика жиынтығына келесі операцияларды қолдануға болады:



Фаззификация - х мәнінің жиынтығы М(х) функциясында жататындығын анықтау болып табылады, демек, х мәнін нақты форматқа ауыстыру.

Дефаззификация - фаззификацияға кері процесс.

Нақты емес логиканың барлық жүйесі бір қағида бойынша жұмыс жасайды: өлшеу құрылғыларының көрсеткізуі фаззификацияланады (нақты емес форматқа ауысады), өңделеді, қалыпты сигнал түріне дефаззификацияланады және құрылғымен орындалуға жіберіледі.

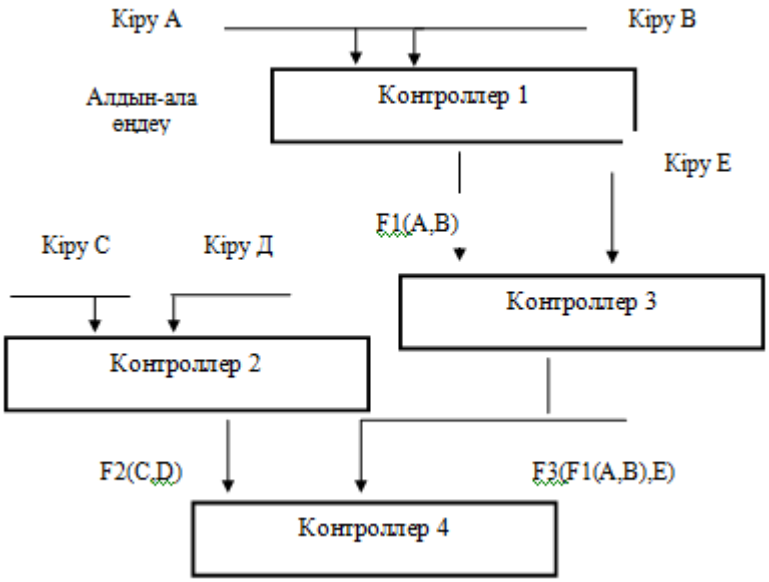
Нақты емес логикада лингвистикалық айнымалы түсініктері енгізіледі, олардың мәндері сан емес; оларды терма деп атайды.

Мысалы, мобильді роботты басқару жағдайында екі лингвистикалық айнымалыны енгізуге тура келеді; диспанция (жаңғырық ара-ақышықтығы) және бағыт (жұмыс жасалып тұрған остің арасындағы бұрыш және жаңғырыққа бағытты ұсыну).

Нақты емес дәстүрлі логиканы қазіргі жүйелерде қолдану келесі факторлармен шектеулі:

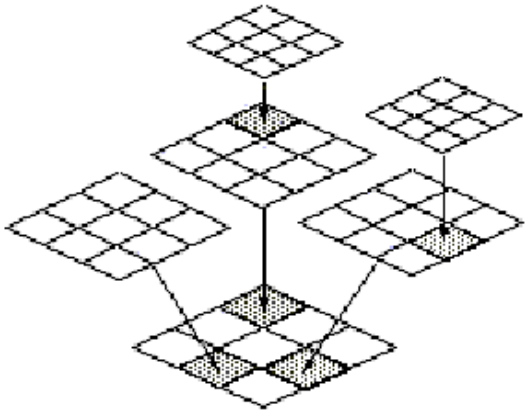
- ереже бойынша басқарудың күрделі жүйесі кірудің үлкен көлеміне ие;
- кіретін айнымалыларды қосу күрделі экспоненциальды есептеуді өсіреді;
- ереже базасы өседі, ол қиын қабылдауға алып келеді (ережелер базасы қолмен теріледі).

Әр элемент, нақты емес желі арқылы алынған әр элемент нақты емес торап ретінде қарастырылады. Осы бір тораптың кіруіне екіншіге кіретін торапты байланыстырсақ, онда есептеу едәуір қысқартылады (Сурет 2). Осы тәсіл нақты емес есептеу алды деп аталады.



Сурет 2. Нақты емес желі арқылы алынатын элемент

Сонымен қатар нақты емес тарапқа шығу мультиплексор көмегімен біріктіруге болады. Ол – жинақ арасындағы ережелер бағасымен өңделеді.



Сурет 3. Нақты емес мультиплексор

Нақты емес логикасының құрылғының жалпы құрылымы.

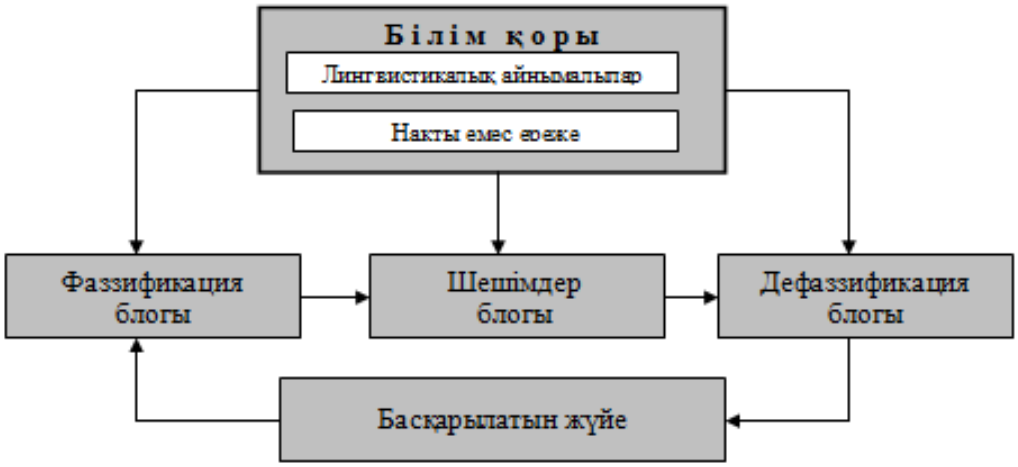
Жалпы микроконтроллердің құрылымы сурет 3-те көрсетілген. Ол келесі негізгі бөліктерден тұрады:

- фаззификация блогынан;
- білім қорынан;
- шешімдер блогынан;
- дефаззификация блогынан.

Фаззификация блогы нақты бірлікті (crisp) объектіні басқарудағы нақты емес бірлікке ауыстырылады немесе түрлендіріліледі. Ол мәліметтер білім қорында лингвистикалық айнымалымен сипатталады.

Шешімдер блогы нақты емес шарттан (if - then) білім қорындағы ереже, ол нақты емес кіру мәліметтерін түрлендіру үшін қолданылады (Сурет 4).

Дефаззификация блогы нақты емес мәліметтерді шешімдер блогынан шығу кезінде нақты бірлікке түрлендіру кезінде қолданылады.



Сурет 4. Микроконтроллердің негізгі жұмыс жасауы

Қарапайым кластерлеу келесі қадамдардан тұрады:
- үлгіні ұсыну (белгілерді таңдау немесе белгілеу);
- үлгілердің ұқсастығын анықтау – мәліметтердің аймағына сәйкестігін өлшеу;
- кластеризация немесе топтау;
- мәліметтерді абстракциялау (қажет болған жағдайда);
- шығару бағасы (қажет болған жағдайда).

Әдебиеттер

1. Фролов А.А., Муравьев И.П. Информационные характеристики нейронных сетей. - М.: Наука, 1988 – 289 ст.
2. Фролов А.А., Муравьев И.П. Нейронные модели ассоциативной памяти.- М.: Наука, 1987.- 160 ст.
3. Фу К. Структурные методы в распознавании образов.- М.: Мир, 1977.- 320 ст.
4. Фукунга К. Введение в статистическую теорию распознавания образов.- М.: Наука, 1982.- 367 ст.

МӘТІНДЕРДІ СЕМАНТИКАЛЫҚ ӨҢДЕУ ЖҮЙЕЛЕРІ
СИСТЕМЫ СЕМАНТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ
SYSTEMS OF SEMANTIC TEXT PROCESSING

Г.Т. БЕКМАНОВА, Л. ЖЕТКЕНБАЙ

Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
«Жасанды зерде» ҒЗИ, Астана, Қазақстан

ҚАЗАҚ ТІЛІНІҢ КҮРДЕЛІ СӨЗДЕРІН ФОРМАЛДАУ НЕГІЗІНДЕ ЖАСАУ

1980 жылдарға дейінгі еңбектерде сөз тұлғасының бір түрі – күрделі сөз деген ұғым морфологиялық категория ретінде қарастырылып келгені, 1989 жылдан бастап сөзжасам жүйесінде қарала бастағаны белгілі [1].

Күрделі сөздер қазіргі тіл біліміндегі ең күрделі мәселе болып табылады, оның анықталмаған сырлары мен қырлары өте көп. Күрделі сөздер жоқ тілді көрсету өте қиын, бірақ күрделі сөздер, әсіресе, неміс, ағылшын, жапон, хинди, орыс тілдерінде кең тарағанын жазба деректерден көруге болады. Күрделі сөздердің тілдерде алатын орны бірдей емес. Ол тілдердің лексикалық жүйесінен әркелкі орын алады. Кейбір тілдерде сөздерді біріктіру қосымша тәсіл болып, оның нәтижесі елеусіз болады да, сөздіктің периферия бөлімінен орын алады. Келесі бір тілдерде ол кең тараған, негізгі тақырыпқа жатады. Соңғы топтағы тілдерге жапон тілі жатады. Лексикадан алатын орнына қарағанда жапон тілін тіпті күрделі сөздің тілі деуге де болады. Осындай күрделі сөздерге бай тілдерге түркі тілдері, оның ішінде қазақ тілі де жатады. Ғалымдар күрделі сөздердің түркі тілдерінде жалпы көне құбылыс екенін, олардың орхон жазба ескерткіштерінің тілінде де барын айта келіп, ескерткіштер тіліндегі кісі аттарынан, жер су аттарынан көптеп мысал келтіреді. Күрделі сөздер барлық басқа туынды сөздер сияқты сөзжасам әрекетінің нәтижесіне жатады. Анығырақ айтқанда, ол – туынды сөздер тобындағы сөздер лексикалық бірлік – деп танылатыны белгілі. Олай болса, аналитикалық синтаксистік тәсіл арқылы жасалған күрделі сөздер тілге белгілі. Күрделі сөздер сөзжасамның тілден үлкен орын алатын аналитикалық синтаксистік тәсіл арқылы жасалады. Сөз жасамның аналитикалық тәсілі ғылымдағы көне әдістердің бірі болып табылатыны баршамызға мәлім. Осы тәсілдің түркі тілдерінде атам заманнан бері қолданылып келе жатқанын орхон жазба ескерткіштердің тілдерінде қолданылған йашыл үгүз – көк өзен, күнтүз – күндіз, Беш балық – Бес балық (қала аты), Қара құм (жер аты), Білге қаған (хан аты) т.б. күрделі сөздерден көруге болады. Келтірілген мысалдар күрделі сөздердің көне замандардан бері қолданылып келе жатқаны екіншіден, олардың географиялық атаулар, яғни өзен, көл, жер, қала және кісі аттарында жиі кездесетіні айқын көрінеді. Осыған қарап, кейбіреулер аналитикалық тәсіл арқылы сөз жасау тәжірибесі түркі тілдерінде осы жер – су, кісі аттарынан басталмады ма екен деген ой айтады[2].

1 Қазақ күрделі сөздерінің түсінігі

Қазақ тілі сондай-ақ басқа түркі тілдерінде күрделі сөз мәселесі Қазан төңкерісіне дейін арнаулы зерттеудің объектісіне айнала қоймаған еді. Түркі тілдерінің грамматикалық құрылысын зерттеген ғалымдардың еңбектерінде ол туралы азды-көпті мағлұмат беріліп келді. Соның өзінде күрделі сөз лексикалық емес, негізінен, тілдің грамматикалық құрылысы, оның ішінде сөз табы категориясына қатысты жағынан қарастырылатын.

Ал қазақ тілі мамандарынан бұл мәселені алғаш сөз еткен профессор Қ.Жұбанов. Ол кісі 1930 жылдардың өзінде-ақ күрделі сөздер қанша сөзден біріксе, кіріксе де тұтас бір ғана мағынаны білдіреді, бір ғана заттың атауы болады, сөйлемде де сол жұбын жазбай, бір ғана мүше болады, кейбіреулерінің құрамындағы сыңарларының дыбыстық өзгеріске түсуінен сыртқы түр-тұрпаты басқаша болып қалыптасады, бір ғана екпінді иеленеді деп дұрыс тұжырымдайды.

Профессор Н.Т.Сауранбаев қазақ тіліндегі йзафеттік топтың түрлеріне тоқталған жерінде бірінші топтың өзгешелігі – синтаксистік шеңберден шығып, морфология, оның ішінде сөз

тудыру жүйесіне айналған тәсіл екенін көрсете келіп, йзафеттің бұл түрі қазір де тұтас бір атау (лексема) болып ұғынылатынын айтады.

Біріккен сөзге арнайы еңбек жазған А.Ермеков біріккен сөздердің сыртқы түр-тұрпаттық белгілі қасиеттеріне, біріншіден, тұтас тұлғалануын, екіншіден, сөйлемнің бір ғана мүшесінің қызметін атқаруын, үшіншіден, белгілі сөз табының құрамына кіруін, төртіншіден, бірұдайы интонациямен айтылуын, бесіншіден, кейбір компоненттерінің бүтіндей, ия жартылай дербес мағынасынан айрылуын, алтыншыдан, бірге жазылуын, жетіншіден, сөздікте түтін бір сөз ретінде берілуін жатқызады.

Қазақ тіліндегі біріккен сөзбен шұғылданған екінші бір тіл маманы Г.Жәркешова біріккен сөздердің ерекшеліктері дегенге мағына тұтастығын, сөйлемге дербес бір ғана мүше болуын, компоненттерінің қалыптасуына дыбыс үндесу заңы қатысуын жатқызады.

Профессор А.Ысқақов соңғы кезде шыққан еңбектерінде күрделі сөздердің аса маңызды мәселелерін қамтып, дұрыс тұжырымдар жасады, бірқатар өзекті мәселелерді алға қойды. Күрделі сөздердің әлі де анықталмаған, айқындалмаған мәселелері көп екенін айта отырып, мәселелердің негізгі түйіндері біріккен сөздер мен фразалық тіркестердің, біріккен сөздер мен идиомалық тіркестердің, біріккен сөздер мен күрделі атау сөздердің, фразалық тіркес пен идиомалық тіркестің араларын ашу, олардың ара қатынастарын анықтау мәселелерімен байланысып жатқанын көрсетеді.

Профессор Қ.Аханов «Грамматика теориясының негіздері» деген еңбегінде сөзге, соның ішінде күрделі сөзге де тән және оны сөз тіркесінен ажыратын жалпылама белгілер делініп семантикалық тұтастық, морфологиялық тұтастық, синтаксистік тұтастық белгілері айтылып жүр дей келіп, күрделі сөздердің сөз тіркесінен айырмашылығын осы тұрғыдан сөз етеді.

Басқа түркі тілдері сияқты, қазақ тілінде де күрделі сөздерді компоненттерінің бір-бірімен байланысу тәсілі тұрғысынан әдетте екіге бөледі. Құрамындағы сыңарлары сабақтаса байланысып құрылғандарын біріккен сөзге, ал салаласа байланысып құрылғандарын қос сөздерге жатқызады.

Күрделі сөздердің арғы тегі, негізінен, сөз тіркесі; некен-саяқ болмаса, көпшілік күрделі сөздерге негіз болған сөз тіркесін тауып, оны түсіндіруге болады.

Нағыз күрделі сөздің үлгісіне – бүгін, биыл, сексен, сарала, жарымжан, көк ала, жақсы көру, еңбек ету, шығарып салу т.б. тәрізді сөздер жатады.

Жалаң сөздер сияқты да сыртқы пішін, ішкі мән-мағына, яғни бірден көзге түсіп, көңілге қонатын құрылым болады. Ол оны құрастырушы компоненттердің тұлғалық, мағыналық бірлігінен тұрады.

Тарихи тұрғыдан алып қарағанда күрделі сөздердің арғы тегі, негізінен, сөздердің еркін тіркесі болады дедік. Осы грамматикалық категорияға жататын күрделі сөздердің пайда болуы, әрине кездейсоқ, өзінен-өзі бола қалатын жай емес; күрделі және белгілі бір заңдылықтарға негізделетін, соның нәтижесінде пайда болатын құбылыс болмақ. Ондай заңдылықтардың бірі тілдік дыбыс жүйесі, екіншісі лексикасы, үшіншісі грамматикасына келіп саяды.

Күрледі сөз деп екі я одан да артық сөз тізбегінен құралып, лексика, лексикалық-грамматикалық мағыналары мен тұлға, сөйлемдегі қызметі жағынан бөлшектенбей, тұтасымен белгілі бір ұғымның атауы ретінде жұмсалатын сөздерді айтамыз [3].

Күрделі сөздер екі не одан да көп сыңарлардан жасалады. Сондықтан күрделі сөздердің негізгі белгісінің біріне оның күрделі құрылымы жатады, сол арқылы ол дара сөздерден ерекшеленеді. Мысалы, ата-ана, елтаңба, елбасы, таң намаз, шұбар ала, ой-пікір, тоқсан жеті, теміржол, күні бүгін, майшабақ т.б [2].

2. Қазақ күрделі сөздерінің түрлері

Күрделі сөздер жасалу жолына қарай кіріккен сөздер, біріккен сөздер, қос сөздер, қысқарған сөздер, тіркесті сөздер болып бөлінеді.

Кіріккен сөз. Қазақ тіліндегі кіріккен сөздерге бәйшешек, белбеу, қонағасы, қолқанат, бүгін, биыл, сексен, қарлығаш және т.б. сөздер жатады. Мәселен, белбеу – белге буынатын

бау, сексен – сегіз он, бүгін – бұл жыл, қоян – қой сияқты аң т.б. дені екі, ал былтыр (бір жыл дүр) үш сөздің тіркесінен пайда болған. Тіл мамандарының қай-қайсысы болмасын кіріккен сөздерге компоненттері дыбыстық өзгеріске ұшырау салдарынан бастапқы түр-тұрпатын өзгерткен сөздерді жатқызады [3].

Енді кіріккен сөздердің құрылымдық сипатының әуел бастағы қалпынан өзгеше болып қалыптасуына ұйтқы болған тілдік заңдылықтарға тоқталайық. Тіл мамандарының көрсетулеріндей, мәселен, бәйшешек, бәйтерек сөздерінің құрамындағы бай деген элемент әуел баста бай формасында болғаны, кейін соңғы компоненттердегі жіңішке дауысты дыбыстардың ықпалынан «а» фонемасы «ә» фонемасына айналып, бүтін сөз бәйшешек, бәйтерек болып кеткені мәлім; Демек, бәйшешек, бәйтерек, белбеу тәрізді сөздердің бастапқы тұлғасын өзгертіп қалыптасуына үндестік заңы ұйтқы болған [3].

Біріккен сөз. Күрделі сөздердің бір түрі – біріккен сөздер. Біріккен сөздердің екі түбірден бірігіп, жаңа бір ұғым тудыратынын айту оңай, ал біріккен сөздерді сөз тіркесінен ажыратып алу, олардың шегін айқындау оңай емес. Олай дейтініміз, біріккен сөздер мен сөз тіркестері бір-бірімен мағыналық жақтан да, құрылымы, құрамы жағынан да ұқсас, өзара астарласып жатады. Біріккен сөздер мен тіркес сөздердің құрамын ажыратудың жолы салыстыру әдісі арқылы іске асырылады. Мысалы, ақсақал, Ақбозат, Темірқазық, соқыртеке, итмұрын деген сөздерді бірге жазып,әрқайсысының мағыналарын ажыратуда олардың синонимдерін жарыстыра айтып түсіндіру керек. Демек, ақсақал сөзі қарт, Ақбозат, Темірқазық жұлдыздың аты.

Біріккен сөз бен күрделі сөз тіркестерінің арасындағы айырманы нақтылаған соң, біріккен сөздердің кем дегенде екі немесе үш түбірлі сөздерден бірігіп, жаңа бір ұғым тудыратыны белгілі болды [4].

Қос сөздер. Күрделі сөздердің бір түрі – қос сөздер.

3. Қазақ күрделі сөздерін формалдау ережелері

Күрделі сөздер жасалу жолына қарай кіріккен сөздер (бүгін, жаздыгүні, биыл), біріккен сөздер (баспасөз, елтаңба, елбасы, Темірқазық), қос сөздер (үлкен-кіші, аяқ-табақ), қысқарған сөздер (ЕҰУ, АҚШ, БҰҰ), тіркесті сөздер (таң намаз, қара торғай, қара көк, жүз бес) болып бөлінеді.

Таңбалық амалдардың барлығы «Компьютерлік лингвистика» деген тілдің құрамы мен қасиеттерін зерттейтін ғылым саласында анықталған. Таңбаларда анықталған амалдар арқылы *таңбалық өрнектер* құрастырылады.

Таңбалық өрнектерді өңдеп көрсету үшін таңбалық амалдардың *ассоциативтік, дистрибутивтік* және басқа заңдылықтары пайдаланылады /13/, /14/, /15/. Осындай заңдылықтар берілген күрделі өрнектерді қарапайымдап, ондағы амалдар санын қысқартады және олардың орындалуын жеңілдетеді. Олардың практикадағы маңызы өте зор.

Таңбаларда берілген екі таңбалық тізбеден бір таңбалық тізбе құруға мүмкіндік беретін *құрастыру* амалдары анықталған.

Күрделі сөздердің жасалу жолдарын компьютерлік лингвистика көмегімен формалдауды қарастырайық. Кез келген тілден таңба ретінде бөлінбейтін бірлік (*фонема-дыбыс, графема-әріп, цифр, тыныс белгісі, жақша және т.б.*) немесе құрылымдық бірлік (*морфема, лексема және т.б.*) алынады. Берілген таңбалармен жұмыс істеу үшін осы таңбаларға *амалдар* қолдану керек. Ал амалдарды қолдана білу үшін олардың анықтамаларын, белгілерін және қасиеттерін білу қажет.

Таңбаларда берілген екі таңбалық тізбеден бір таңбалық тізбе құруға мүмкіндік беретін *құрастыру* амалдары анықталған.

Құрастыру амалдарының ішінде ең қарапайымы – *конкатенация (тіркеу)* амалы. Егер бұл амалды ‘·’ таңбасымен белгілесек, онда оны берілген екі айнымалы таңбалық шамалар α_1және α_2үшін былай анықтауға болады: α_1-дың мәнінің оң жағына α_2-нің мәнін тіркегеннен кейін γ тізбесі шығады және ол мына түрде жазылады $\alpha_1 \cdot \alpha_2 = \gamma$.Осы айтылған ереже күрделі сөздердің жасалуына сәйкес келеді. Мысалы, егер α_1– дың мәні ‘ЕЛ’, ал α_2-нің мәні ‘ТАҢБА’

болса, онда γ -ның мәні ‘ЕЛТАҢБА’ болады. Енді күрделі сөздің қысқарған сөздерден жасалатынын өрнектеп көрелік. Бізде $\alpha_1 \cdot \alpha_2 \cdot \alpha_3 \ldots \cdot \alpha_n = \gamma$ болсын. Мысалы, α_1= Е, α_2= Ұ, α_3= У нәтиже $\alpha_1 \cdot \alpha_2 \cdot \alpha_3$=ЕҰУ.

Мұндағы $\alpha_{11} \cdot \alpha_{12} \cdot \alpha_{13} \ldots \cdot \alpha_{1n} = \alpha_1,\ \alpha_{21} \cdot \alpha_{22} \cdot \alpha_{23} \ldots \cdot \alpha_{2n} = \alpha_2,\ \alpha_{31} \cdot \alpha_{32} \cdot \alpha_{33} \ldots \cdot \alpha_{3n} = \alpha_3$

α_{11} =Е, α_{12} =У, α_{13} =Р, α_{14} =А, ... α_{1n} =Я

α_{21} =Ұ, α_{22} =Л, α_{23} =Т, α_{24} =Т, ... α_{2n} =Қ

α_{31} =У, α_{32} =Н, α_{33} =И, α_{34} =В, ... α_{3n} =І

Е·У·Р·А·З·И·Я Ц Ұ·Л·Т·Т·Ы·Қ· Ц У·Н·И·В·Е·Р·С·И·Т·Е·Т·І=ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ. Мұндағы Ц – кетікті көрсетеді, яғни сөздердің бөлек жазылатынын бірақ бір мағына білдіретінін көрсетеді.

Егер α_1= ҚАРА, α_2= КӨК болса, онда $\gamma = \alpha_1 \cdot \alpha_2$= ҚАРА· Ц ·КӨК деген таңбалық өрнек шығады. Бұл өрнек бізде сөздердің тіркесу арқылы жасалуын сипаттайды.

Енді қазақ күрделі сөздеріне байланысты ереже жасап, оны жақшалық әдіс көмегімен формалдауды қарастырайық.

Келісім: Алдымен әріптерді бір жүйеге келтіріп, төмендегідей етіп белгілеп алайық.

АОҰЫЭ!01	ЖЗ!11
ӘӨҮІЕЯИЮ!02	ЖЗ!12
МНҢ!03 //жуан әріпке бітетіндік	П!13
МНҢ!04 // жіңішке	П!14
РУЙ!05	К!15
РУЙ!06	К!16
Л!07	Қ!17
Л!08	СТШ!18
БГҒД!09	СТШ!19
БГҒД!10	

Ереже 1. Дауысты дыбыстан басталатын сын есімнің алдыңғы әрпіне п – дауыссызын жалғау арқылы күрделі сөз жасауға болады. Жоғарыдағы келісім бойынша:

01Х! ((01п)-Х) - ап-аласа
01Х! ((01п)-Х) - ап-алыс
01Х! ((01п)-Х) - ап-аласа
01Х! ((01п)-Х) - ап-арық
01Х! ((01п)-Х) - ап-ащы
01Х! ((01п)-Х) - оп-оңай
01Х! ((01п)-Х) - ұп-ұзын
01Х! ((01п)-Х) - ып-ыстық
02Х! ((02п)-Х) - әп-әдемі
02Х! ((02п)-Х) - үп-үлкен
02Х! ((02п)-Х) - өп-өңді
02Х! ((02п)-Х) - еп-ерсі

Ереже 2. Дауыссыз дыбыстан басталатын сын есімнің алдыңғы екі әрпінен кейін п – дауыссызын жалғау арқылы күрделі сөз жасауға болады.

Жоғарыдағы келісім бойынша:

0301Х! ((0301п)-Х) - моп-момын
0301Х! ((0301п)-Х) - нып-нығыз
0402Х! ((0402п)-Х) - мөп-мөлдір
0402Х! ((0402п)-Х) - нәп-нәзік
0901Х! ((0901п)-Х) - боп-бос
0901Х! ((0901п)-Х) - боп-боз
0901Х! ((0901п)-Х) – дап-дайын
0901Х! ((0901п)-Х) - доп-домалақ
1002Х! ((1002п)-Х) - біп-биік

1002Х! ((1002п)-Х) - бөп-бөтен
1002Х! ((1002п)-Х) - бөп-бөлек
1002Х! ((1002п)-Х) - дәп-дәл
1002Х! ((1002п)-Х) - дөп-дөңгелек
1101Х! ((1101п)-Х) - жап-жасыл
1101Х! ((1101п)-Х) - жап-жақсы
1101Х! ((1101п)-Х) - жып-жылы
1101Х! ((1101п)-Х) - жұп-жұмсақ
1101Х! ((1101п)-Х) - жап-жаңа
1101Х! ((1101п)-Х) - жап-жақын
1202Х! ((1202п)-Х) - жіп-жіңішке
1202Х! ((1202п)-Х) - жеп-жеңіл
1602Х! ((1602п)-Х) - көп-көрім
1602Х! ((1602п)-Х) - кәп-кәрі
1602Х! ((1602п)-Х) - кәп-кәрі
1701Х! ((1701п)-Х) - қап-қара
1701Х! ((1701п)-Х) - қып-қызыл
1701Х! ((1701п)-Х) - қап-қатты
1701Х! ((1701п)-Х) - қоп-қою
1701Х! ((1701п)-Х) - қоп-қоңыр
1801Х! ((1801п)-Х) - сап-сары
1801Х! ((1801п)-Х) - сұп-сұры
1801Х! ((1801п)-Х) - шып-шымыр
1801Х! ((1801п)-Х) - шұп-шұбар
1902Х! ((1902п)-Х) - сеп-семіз
1902Х! ((1902п)-Х) - тәп-тәтті
1902Х! ((1902п)-Х) - тәп-тәуір
1902Х! ((1902п)-Х) - түп-түзу

Ереже 3. Жалаң сөздің қайталануынан күрделі сөз (қайталама қос сөз) жасауға болады.

Х! ((Х)-Х) - арбаң-арбаң
Х! ((Х)-Х) - әрең-әрең
Х! ((Х)-Х) - бір-бір
Х! ((Х)-Х) - биік-биік
Х! ((Х)-Х) - белес-белес
Х! ((Х)-Х) - қап-қап
Х! ((Х)-Х) - мая-мая
Х! ((Х)-Х) - гүрс-гүрс
Х! ((Х)-Х) - тал-тал
Х! ((Х)-Х) - сылқ-сылқ
Х! ((Х)-Х) - тез-тез
Х! ((Х)-Х) - ербең-ербең
Х! ((Х)-Х) - күрс-күрс
Х! ((Х)-Х) - қарын-қарын
Х! ((Х)-Х) - тау-тау
Х! ((Х)-Х) - бәрі-бәрі
Х! ((Х)-Х) - қандай-қандай
Х! ((Х)-Х) - қалай-қалай
Х! ((Х)-Х) - не-не

Ереже 4. Бұйрық райлы етістікке қосымшасын (етістікке -а, -е, -й көсемше жұрнағын жалғау арқылы) сөздің қайталануынан күрделі сөз (қайталама қос сөз) жасауға болады.

((Х06))! (((Х06)е)-Х06(е))- Көре-көре
((Х06))! (((Х06)е)-Х06(е))-жүре-жүре
((Х05))! (((Х05)а)-Х05(а))- бара-бара
((Х02))! (((Х02)е)-Х02(е))-билей-билей
((Х06))! (((Х06)е)-Х06(е))-жүгіре-жүгіре
((Х05))! (((Х05)а)-Х05(а)) - тұра-тұра
((Х08))! (((Х08)е)-Х08(е)) - күле-күле
((Х08))! (((Х08)е)-Х08(е)) - біле-біле
((Х02))! (((Х02)е)-Х02(е))-сөйлей-сөйлей

Ереже 5. Жалаң сөзді қайталау арқылы және оның алдыңғы сыңарына -па, -пе, -та, -те, -ма, -ме, -ба, -бе, -да, -де, қосымшасын жалғау арқылы күрделі сөз жасауға

болады. Егер сөз қатаң дауыссызға аяқталса онда -па, -пе, ұяң дауыссызға аяқталса -ба, -бе, үнді дауыссызына аяқталса -ма, -ме, -да, -де қосымшалары жалғанады.

((X19))! (((X19)пе)-X) - бетпе-бет
((X06))! (((X06)де)-X) - бірде-бір
((X11))! (((X11)ба)-X)–ауызба-ауыз
((X12))! (((X12)бе)-X) - жүзбе-жүз
((X07))! (((X07)ма)-X) - жолма-жол
((X07))! (((X07)ма)-X) - қолма-қол
((X08))! (((X08)ме)-X) – дәлме-дәл
((X04))! (((X04)де)-X) - кімде-кім

Ереже 6. Екі сөздің бірігуі арқылы және бастапқы компонент дауыстыға аяқталып, соңғы компонент дауысты фонемадан болғандықтан, бастапқы сөздің соңғы дауысты дыбысы ығысып шығып қалуы арқылы күрделі сөз жасауға болады.

$$\begin{cases} У_1 + У_2 = У \\ X_1 + X_2 = X \end{cases}$$

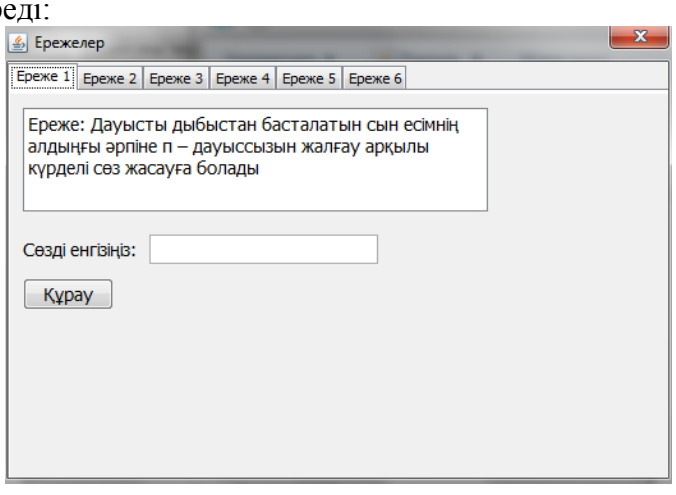
Жүйені қосу тәсілімен шешетін болсақ: $У_1 + У_2 + X_1 + X_2 = У + X$

Мұндағы $У_2$, X_1- дауысты дыбыстар.

Енді

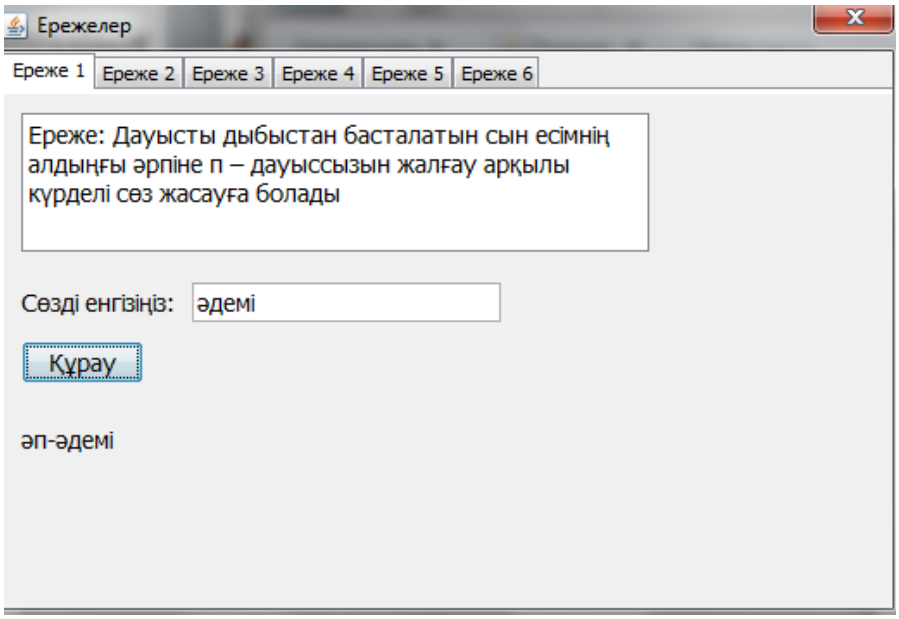
Алма+аты=Алматы
ала+аяқ=алаяқ
алты+атар=алтатар
бие+емшек=биемшек
Қара+аспан= Қараспан
қара+ала=қарала
Қанды+ағаш=Қандағаш
сапты+аяқ=саптаяқ
сары+ала=сарала
сары+ағаш=сарағаш
шыны+аяқ=шынаяқ
өлі+ара=өлара
жалма+ауыз=жалмауыз
құла+ала=құлала
Мақта+арал=Мақтарал

Қазақ күрделі сөздерін формалдау ережелері бізге күрделі сөздерін автомат түрінде шығару мүмкіндік береді:



Сурет 1. Ережелер терезесі

Жұмысты жалғастыру үшін «Сөзді енгізіңіз» өрісіне «Ереже 1» - ге сәйкес өзімізге қажетті сөзді енгіземіз (Сурет 2).



Сурет 2. Ереже 1

Жоғарыда көріп тұрғандай, «әдемі» сөзін енгізген соң құрау батырмасын басамыз. Сонда бізге ол автоматты түрде «Ереже 1» - ге сәйкес күрделі сөзді жасап береді.

Әдебиеттер

1 www.egemen.kz/?p=3486
2 Жанпейісов Е., Хұсайын К. және т. б. Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис. – Астана: Астана, 2002. 152б.
3 Шәкенов Ж. Қазақ тіліндегі күрделі сөздер мен күрделі тұлғалар. – Алматы: Ана тілі, 1991, Б. 3-20
4 Аханов К., Б.Кәтенбаева , Әбдіғалиева Т. Қазақ тілі оқулығының методикалық нұсқауы. Алматы: Рауан, 1990, Б. 19-27

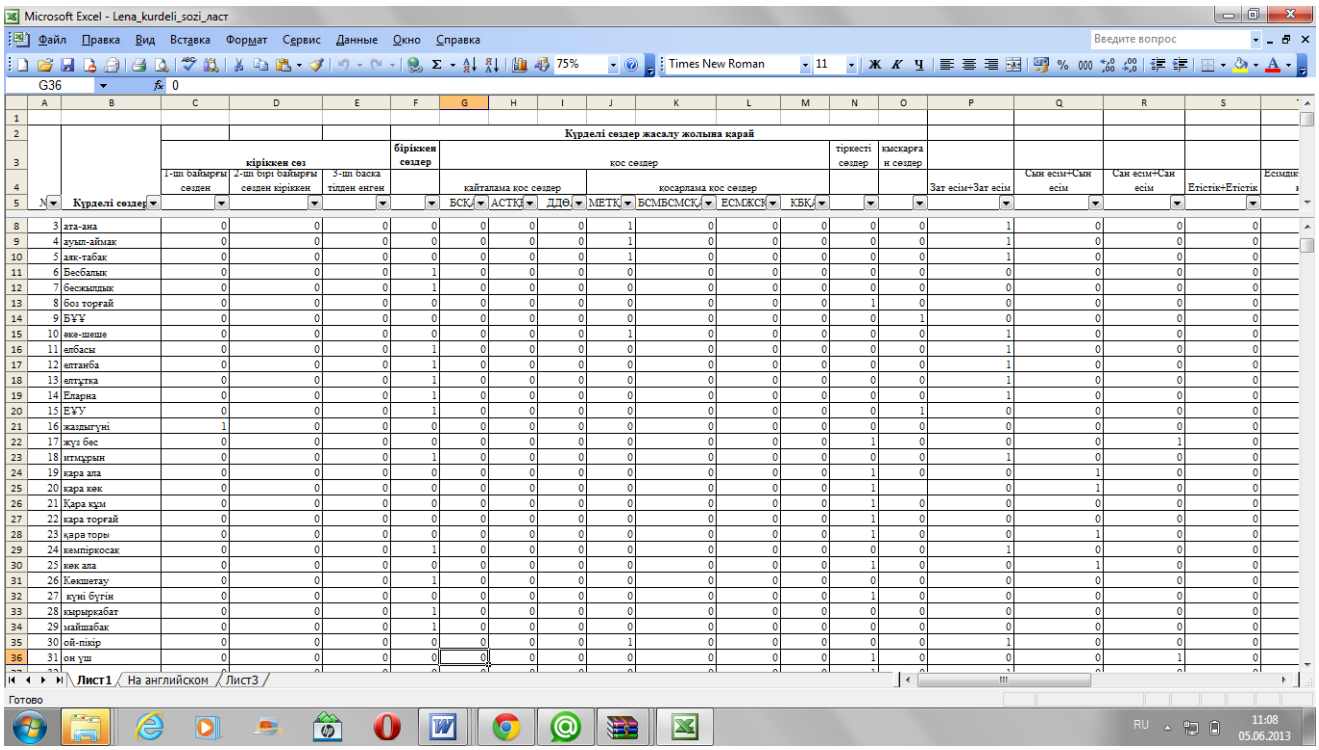
Г.Т. БЕКМАНОВА, Л. ЖЕТКЕНБАЙ

Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
«Жасанды зерде» ҒЗИ, Астана, Қазақстан

ҚАЗАҚ КҮРДЕЛІ СӨЗДЕРІН ТҮРЛЕНДІРУДІҢ СЕМАНТИКАЛЫҚ МОДЕЛІ

1 Қазақ күрделі сөздерінің семантикалық базасын құру

Семантикалық белгілер ретінде күрделі сөз, күрделі сөздердің түрлері, сөз таптары алынады.



Сурет 1. Семантикалық кесте

Ең бірінші әрбір күрделі сөздердің түрлері бойынша сөздің қай топқа жататынын, содан кейін сөздің сыңарлары қай сөз табынан жасалғанын анықтаймыз. Әрі қарай, осы топтардың қиылысы ізделінеді, егер олар болатын болса, онда осы топтар қиылысу болмайтындай етіп бөлінеді, яғни декомпозиция жасалынады.

Сөз түрлендіру және сөз тудыру процесі сөздің бастапқы қалпын оның морфологиялық белгілерін айқындау және оның семантикалық белгілерін білімдер базасынан оқу мақсатында бөлшектеп талдау арқылы табуға негізделеді. Одан әрі қарай сөзтүрленудің траекториясы анықталады, ал сөзтүрлену процесі семантикалық желі негізінде жүзеге асырылады, сосын сөзжасамды және оның морфологиялық ақпараттарын сөзжасамдар сөздігіне жазу.

2 Қазақ күрделі сөздерін граф түрінде бейнелеу

Табиғи тілдегі мәтінді өңдеу процесін бірнеше деңгейге бөлуге болады: талдау, анализ және синтез. Талдау деңгейін табиғи тілдегі мәтінді формалданбаған түрден формалданған ішкі ұғымға түрлендіретін функция ретінде анықтаймыз. Анализ деңгейі – ішкі ұғымда бар деректерді түрлендіру және соның негізінде формалды түрде жаңа деректер шығару функциясы. Синтез деңгейі – ішкі формалданған ұғымға бара бар табиғи тілде жауап құрастыру функциясы.

Талдау деңгейін бірнеше тізбектелген операцияларға бөлуге болады: морфологиялық, синтаксистік және семантикалық талдау. Морфологиялық талдау мәтіндегі бөлек сөздерді алып, оларды морфемаларға бөлу арқылы жасалады. Табиғи тілде мәтіннің синтаксистік талдау операциясы оның барлық синтаксистік белгілерін және семантикалық талдау үшін керек болатын сол сөздердің синтаксистік байланысын анықтаудан тұрады. Семантикалық талдау жүйенің жадында сыртқы әлемнің ұқсас моделінің болғанын талап етеді. Бұл талдауда ұқсас моделдің білімдерімен бірге мәтінде тікелей орналасқан деректерді салыстыру арқылы мәтіннің ішкі формалданған ұғымын тиянақты құрастырылады.

Морфологиялық сөз түрлену және сөзжасамдарды талдау есептерін шешу үшін формалды Маккаллок-Питтстің нейронды желісінің қасиетіне жақын семантикалық нейронды желіні қолданамыз.Маккаллок-Питтс нейронды желісіндегі бөлек нейрондар *және, немесе* және *не*

логикалық операциялары сияқты бейнеленеді. Логикалық операцияның орындалуының нәтижесіне байланысты нейрондар *ақиқат* және *жалған* логикалық мәндеріне сәйкес келетін *қозу* немесе *тыныштық* күйінде болады. Әрбір нейронның жұмысымен байланысты уақытылы кідіру осында локикалық жүйемен алынатын тиімді және конструкциялық сипатына кепілдік береді.Желідегі нейрондар ақырлы және нейронның күйлер саны да ақырлы болғандықтан, мұндай нейронды желілер күйі ақырлы болатын автоматтар болып табылады. Маккаллок-Питтс желісіндегі нейрон тек екі логикалық күйде болуы мүмкін және логика алгебрасының функциясын орындауды ғана қамтамасыз етеді. Табиғи тіл нақты емес және толық емес түсініктермен жұмыс істейді Семантикалық нейронды желінің Макколлок-Питтстің желісінен айырмашылығы мынада, Макклок-Питтс желісінде Бульдік алгебраның логикалық операциялары орындалады, ал семантикалық нейронды желіде бұлдыр логиканың операциялары орындалады. Бұлдыр логикада тұжырымдаманың ақиқаттығы дәрежесін анықтау үшін сенімділік факторын қолданамыз – кейбір интервалда орналасқан сан, мысалы 0 мен 1 аралығы. Бұл интервалдың максималды мәні оқиғаның пайда болғандығының толық сенімділігі ретінде түсіндіріледі, ал минималды мәні – ол оның толық жоқтығына сенімділік. Ықтималдықтар теориясына қарағанда сенімділік факторы оқиғаның пайда болатындығына субъективті сенімділікті сипаттайды және ешқандай статистикалық мағынасы жоқ. Семантикалық нейронды желілердегі нейрондар табиғи тілдің қарапайым түсініктеріне сәйкес келеді және дискретті градиентті мәндерді өңдейді. Осы желінің әрбір нейронында ақырлы күйлер саны болады. Сондықтан, семантикалық нейронды желі ақырлы автомат түрінде қарастырыла алады.

Морфологиялық және семантикалық талдау жасайтын семантикалық нейронды желінің құрылысы ретінде синхрондалған ағаш немесе графты таңдаймыз.

Мәнін синхрондалған сызықтық ағаш түрінде шығарып алу қабатын ақырлы автомат ретінде қарастыруға болады, себебі желідегі нейрондар саны шектеулі және оларда ақырлы күйлер мен байланыстар саны бар. Бір күйден екінші күйге өту мәнін шығарып алу қабатына кіріс тізбегінің кезекті символын беру кезінде болады. Мәнін шығарып алу қабатын бір автомат түрінде емес, бірнеше сөздік мақалалардың саны сияқты ақырлы субавтоматтар ретінде қарастырған ыңғайлы. Сонымен қатар, бір нейронда тыныштық күйден қозу күйіне дейінгі аралығында бір градиентті субкүйі бар деп есептеген ыңғайлы. Әрбір осындай субкүй қарапайым мағына болсын. Нейроавтоматтың бір белсенді субкүйіне бір немесе бірнеше қозған нейрон сәйкес келеді. Сонда бір синхрондалған сызықтық ағаштың бір үзіндісінде біз бірнеше субавтомат аламыз – сөздік мақала санына байланысты немесе бір мезетте бір автоматта бірнеше күй болады. Бұл шешім кейін табиғи тілдің көптеген есептерін шешуге көмегін тигізеді.

Сөздік мақаланың нейронының моделін қарастырып өтейік. Семантикалық нейронды желінің тізбектей есептеуіш жүйесінде жүзеге асырылуы нейронның жылдамдығына қосымша талаптар қояды. Мүмкіндігінше, бөлек нейронның жұмысының жылдамдығын арттырып, желідегі нейронның санын азайту керек, себебі бұл жағдайда нейрондар, бірінен кейін бірі тізбектей өңделеді, сондықтан жүйенің бір тактысының жалпы уақыт өлшемі өңделетін нейрондарды бір нейронды өңдеуге кететін уақытқа көбейткендегі санына тең. Нейрондардың санын азайту және олардың жұмысын арттыру үшін дизъюнктор мен конъюнкторды бір нейронға біріктіреміз. Бұл кезде әрбір нейронда екі дендритті ағаштан болады: біріншісі – кіріс градиентті мәннің дизъюнкциясының функциясын орындайтын, басқасы – кіріс градиентті мәннің конъюнкциясының функциясын орындайтын және дизъюнкция функциясының нәтижесін орындайтын. Нейронды дөңгелекпен белгілейміз, ал оның дизъюнкциясының дендритін осы дөңгелектің мол жағына қоямыз, конъюнктор дендритін дөңгелектің жоғары немесе төмен жағына орналастырамыз, аксон – дөңгектің оң жағында орналасады. Ыңғайлы болу үшін дөңгелектің ішіне конъюнктордың дендритіне сәйкес келетін символдарды жазамыз.

Сөздік мақаланың моделін қарастырайық. Бөлек сөздік мақала ол мақаланың негізгі мағыналық күші болатын негізгі сөз, және сөзтүрлену (септелу немесе жіктелу) арқылы

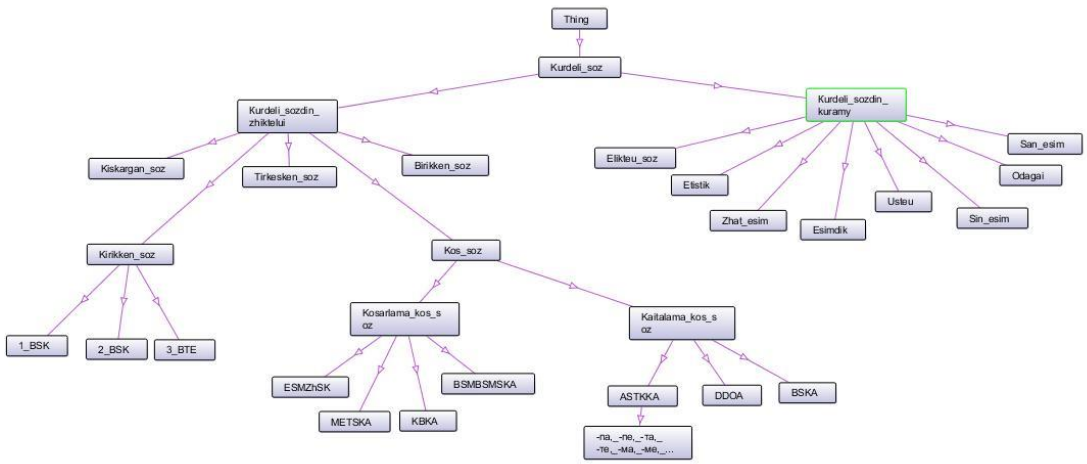
негізгі сөзден алынған сөз формасының тобы болады. Бір сөздік мақала – ол нейрондар тобы, немесе мәнін шығарып алу қабаттағы бір нейронды субавтомат. Сөздік мақаланың субкүйлерінің жалпы саны сол мақаланың сөзжасамдардың санына тең болсын. Осындай субавтоматтың әрбір субкүйі бір қозған нейрон болады. Бұл кезде, бір мезгілде бір субавтоматтың екі түрлі нейроны қозған жағдайда субавтоматта біруақытта екі түрлі субкүйі бар деп айтамыз. Әрбір мақалада сол мақалаға сәйкес келетін негізгі нейрон болады. Сөздік мақаланың негізгі нейроны оның сөздік мақаласына жататын сөз танылған кезде үнемі қозған күйде болады. Әрбір сөз формасына бөлек нейрон сәйкес келеді. Ол сөз формасы танылған кезде қозады.

Мәнін шығарып алу қабатта бөлек сөздік мақалаларға жатпайтын нейрондар болады. Бұл нейрондар көптеген сөздік мақалаларға тән септік, шақ, жіктеу сияқты сөзжасамдардың белгілеріне сәйкес келеді. Олар сәйкес белгілері бар сөз формалары қозғанда қозады. Сөзжасамдардың белгілеріне сәйкес келетін нейронның күйлері сол нейрондар байланысатын сөздік мақаланың субавтоматтарына жатады. Сонда, бірнеше сөздік мақалалар бір мезгілде дәл сол күйде болуы мүмкін.

Субавтоматтың қозған нейрондар жиыны субавтомат танитын бөлек нейронға жататын белгілер жиынына сәйкес келеді. Жіктеу немесе берілген символдық тізбек бойынша сөздік мақаланы және сөзжасамды анықтау есебі мәнін шығару қабаты арқылы қозу толқынының өтуіне және сәйкес сөздік мақала үшін сәйкесінше субавтоматтың қозуына алып келеді. Сөз түрлену есебі мұндай субавтоматтың бастапқы сөз түрлену басталатын сәйкес сөз формасының күйінен алғашқы сөзжасамына түрлендіру керек болатын сәйкес сөз формасының ақырғы күйіне өзгеруіне алып келеді.

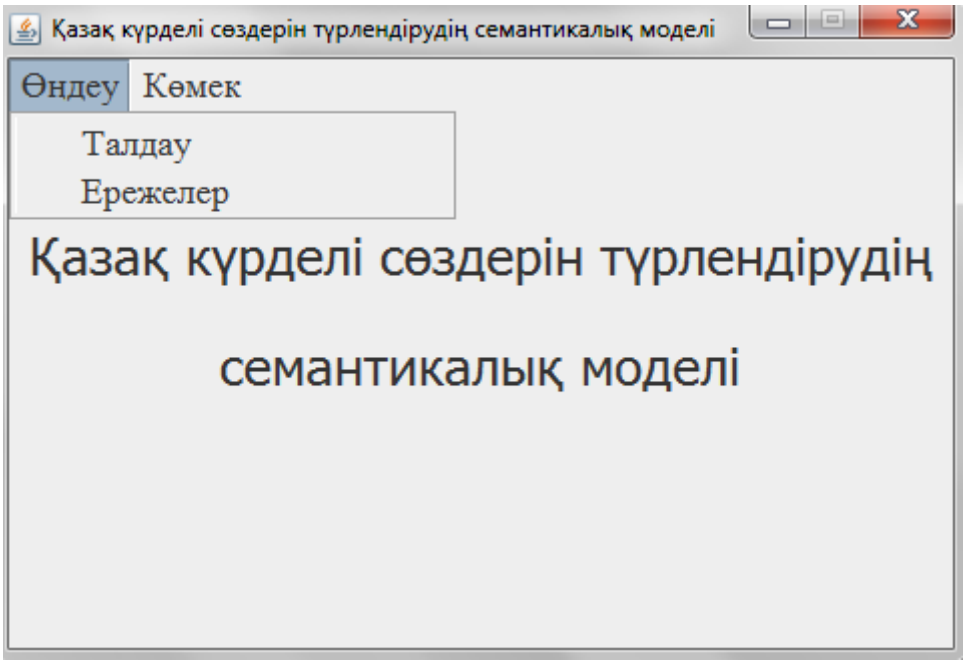
Сипатталған есептердің шешімін қамтамасыз ететін нейронды желілердің байланысының құрылысын қарастырайық. Синхрондалған сызықты ағаш сөздік мақала бойынша сөз формаларының жіктелу және сол сөз формасының белгілерін анықтау есептерінің шешімін қамтамасыз етеді. Егер көпмәнді жағдай туса синхрондалған сызықтық ағашта немесе графта сөз формасының барлық бөлек бөлек мәндеріне сәйкес келетін барлық сөздік мақалалар мен сөз формалары қозады.

Сөзтүрлену және сөзтудыру есептерін шешу үшін де синхрондалған сызықтық есепті қолдануға болады. Бұл жағдайда ол субавтоматты бір күйден екінші күйге ауыстыратын, қоздыруды тудыратын ауыстырғыш тізбек ретінде болады. Субавтоматтың күйлерінің ауысуы синхрондалған сызықтық ағаштың кірісіне арнайы командалар беру кезінде болады. Бұл командаларды синхрондалған сызықтық ағаш таниды және оларға сәйкес келетін нейрон-эффекторлардың шығысында градиентті мәнге түрленеді, бұл сөздік мақаланың күйлеріне сәйкес келетін нейрондардың қозуына немесе тежелуіне алып келеді.



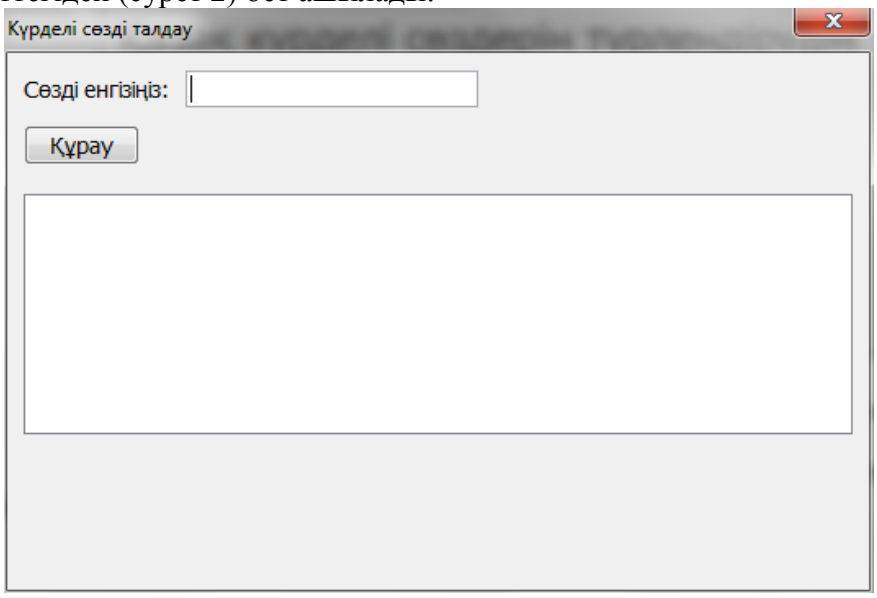
3 Қазақ күрделі сөздерінің семантикалық моделін программалық жүзеге асыру

Қазақ күрделі сөздерінің семантикалық моделін программалық жүзеге асыратын, оның ішінде қазақ күрделі сөздерін құруды және талдауды автоматтандыратын ақпараттық жүйесі JAVA тілінде Netbeans программалық ортасында жасалынған.



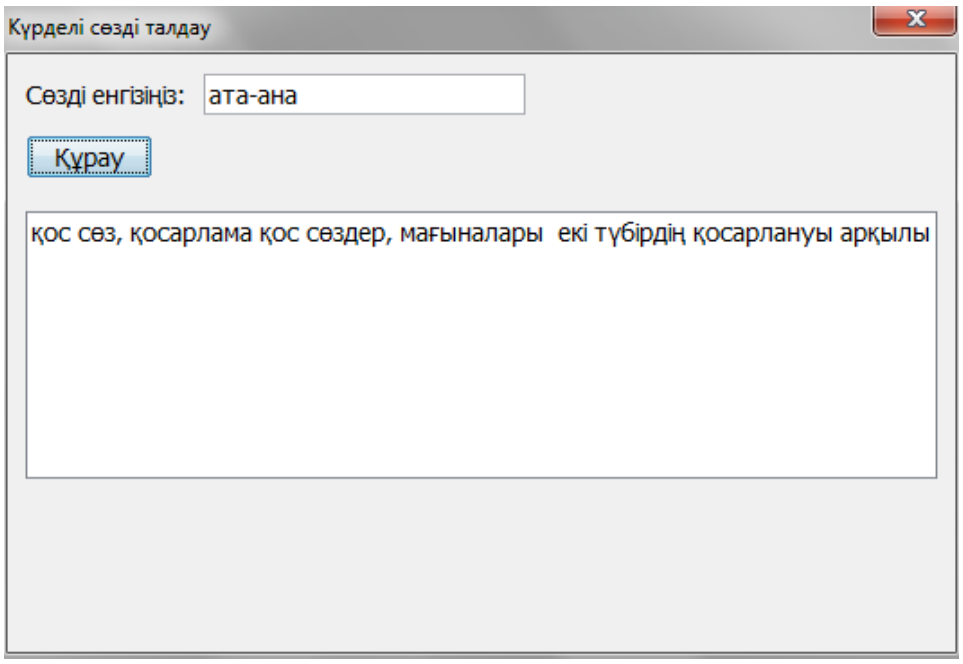
Сурет 1. Күрделі сөздерді талдау және ережелер
бойынша құру терезесі

Жұмысты жалғастыру үшін «Өңдеу» мәзірінен «Талдау» таңдаймыз. Таңдағаннан кейін төмендегі суреттегідей (сурет 2) бет ашылады.



Сурет 2. Күрделі сөздерді талдау терезесі

«Сөзді енгізіңіз» өрісіне күрделі сөзді енгіземіз, содан кейін құрау батырмасын бассақ, ол базадан сол күрделі сөзді тауып күрделі сөзге талдау жасайды. Мысалы төмендегі сурет 3 көрсетілген.



Сурет 3. Күрделі сөздерді талдау терезесі

Әдебиеттер

1 www.egemen.kz/?p=3486
2 Жанпейісов Е., Хұсайын К. және т. б. Қазақ грамматикасы. Фонетика, сөзжасам, морфология, синтаксис. – Астана: Астана, 2002. 152б.
3 Шәкенов Ж. Қазақ тіліндегі күрделі сөздер мен күрделі тұлғалар. – Алматы: Ана тілі, 1991, Б. 3-20
4 Аханов К., Б.Кәтенбаева , Әбдіғалиева Т. Қазақ тілі оқулығының методикалық нұсқауы. Алматы: Рауан, 1990, Б. 19-27

М. ЕРГЕШ

Л.Н.Гумилев атындығы Еуразия ұлттық университеті, Астана, Қазақстан

ҚҰЖАТТАРДАҒЫ КІЛТТІК СӨЗДЕРДІ ВЕКТОРЛЫҚ МОДЕЛЬ АРҚЫЛЫ ІЗДЕУ

Электронды түрдегі мәтіндік ақпараттар көлемі күн өткен сайын еселеніп көбейіп келеді. Сондықтан қазіргі таңда ақпараттық іздеу есептерін шешетін жүйелер жасау қажетілігі туып тұр. Қолданушының ақпараттық сұранысын қанағаттандыратындай құжаттар жиынынан іздеу процесі ақпараттық іздеудің классикалық мәселесі. Кілттік сөздерді анықтап, олардың маңыздылығын анықтау ақпараттық іздеу үшін қажетті мәселелердің бірі. Кілттік сөздің салмағы сөз формасының ақпараттылығын анықтайды және ол қолданушының сұранысына байланысты есепке алынады.

Ақпараттық іздеу әдістерінің белгілі бірнеше тәсілдері бар: бульдік модель, векторлық модель, ықтималдық модель. Бұл жұмыста құжаттардағы кілттік сөздерді табуға векторлық моделдің қолданылуын қарастырамыз. Қазақ тілді құжаттардағы кілттік сөздерді табу арқылы ақпараттық іздеу жүйелерінің қазақ тілді мәтіндерді іздеудің толықтығын және релевантылығын арттыруға болады.

Ақпараттық жүйелердің тиімдігінің басты белгісі 1960-шы жылдары енгізілген толықтық пен нақтылық. Іздеудің толықтығы берілген релеванттық құжаттырдың релеванттық құжаттардың жалпы санына қатынасы ретінде анықталған, ал іздеудің нақтылығы берілген релевантты құжаттардың шығарылған құжаттардың жалпы санына қатынасымен анықталады.

Векторлық модель – ақпараттық іздеуде құжаттар жиынын векторлық кеңістікте векторлармен сипаттау.

Векторлық моделде құжаттар реттелмеген термдер жиыны ретінде қарастырылады. Ақпараттық іздеуде *термдер* деп мәтіннің сөздері мен элементтері аталады, мысалы: кітап, ақпарат, 2010.

Құжаттағы термдердің салмағын түрлі тәсілдермен анықтауға болады - берілген мәтін үшін сөздің «маңыздылығы». Мысалы, термнің жиілігі (tf) деп аталатын құжаттағы термнің қолданылу санын жай ғана есептеуге болады, яғни құжатта сөз көбірек кездескен сайын сөздің салмағы да үлкен болады. Сәйкесінше, құжатта терм кездеспесе, сол құжаттағы салмағы нөлге тең болады.

Өңделіп жатқан жиындағы құжаттарда кездесетін барлық термдерді реттеуге болады. Егер кейбір құжат үшін ретімен салмағы бойынша барлық термдерді кездеспесе де жазып шығу керек.

Сол вектор құжаттың векторлық кеңістіктегі көрінісі болады. Вектордың өлшемі кеңістіктің өлшемі сияқты, барлық жиындағы түрлі термдердің санына тең болады және барлық құжаттар үшін бірдей болады.

Құжаттың векторлық көрінісі

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$$

мұнда d_j —*j*-шы құжаттың векторлық көрінісі, w_{ij} —*i*-ші термнің *j*-ші құжаттағы салмағы, *n* —жиындағы құжаттардағы түрлі термдердің жалпы саны.

Құжаттардың осындай векторлық көрінісі арқылы кеңістіктегі нүтелердің арасындағы ара қашықтықты тауып, құжаттардық ұқсастығын табу мәселесін шешуге болады. Нүктелер жиі орналасқан сайын, сәйкесінше құжаттар ұқсас болады.

Құжаттағы сөздің маңыздылығын анықтаудың қарапайым тәсілі құжаттағы сөздің қолдану жиілігін анықтау.

Жиындағы сөздің қолдану жиілігі сол сөз бар құжаттардың жиындағы санын (df) есептеу арқылы шығаруға болады. df өскен сайын құжаттағы сөздің салмағы төмендей береді. Оны tf құжаттағы сөздің қолдану жиілігін оның кері шамасы idf-қа көбейту арқылы шығаруға болады. Сөйтіп құжаттағы сөздің салмағы tf*idf формуласымен есептеледі. idf төмендегі формула арқылы есептеледі:

$$\mathrm{idf}_{tj} = \log(N/n_j)$$

мұнда, N – жиындағы құжаттар саны, n_j - t_j кездескен құжаттар саны.

Сонымен, $D = (d_1,...,d_n)$ –жиындағы құжаттар жиыны, $T = (t_1,..,t_M)$ – сөздер жиыны. Әрбір тұрақты i үшін d_i құжаты төмендегі салмақ векторы арқылы сипатталады:

$W_{ij} = tf_{ji*}\ idf_{ji} = 1 \ldots M,$

мұнда tf_{ji} - dj құжатындағы t_j сөзінің кездесі жиілігі, idf_{ji} – барлық құжаттардағы t сөзінің кездесу жиілігіне кері шама.

Құжаттағы барлық сөздердің салмағын есептегеннен кейін құжатты вектор ретінде көрсетеміз, ондағы әрбір компонент құжаттағы бөлек сөздерге сәйкес келеді. Құжаттарды ондағы сөздердің векторы түрінде көрсету ақпараттық іздеудің векторлық моделінің негізі болып табылады.

Ақпараттық іздеудің векторлық моделінің артықшылығы реттелген ақпараттық жүйені жасау үшін қарапайым модель береді. Сонымен қатар, шешіліп жатқан мәселеге және жұмыс

жиынына байланысты құжаттағы сөздер салмағын есептеудің тәсілдері өзгере беуі мүмкін. Мәтіндегі сөздердің бір біріне тәуелді болмайды деп қарастыру векторлық моделдің кемшілігі болып табылады, себебі мәтіндегі сөздер бір бірімен мағына қатысты байлынысып тұрады.

Әдебиеттер

1. Daniel Jurafsky, James H. Martin Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Pearson Education International. - 2009. - 1024 pp.

2. Peter D.Turney, Patrick Pantel. From frequency to meaning: vector space models of semantics // Journal of artificial intelligence research 37 (2010) 141-188

3. А.А.Мамчич. Алгоритмы индексирования и поиска документов на основе динамических корпусов текстов // информатика. - 2010. № 1.

М.Х. ХАКИМОВ, М.М.АРИПОВ

Национальный Университет Узбекистана им. Мирзо Улугбека,
г. Ташкент, Республика Узбекистан

СЕМАНТИЧЕСКИЕ БАЗЫ РУССКОГО ЯЗЫКА

Многоязычная ситуация компьютерного перевода текста [1] требует разработку семантических баз данных и математических моделей естественных языков [3]. Предшествующим этапом построению математических моделей русского языка (РЯ) явились исследования лексического, синтаксического, семантического анализа и построения на их основе логико-лингвистических моделей. При описании семантических баз данных и математического моделирования РЯ используем элементы расширенного входного языка множества Λ терминальных символов [2].

В результате семантического анализа когда основой слова является существительное было выведено 16 вида приставок (табл. 1) формирующих семантическую базу приставок существительного - C(T), 120 вида суффиксов (табл. 2) формирующих семантическую базу суффиксов существительного - C(S) и 28 вида окончаний (табл. 3) формирующих семантическую базу окончаний существительного - C(O):

Таблица 1.

без-	за-	на-	о-	пере-	под-	при-	с-
бес-	между-	не-	об-	по-	пред-	про-	со-

Таблица 2

-ак	-ак-а	-ан	-ани-е	-анин	-ар	-овщин-а	-ш-
-арь	-аци-я	-ач	-бищ-е	-в-а	-ек	-ичеств-	-ан-
-емость	-енец	-ени-е	-енк-а	-енок	-еныш	-ят-	-иц-
-еньк-а	-есть	-ец	-ец-о	-ечк-а	-ечк-о	-ч-	-щиц-
-и-е	-изм	-изн-а	-ик	-имость	-ин-	-и-	-ниц-
-ин-а	-ист	-итель	-иц-а	-ич	-ичк-а	-овств-	-ух-
-ишк-а	-ишк-о	-ищ-а	-ищ-е	-к-а	-к-о	-еств-	-их-
-л-а	-лец	-лиц-е	-лк-а	-льник	-льн-я	-ни-	-к-
-льщик	-ник	-ниц-а	-ность	-н-я	-няк	-ани-	-янк-
-овец	-ович	-овщик	-ок	-онк-а	-онок	-ун-	-енк-
-оньк-а	-ость	-от-а	-отн-я	-очек	-очк-о	-ств-	-ян-
-ств-о	-тель	-ти-е	-ул-я	-ун	-ура	-ени-	-анк-

-ус-я	-ушк-а	-ушк-о	-ц-а	-ц-е	-ц-о	-ти-	-льщиц-
-честв	-чик	-щик	-щин-а	-ыш	-ышк-о	-еч-	-чиц-
-юшк-а	-яг-а	-як	-як-а	-ян	-янин	-тельств-	-й-

Таблица 3.

-а	-е	-и	-й	-ом	-ьев	-я
-ам	-ев	-ие	-о	-у	-ью	-ям
-ами	-ей	-ий	-ов	-ы	-ья	-ями
-ах	-ем	-ия	-ой	-ь	-ю	-ях

В результате семантического анализа когда основой слова является прилагательное было выведено 25 вида приставок (табл. 4) формирующих семантическую базу приставок прилагательного - P(T), 65 вида суффиксов (табл. 5) формирующих семантическую базу суффиксов прилагательного - P(S), 42 вида окончаний (табл. 6) формирующих семантическую базу окончаний прилагательного - P(O):

Таблица 4.

анти-	без-	вне-	внутр-	до-
за-	интер-	между-	на-	над-
наи-	не-	небез-	небес-	по-
под-	после-	пре-	пред-	при-
про-	противо-	раз-	сверх-	ультра-

Таблица 5.

-ав-	-айш-	-ан-	-аст-	-ат-	-ач-	-ащ-	-е	-ебн-
-еват-	-ее-	-ей	-ейш-	-енн-	-еньк-	-ёхоньк-	-ешеньк-	-ив-
-ик-	-им-	-ин-	-инск-	-инск-	-ист-	-ит-	-ич-	-ическ-
-ическ-	-ичн-	-й-	-к-	-л-	-лив-	-льн-	-ляв-	-н-
-ов-	-ов-	-оват-	-овит-	-овн-	-овск-	-овск-	-онн-	-оньк-
-охоньк-	-ошеньк	-ск-	-тельн-	-уч-	-ущ-	-ущ-	-чат-	-ческ-
-чив-	-ше	-ък-	-ьн-	-юч-	-ющ-	-ющ-	-яв-	-ян-
-яч-	-ящ-							

Таблица 6.

-а	-ая	-е	-ё	-его	-ее	-её
-ей	-ем	-ём	-еми	-ему	-ех	-ею
-и	-ие	-ий	-им	-ими	-их	-й
-о	-ого	-ое	-ой	-ом	-ому	-у
-ую	-ы	-ые	-ый	-ым	-ыми	-ых
-ье	-ьи	-ья	-ю	-юю	-я	-яя

В результате семантического анализа для основы слова типа глагола было выведено 30 вида приставок (табл. 7) формирующих семантическую базу приставок глагола - G(T), 43 вида суффиксов (табл. 8) формирующих семантическую базу суффиксов глагола - G(S), 37 вида окончаний (табл. 9) формирующих семантическую базу окончаний глагола - G(O):

Таблица 7

в-	-в-	вз-	взо-	-во-	вс-
вы-	вы-	-вык-	до-	за-	из-
-каз-	-лож-	на-	над-	-ня-	о-
обез-	обес-	от-	пере-	по-	под-
при-	про-	раз-	рас-	с-	у-

Таблица 8

-а-	-ач-	-ащ-	-в	-ва-	-вш-	-вши	-е-
-ева-	-ем-	-енн-	-ере-	-и	-и-	-ива-	-изирова-
-им-	-ирова-	-ича-	-ка-	-л-	-л-ый	-нича-	-нн-
-ну-	-ова-	-оло-	-ом-	-ствова-	-ти	-ть	-уч-
-учи-	-ущ-	-чь	-ши	-ыва-	-юч-	-ючи	-ющ-
-я-	-яч-	-ящ-					

Таблица 9

-а	-ат	-ать	-ая	-ее	-ем	-ет	-ете
-еть	-ешь	-ёшь	-и	-ие	-им	-ит	-ит
-ите	-и-те	-ить	-ишь	-о	-сь	-ся	-ти
-ть	-у	-ут	-чь	-ые	-ь	-ь-те	-ью
-ю	-ют	-я	-ят	-яя			

В результате семантического анализа для основы слова типа местоимения было выведено 35 вида окончаний (табл. 10) формирующих семантическую базу окончаний местоимения - M(O):

Таблица 10

-а	-ая	-е	-ё	-его	-её	-ей
-ем	-ём	-еми	-ех	-и	-ие	-ий
-им	-ими	-их	-й	-ого	-ое	-оё
-ой	-ом	-ому	-у	-ую	-ые	-ый
-ым	-ыми	-ых	-ью	-ю	-юю	-я

В результате семантического анализа для основы слова типа наречия было выведено 21 вида приставок (табл. 11) формирующих семантическую базу приставок наречия - N(T), 13 вида суффиксов (табл. 12) формирующих семантическую базу суффиксов наречия - N(S), 21 вида окончаний (табл. 13) формирующих семантическую базу окончаний наречия - N(O):

Таблица 11

в-	д-	до-	е-	еже-	за-	и-
из-	к-	ка-	на-	не-	о-	об-
по-	про-	с-	через-	черес-	чрез-	чрес-

Таблица 12

-е	-ему	-енечк-	-еньк-	-жды
-и	-мя	-о	-оват-	-ому
-онечк-	-оньк-	-у		

Таблица 13

-а	-е	-ё	-ем	-ём	-ех	-ею
-и	-им	-их	-й	-о	-ом	-у
-ую	-ы	-ым	-ых	-ю	-юю	-я

В результате семантического анализа с основанием числительных было выведено 43 вида суффиксов (табл. 14) формирующих семантическую базу суффиксов числительных - F(S):

Таблица 14

-а	-ами	-ая	-дцать	-е	-ей	-ем	-емя
-емя	-еро	-ёх	-и	-им	-ими	-их	-мя
-надцать	-о	-ого	-ое	-ой	-ом	-ому	-сот
-ста	-стам	-стами	-стах	-сти	-у	-ум	-умя
-ух	-ы	-ые	-ый	-ым	-ых	-ьмя	-ью
-ю	-ям	-ями					

В результате семантического анализа грамматики РЯ было выявлено 48 типа предлогов (табл.15) формирующих семантическую базу предлогов D, 83 вида союза (табл.16) формирующих семантическую базу союзов Y, 85 типов частиц (табл.17) формирующих семантическую базу частиц U, 84 типа междуметий (табл.18) формирующих семантическую базу междуметий E, 50 вида модальных слов (табл.19) формирующих семантическую базу модальных слов L, 8 вида постфиксов (табл.20) формирующих семантическую базу постфиксов B и два вида морфем (-о-, -е-) формирующих семантическую базу морфем W:

Таблица 15

без	близ	в	вдоль	вне	внутри
возле	вокруг	впереди	для	до	за
из	из-за	из-под	к	кроме	кругом
между	мимо	на	над	напротив	о
об	обо	около	от	относительно	оть
перед	по	под	подле	поперек	после
пред	прежде	при	про	ради	с
сзади	спустя	сь	у	через	чрез

Таблица 16

а	а то	благодаря тому что	будто	в то время как	ввиду того	вследствие того что
где	да	дабы	для того чтобы	едва	ежели	если
если – то	затем	и	и – и	и да	ибо	или
или – или	итак	к тому же	как	как	как – так и	как – то
как будто	какой	когда	коли	который	кто	куда
либо	либо – либо	лишь	лишь бы	лишь только– как	наконец	напротив
не то – не то	не только – но и	несмотря на то	ни	ни – ни	но	но
но и	однако	откуда	оттого	после того как	потом	потому что
правда	прежде того как	прежде чем	пускай	пусть	раз	с тем чтобы
с тех пор как	словно	так и	так как	так как – то	так только	так что
также	то – то	то есть	тогда – так	тоже	только что – как	точно
хоть	хотя	чей	что	что	что бы	

Таблица 17

а	а ну	б	бишь	будто	бы	ведь	вишь
вон	вон и	вот	вот и	вот как	вот так	все	все же
все таки	всего	да	да и	давай	давайте	даже	де
едва	единственно	еще бы	ж	же	и	именно	исключительно
ишь как	ишь какой	как	как будто	как раз	кое	-кое	кое-
-либо	-либо	лишь	лишь только	мол	-на	не	не
нет	неужели	ни	нибудь	-нибудь	ну	ну и	оно
отнюдь	почти	просто	прямо	пускай	пусть	равно	разве
разве	ровно	словно	-сь	-ся	-таки	-те	то
-то	-то	только	точно	точно	уж	хотя бы	что
что за	что ли	чуть не	это	якобы			

Таблица 18

а	а ну тебя	ага	алло	ась	ату	ах
аха	баста	благодарю	боже мой!	брысь	бух	виноват
вон	вот еще!	всего	всех благ	га – га –га	глупости!	да

		хорошего				
динь–динь–динь	до свидания	добрый день	здравствуйте	извините	извиняюсь	к чёрту!
-ка	кхе – кхе –кхе	кши	марш	мах	мерси	миг – миг
на – ка	на – те	на – те – ка	нет	ну – ка	Ну – ну!	ну – те
ну – те – ка	Ну!	о	ой	ой ли	ох	ох
пардон	право	простите	прочь	прыг	ррр… нга – нга	спасибо
стоп	так – так	-те	тик – так	толк	тпру	Тр – тр
трах	Трра!	Тррах! Та, тах!	тс	тьфу	увы	угу
уж	ура	уф	фи	фу	фьюить	ха – ха – ха
хи – хи – хи	хлоп	цып!	цыц	шш	щелк	эх

Таблица 19

в частности	верно	вероятно	видать	видимо	видно	вне всякого
во-вторых	возможно	во-первых	действительно	дело	добро	должно быть
думается	желательно	желать	значит	известно	итак	к несчастью
к радости	к счастью	к удивлению	кажется	как будто	конечно	может быть
мочь	наверное	наверняка	надо	наконец	например	необходимо
необходимо	несомненно	нужно	нужно	признаться	разумеется	самом деле
следовательно	слышно	сомнения	стало быть	факт	хотеть	шутка
шутка сказать						

Таблица 20

-ся	-сь	-то	-либо	-нибудь	-таки	-ка	-те

Вышеизложенные семантические базы данных применяются математических моделях вывода слов и предложений по типам.

Литература

1. Хакимов М.Х. Формальные системы машинного перевода в многоязычной ситуации. Материалы республиканской научной конференции «Современные проблемы математики, механики и информационных технологий», НУУз, Институт Математики и ИТ АН РУз, Т, 2008, с.297-301

2. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. ЎзМУ хабарлари, № 1, 2009, с.75-80.

3. Хакимов М.Х. Математические модели узбекского языка. ЎзМУ хабарлари, № 3, 2010, с.187-191.

МАШИНАЛЫҚ АУДАРУ ЖҮЙЕЛЕРІ
СИСТЕМЫ МАШИННОГО ПЕРЕВОДА
MACHINE TRANSLATION SYSTEMS

Д.Ш. СУЛЕЙМАНОВ, А.Р. ГАТИАТУЛЛИН, Р.А. ГИЛЬМУЛЛИН, М.М. АЮПОВ

НИИ "Прикладная Семиотика" Академии наук Республики Татарстан

К РАЗРАБОТКЕ ТАТАРСКО-ТУРЕЦКОГО МАШИННОГО ПЕРЕВОДЧИКА[6]

Введение

В настоящее время большинство систем машинного перевода, особенно для языков индо-европейской группы, основано на статистическом подходе. Это объясняется, как рядом очевидных преимуществ такого подхода, таких как, возможность «самообучения», гладкость перевода, переносимость технологии на любые языковые пары, так и, главным образом, наличием достаточного количества параллельных корпусов. Собственно, наличие паралельных корпусов является важным и непременным условием для эффективного применения статистического метода перевода. Ситуация для языков тюркской группы в настоящее время совершенно иная - для большинства языков этой группы практически не имеется параллельных национальных корпусов. Как показывает анализ электронных корпусов языков в сети Интернет, только электронные параллельные корпуса для турецкого и уйгурского языков (уйгурско-китайский параллельный корпус) обладают достаточным объемом параллельных текстов, которые могут быть использованы для создания машинных переводчиков на основе статистического подхода.

С середины 90-х годов началась активная работа по созданию машинных переводчиков для тюркских языков. В частности, в Интернете сегодня доступен целый ряд таких переводчиков - русско-узбекский (www.spells.uz), русско-казахский, казахско-русский (www.sanasoft.kz), азербайджано-английский, азербайджано-турецкий (www.dilmanc.az), уйгурско-китайский (www.jofcis.com/downloadpaper.aspx?), уйгурско-японский [Muhtar M., 1994] и турецко-крымско-татарский [Altıntas, 2000] переводчики. Вместе с тем, в списке языков перевода, осуществляемых системой Google, из тюркских языков представлены только турецкий и азербайджанский языки.

В двух из этих систем машинного перевода производится перевод для близкородственных языков: азербайджано-турецкий [Fatullayev, 2008] и турецко-крымско-татарский [Altıntas, 2000]. В обоих проектах используется RBMT (Rule Based Machine Translation) подход, где для решения задач морфологического анализа и синтеза словоформ тюркских языков использованы фонологические и морфотактические правила автоматического морфологического анализа в двухуровневой модели морфологии, реализованной в системе PC KIMMO. PC KIMMO – это компьютерная программа, которая использует лингвистическое описание фонологии и морфологии естественного языка и специальным образом размеченный словарь (Лексикон) для распознавания и генерации слов на этом языке. Использование RBMT подхода, скорее всего, объясняется тем, что статистический подход плохо справляется с анализом агглютинативных конструкций морфологии тюркских языков.

Хотя RBMT подход тоже имеет свои слабые стороны, среди которых можно отметить трудоемкость и длительность разработки, а также необходимость постоянно поддерживать и актуализировать лингвистические базы данных, для реализации систем татарско-турецкого и татарско-казахского машинного перевода в НИИ Прикладная семиотика был выбран именно

[6] Исследование выполнено в рамках научно-исследовательского проекта РФФИ («Математические модели, методы, технологии и системы обработки многоязыковых текстов тюркских языков для задач машинного перевода»), проект № 12-07-97015

RBMT метод, в первую очередь, в силу отсутствия базы параллельных текстов, а также исходя из желания добиться наибольшей точности перевода.

Система перевода для близкородственных языков строится на основе прагматически-ориентированного подхода к разработке лингвистических моделей [Сулейманов, 1998]. Прагматически-ориентированный подход позволяет более детально прорабатывать модели определенного языкового уровня в зависимости от целевой ориентированности разрабатываемой системы и определять минимальный набор средств для решения определенного круга лингвистических задач. Эффективность системы перевода, разрабатываемая на основе этого подхода, может быть обеспечена на уровне формирования лингвистических моделей разного уровня, за счет учета близости структурных и типовых характеристик языков внутри одной языковой группы.

Языки внутри одной тюркской языковой группы, в число которых входят татарский, казахский и турецкий, обладают большим сходством на всех языковых уровнях. Поэтому нами выдвинута гипотеза, что при разработке систем перевода внутри тюркских языков основную часть перевода будут обеспечивать лингвистические модели морфологического и морфо-синтаксического уровней. Исходя из этой гипотезы и в соответствии с прагматически-ориентированным подходом разработана общая архитектура системы машинного перевода (рис.1).

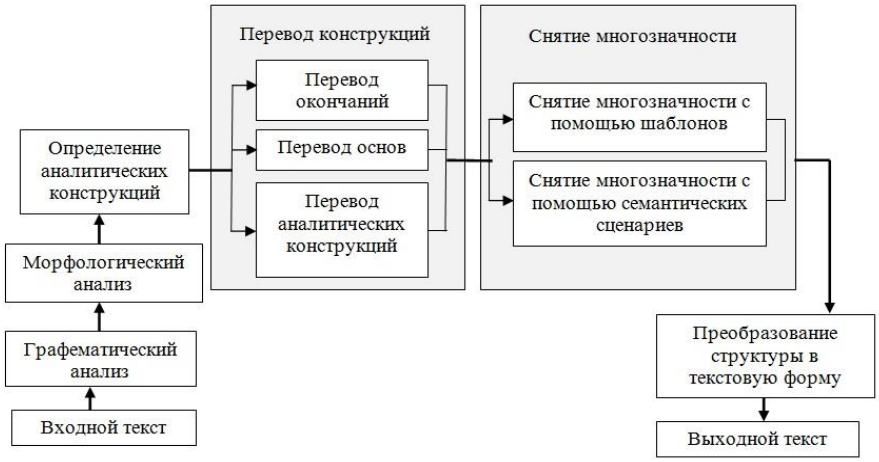


Рис.1. Общая архитектура работы СМП для близкородственных языков

Как правило, благодаря практической идентичности синтаксической структуры предложений, при переводе между близкородственными языками имеющиеся неоднозначности в исходном тексте в том же виде переходят в переводной текст на другом языке. Такая же ситуация со словоформами, в которых имеет место совпадение многозначности в корневых и аффиксальных морфемах. На рис.2 приведен пример перевода словоформы с казахского языка на татарский, при котором в результате перевода неоднозначность, изначально содержащаяся в казахской словоформе, сохраняется и в татарском.

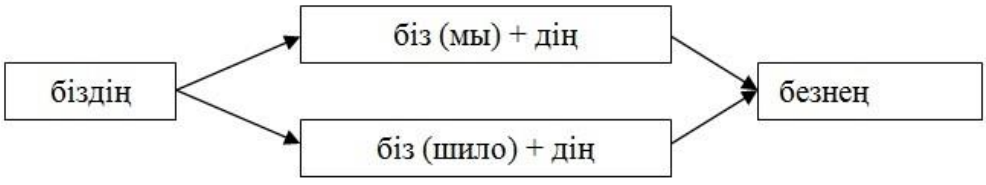


Рис 2. Совпадающие варианты перевода

Тем не менее, на уровне слов, словосочетаний, а также на уровне соответствия аффиксальных морфем, а в ряде случаев и на уровне синтаксических структур, возникают несоотвествия в интерпретации их значений, и тем самым, возникают нетождественные неоднозначности, что требует создания соотвествующих схем их разрешения.

Такая система, реализующая схемы разрешения, и показанная на рис.1 – система перевода конструкций с одного языка на другой, осуществляет перевод на трех уровнях в зависимости от результатов анализа: перевод морфемы в морфему, словоформы в словоформу, многословной конструкции в многословную конструкцию.

В том случае, когда применение простых конструкций не дает однозначного перевода, осуществляется перевод с помощью более сложных языковых конструкций. Такая ситуация может возникнуть в случае несовпадения порядка следования морфем в словоформе. В качестве примера приведем перевод словоформ татарского и турецкого языков, когда в турецком языке морфема персональности следует справа от модальной вопросительной морфемы, а в татарском наоборот:

тур.: *Ben biliyor **mu**-yum? ‘Я знаю?’*

тат.: *Мин белә-м-**ме**? ‘Я знаю?’*

В этом случае перевод морфемы в морфему невозможен.

2. Лингвистические ресурсы

Для решения задачи создания переводчика требуется большое количество лингвистических ресурсов, соответственно, нами был произведен анализ имеющегося на сегодня материала, готового для использования в проекте. Анализ показал, для использования в задачах перевода между татарским и другими тюркскими языками в настоящее время реально доступны только татарско-турецкий и турецко-татарский словари объемом около 20 000 словарных статей. Вместе с тем, даже эти словари изданные в бумажном виде, не представлены в Интернете в электронном виде. Из этого следует, что необходимо активизировать работу по созданию многоязычного словаря тюркских языков.

Авторами предложена структура многоязычного словаря, в соответствии с которой словарь должен содержать в себе не только лексическую, но и морфологическую информацию (рис.3).

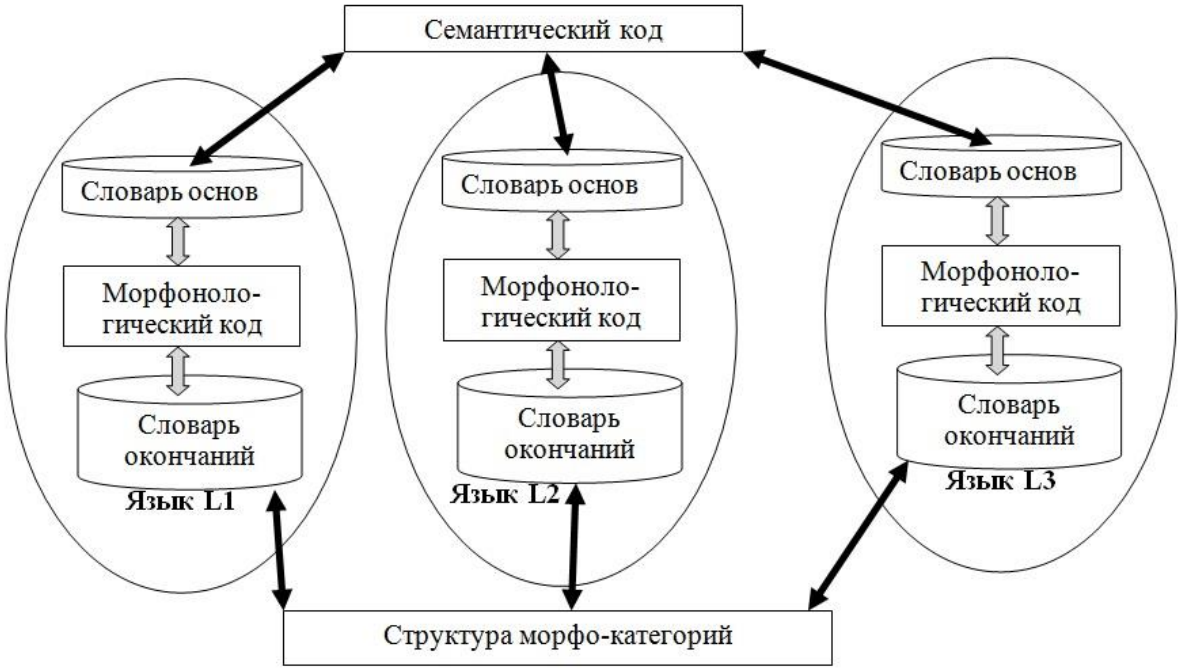


Рис.3. Архитектура базы данных с многоязычным словарем

Согласно этой модели база данных состоит из N взаимосвязанных компонент для каждого из используемых языков. Каждая из компонент имеет независимую структуру, обусловленную языковой спецификой. Компоненты взаимосвязаны между собой на уровне лексических эквивалентов (семантический код) и морфологических категорий.

Компонента для каждого языка содержит как лексическую, так и морфологическую информацию. Это сделано с целью оптимизации поиска в базе данных, чтобы в одном словаре базы данных производить поиск требуемой компоненты только один раз, одновременно вытаскивая из него как морфологическую информацию, так и информацию о лексических эквивалентах найденной лексемы.

Структуру словаря основ для одного языка можно видеть в таблице 1, где представлен фрагмент заполнения словаря основ для татарского языка.

Таблица 1

МТ	**Сем.код**	**Словарная основа**	**Неизменяемая часть**	**МФТ**
N	734	Арлекин (рус.: Арлекин)	арлекин	7
N	734	Шамакай (Грязнуля)	шамака	15
N	734	мәзәкче (юморист)	мәзәкче	2
N	734	балаганчы (балаганщик)	балаганчы	1
N	734	кәмитче (карусельщик)	кәмитче	2
N	734	мәсхәрәче (юморной)	мәсхәрәче	2

Рассмотрим содержание столбцов таблицы словаря основ.

По морфонологическим правилам татарского языка, после присоединения к основе некоторых алломорфов, в самой основе могут происходить определенные изменения, как правило, это чередования или пропуск отдельных букв. Для того, чтобы отразить эти изменения в словарях, в словаре основ содержится два варианта основ: Словарная форма, Наибольшая неизменяемая форма. Сами возможные изменения представлены в словаре окончаний.

Все основы в словаре основ представляются в безаффиксальной форме, для глаголов это форма повелительного наклонения.

Связь лексем в разных языковых компонентах словаря осуществляется с помощью семантического кода, который содержится в поле Семантический код. Слова с одним и тем же значением имеют один и тот же семантический код. Так, на примере в таблице 1 приведен список лексем с одинаковым семантическим кодом. Это, с одной стороны, позволяет в поисковых системах искать не только тексты с заданными словами, но и его синонимами, с другой стороны, невозможно всегда однозначно определить вариант перевода, поэтому на этом этапе перевода выдаются все возможные варианты, а наиболее вероятный вариант перевода уже будет выбираться на этапах снятия неоднозначности, используя соответствующие синтаксические и семантические механизмы.

В таблицах 2-3 (Таблица 2 – казахский язык, Таблица 3 – турецкий язык) приведены фрагменты словарей основ тюркских языков с тем же самым семантическим кодом, что и во фрагменте татарского словаря. Из этих примеров видно, что словари для всех тюркских языков имеют аналогичную структуру. Для каждого языка количество синонимичных основ может быть разное.

Таблица 2

МТ	Код	Словарная основа	Неизменяемая часть	МФТ
N	734	арлекин	арлекин	9
N	734	қуақы	қуақы	1
N	734	масқарапаз	масқарапаз	3

N	734	қалжыңқой	қалжыңқой	5
N	734	қылжақпас	қылжақпас	7

Таблица 3

МТ	Код	Словарная основа	Неизменяемая часть	МФТ
N	734	soytarı	soytarı	1
N	734	arleken	arleken	7

Следующим лингвистическим ресурсом являются словари окончаний для каждого из этих языков. Процесс создания словарей окончаний состоит из двух этапов:

1. Сравнительный анализ аффиксальных морфем, используемых тюркских языков;
2. Определение морфонологических типов.

Сравнительный анализ морфологических категорий татарского, казахского и турецкого языков показал, что для именных конструкций системы морфологических категорий во всех этих языках очень близки и основная масса морфологических категорий и выражающих их аффиксальных морфем полностью совпадают. Между ними существует только морфонологическая разница, как например количество алломорфов аффиксальных морфем. Есть только небольшие различия, как например наличие инструментального падежа в казахском и турецком языках, выражаемого с помощью аффикса, а в татарском аналогичная категория выражается с помощью послелога.

Основное различие между морфологическими категориями татарского, казахского и турецкого языков заключается в глагольных категориях. Результаты этого анализа будут отражены в отдельной публикации.

3. Реализация системы в виде Web-ресурса

Система перевода для близкородственных языков реализуется в виде отдельного веб-ресурса, рабочий вариант которого доступен в сети Интернет по адресу: http://sem.antat.ru/.

Рассмотрим пользовательский интерфейс ресурса на данном этапе. Главная страница состоит из 3 вкладок (рис.4):

– «Перевод словоформы»,
– «Перевод предложений»,
– «Перевод текста из файла».

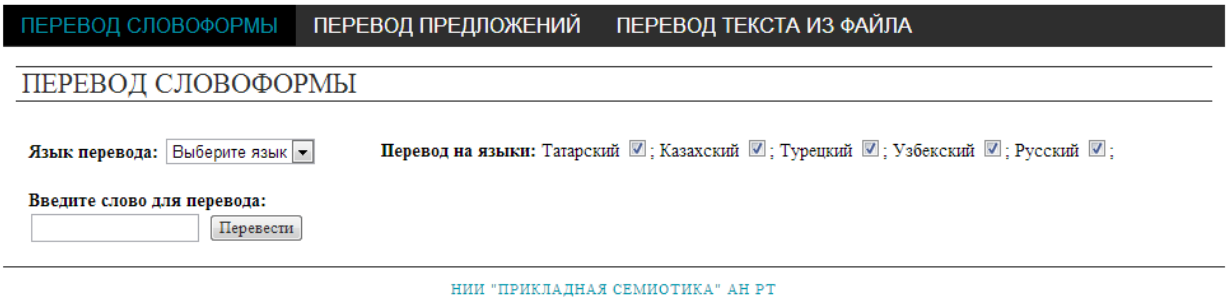


Рис. 4. Главная страница сайта

На вкладке *«Перевод словоформы»* (рис.4) сначала из списка *«Язык перевода»* выбирается язык исходной словоформы. Исходная словоформа может быть на татарском, казахском или турецком языке. Затем в поле *«Перевод на языки»* необходимо отметить языки, на которые следует перевести исходную словоформу. В поле *«Введите слово для перевода»* необходимо ввести словоформу для перевода, после чего нажать на кнопку *«Перевести»*. Для примера

возьмем словоформу на татарском языке *«өйдәге»* (рус.: *то, что в доме)* и выполним его перевод. Результат перевода показан на рис.5.

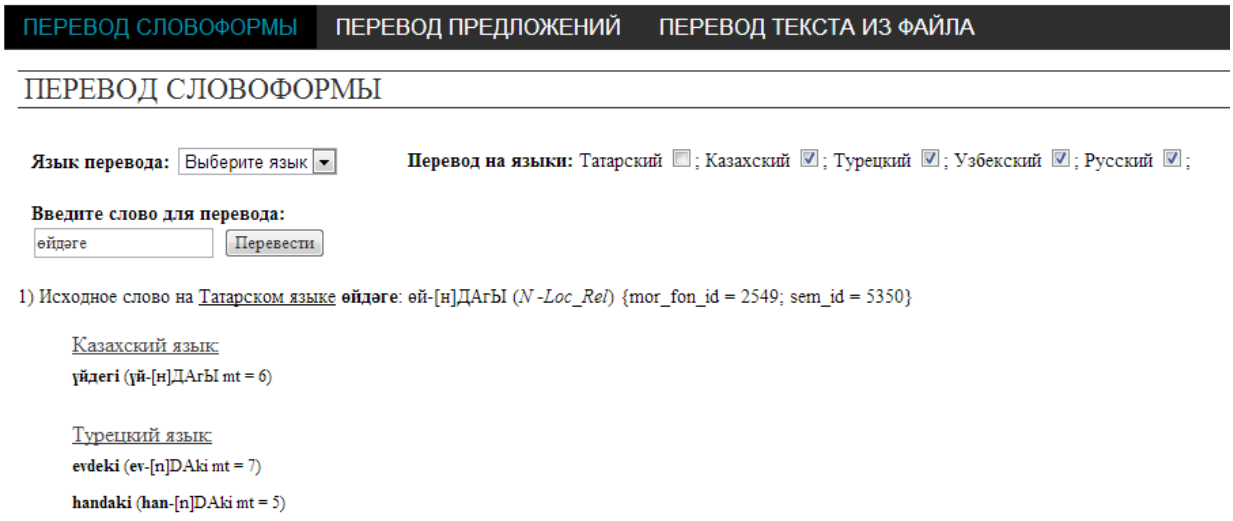


Рис. 5. Перевод словоформы

На вкладке *«Перевод предложений»* (рис.6) можно перевести целое предложение. Для этого сначала из списка *«Язык перевода»* выбирается язык исходного текста. Затем в поле *«Перевод на языки»* необходимо отметить языки, на которые следует перевести набранный текст. В поле *«Введите текст для перевода»* вводится текст, и нажимается кнопка *«Перевести»*.

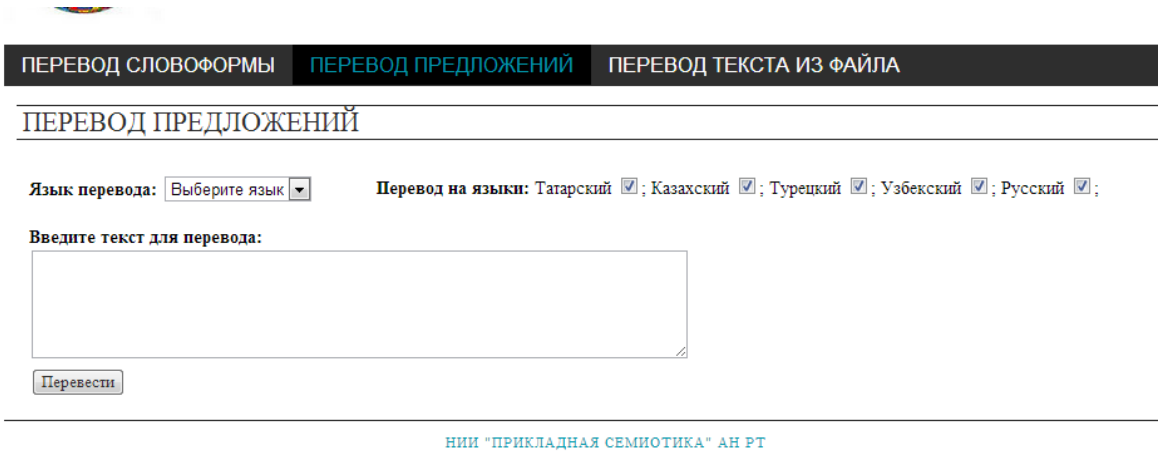


Рис. 6. Перевод предложений

Для примера возьмем предложение на татарском языке *«Абый авылдагы дусларына кайтты»* (рус.: *Старший брат приехал к друзьям в деревню*) и выполним его перевод. Результат показан на рис.7.

Как видно из рис.7, на сегодняшний день реализация программы находится на начальном этапе и в ней предстоит еще реализовать целый ряд планируемых модулей, в частности, механизмы снятия возникающих многозначностей, и поэтому на каждом из целевых языков получается целый набор переводных вариантов.

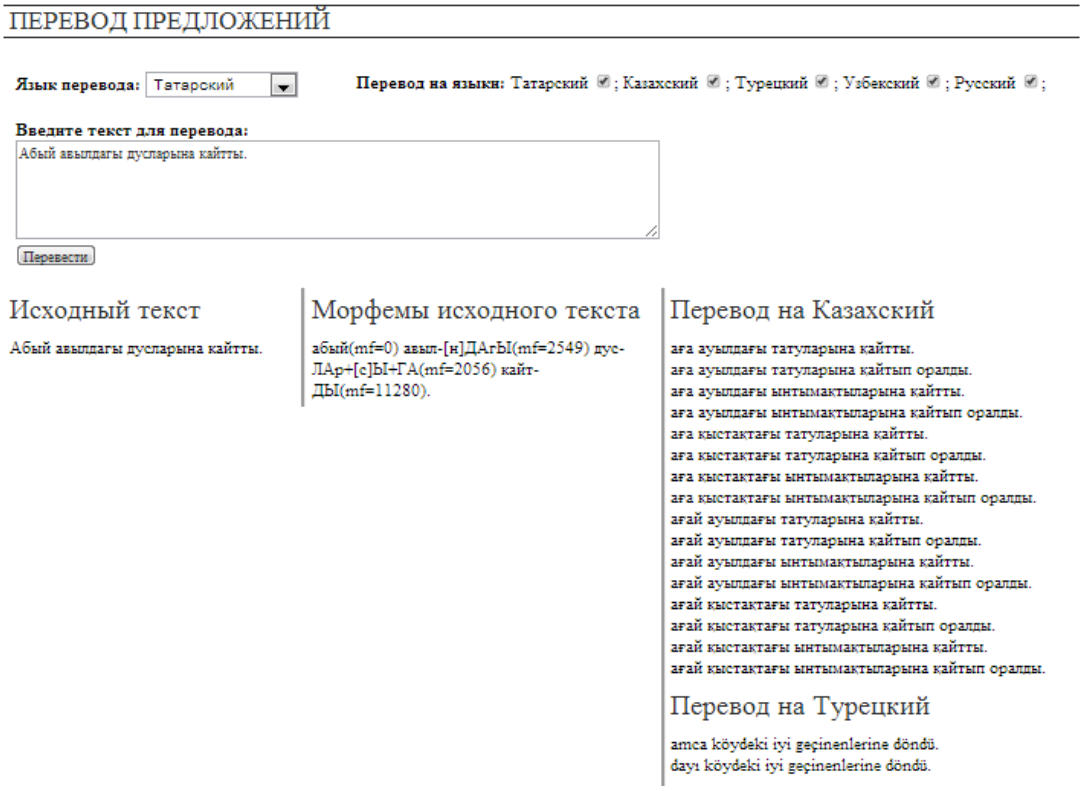


Рис. 7. Пример перевода предложения

Литература

1. [Muhtar M., 1994] Muhtar, M., Casablanca, F., Toyama, K., Inagaki, Y.: Particle-Based Machine Translation for Altaic Languages: the Japanese-Uighur Case. In: Proceedings of the 3rd Pacific Rim International Conference on Artificial Intelligence, Beijing, China, vol. 2, pp. 725–731 (1994).

2. [Altıntas, 2000] *Kemal Altıntas* Turkish to Crimean Tatar Machine Translation System. MSc Thesis, Bilkent University, Ankara, 2000.

3. [Fatullayev, 2008] *Fatullayev R, Abbasov A, Fatullayev A.* Dilmanc is the 1st MT system for Azerbaijani. In: Proc. of SLTC-08, Stockholm, Sweden, 2008. pp.63-64.

4. [Сулейманов, 1998] Сулейманов Д.Ш. Обработка ЕЯ-текстов на основе прагматически-ориентированных лингвистических моделей // Обработка текста и когнитивные технологии. Вып.3., 1998. С.205-212.

М.Х. ХАКИМОВ

Национальный Университет Узбекистана им. Мирзо Улугбека,
г. Ташкент, Республика Узбекистан

МОДЕЛИРУЕМАЯ ТЕХНОЛОГИЯ МАШИННОГО ПЕРЕВОДА

Осуществление компьютерного перевода производится с помощью специальной среды, составляющими которой являются программы, реализующие алгоритм перевода, в которых разработана последовательность однозначно и строго определенных действий над текстом для нахождения переводных соответствий в данной паре языков L_1 - L_2 при заданном

направлении перевода. Система компьютерного перевода включает в себя двуязычные словари, снабженные необходимыми грамматическими информациями (морфологической, синтаксической и семантической) для обеспечения передачи эквивалентных, вариантных и трансформационных переводных соответствий, а также алгоритмические средства грамматического анализа, реализующие какую-либо из принятых для автоматической переработки текста формальных грамматик.

Коммуникативная эквивалентность текста перевода по отношению к оригиналу должна обеспечить выполнения трех основных требований:

- текст перевода должен в возможно более полном объеме передавать содержание оригинала, что прежде всего означает недопустимость произвольного опущения или добавления информации;
- текст перевода должен соответствовать нормам языка перевода, так как их нарушение, по меньшей мере, создает помехи для восприятия информации, а иногда ведет и к ее искажению;
- текст перевода должен быть примерно сопоставим с оригиналом по своему объему, чем обеспечивается сходство стилистического эффекта с точки зрения лаконичности или развернутости выражения.

Машинный перевод (МП) - это выполняемое на компьютере действие по преобразованию текста с одного естественного языка в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

После машинного или автоматического перевода с помощью редактора осуществляется постредактирование, который исправляет ошибки и недочеты в переведенном на компьютере тексте.

Действующие системы компьютерного перевода ориентированы на конкретные пары языков (например, английский и русский или японский и английский) и используют, как правило, переводные соответствия либо на поверхностном уровне, либо на некотором промежуточном уровне между входным и выходным языком. Качество компьютерного перевода зависит от объема словаря, объема информации, приписываемой лексическим единицам, от тщательности составления и проверки работы алгоритмов анализа и синтеза, от эффективности программного обеспечения. Современные аппаратные и программные средства допускают использование словарей большого объема, содержащих подробную грамматическую информацию. Информация может быть представлена как в декларативной (описательной), так и в процедурной (учитывающей потребности алгоритма) форме.

Мощное внедрение новых информационных технологий дал новый импульс для дальнейшего развития теории и практики машинного перевода (МП). Мировая индустрия МП объединяет исследователей, разработчиков программного обеспечения и пользователей. За последние несколько лет, отмечается небывалый рост интереса к МП, который в основном связывают с развитием Интернета. Никогда ранее МП не был известен столь широкому кругу пользователей. И никогда еще у программного обеспечения этого класса не было пользователей с таким громадным опытом работы. В США сложились особые отношения между разработчиками систем МП и правительством, которое считает МП "ключом в информационный век". Особенно важным считается использование систем МП в научных исследованиях, здравоохранении, в области высоких технологий, охраны окружающей среды.

Перспективы развития компьютерного перевода связаны с дальнейшей разработкой и углублением теории и практики перевода, как компьютерного, так и «человеческого». Для развития теории важны результаты сопоставительного языкознания, общей теории перевода, теории закономерных соответствий, способов представления знаний, оптимизации и совершенствования лингвистических алгоритмов. Новые и более эффективные словари с необходимой словарной информацией, строгие теории терминологизации лексики, теория и практика работы с подъязыками помогут повысить качество перевода лексических единиц. Формальные грамматики, ориентированные на перевод, дадут возможность оптимизировать

алгоритмы нахождения переводных соответствий в данной коммуникативной ситуации, которая может быть описана в рамках соответствующих прикладных теорий представления знаний. Наконец, новые возможности программирования и вычислительной техники также будут вносить свой вклад в совершенствование и дальнейшее развитие теории и практики машинного перевода.

Современный машинный перевод следует отличать от использования компьютеров в помощь человеку-переводчику. В последнем случае имеется в виду автоматический словарь, помогающий человеку быстрее подбирать нужный переводной эквивалент. В содержание термина «машинный перевод» входит представление о том, что главную, большую часть работы машина берет на себя, оставляя человеку лишь контроль и исправление ошибок, в то время как компьютерный словарь в помощь человеку - это чисто вспомогательное средство для быстрого нахождения переводных соответствий; однако при этом, такого рода электронных словарях в ограниченной степени могут быть реализованы и некоторые функции, присущие системам машинного перевода.

Флективно-корневые языки, к которым относится, в частности русский язык, характеризуются по словам Н.С. Трубецкого «…неуловимыми корнями, постоянно меняющими свою огласовку и теряющимися среди префиксов и суффиксов», с трудом поддаются из-за своей идиоматичности модельному представлению и алгоритмизации. Но, тем не менее, в русском языковедении на сегодняшний день достаточно широко представлено теоретическое описание и практическая разработка многих сторон русской языковой системы, что даёт возможность широких обобщений и сопоставлений с языками другой структуры на конкретном и элементарном прикладном уровне.

Между тем агглютинирующие языки (и среди них особенно узбекский язык) с прозрачным построением парадигм и относительно регулярным порождением словоформ, представляющих собой синтагматические цепочки хорошо ограниченных друг от друга корневых словообразующих и формообразующих морфем, гораздо более удобны для применения приёмов современной прикладной лингвистики. Сожаление вызывает тот факт, что в узбекском языковедении всё ещё очень мало исследований и лексикографических произведений, которые так необходимы для нужд логическо-лингвистического моделирования и компьютерного перевода.

Учёт специфических особенностей каждого языка данной пары имеет определяющее значение, как для их системного изучения, так и для логическо-лингвистического моделирования. Следует отметить, что теоретико-языковедческими и инженерно-лингвистическими вопросами индоевропейских и иноструктурных языков занимаются – языковеды, математики, программисты, историки, философы, социологи, психологи и психиатры.

С одной стороны, этот интерес объясняется тем, что проблема человека становится одним из центральных вопросов нашей цивилизации, а исследование его языка превращается в одно из действенных средств изучения мышления человека, его индивидуального и коллективного поведения, а одновременно и истории народа – носителя конкретного языка. С другой стороны, внимание к языку и лингвистике стимулирует характерный для нашей эпохи научно-технической революции – интерес к нечётким, но хорошо приспосабливающимся к любой обстановке и надёжно функционирующим системам большой сложности. Классическим примером является система естественного языка.

Сложные и нечёткие системы не всегда удаётся до конца проанализировать, а затем и смоделировать с помощью традиционного математического аппарата. Здесь нас интересуют два вопроса: пределы применения к языку современного формального аппарата, а также направление, в котором должен развиваться и совершенствоваться этот аппарат с тем, чтобы стать эффективным средством изучения и моделирования таких хорошо адаптирующихся нечётких и сложных систем, какими являются системы разноструктурных (например, русского и узбекского) языков.

Для построения системы компьютерного перевода должен быть решен обширный круг проблем:

1. Лингвистические проблемы - определение состава словника для выбранной области, установление запаса сведений, которые должны содержаться в словаре, и построение словаря, выбор типа грамматики и построение грамматической модели.

2. Математические проблемы - разработка общей структуры алгоритма перевода. Разработка алгоритмов отдельных этапов, разработка формализмов для записи лингвистических данных и для разработки алгоритма.

3. Проблемы машинной реализации - разработка способов хранения данных, создание системы программирования, разработки комплекса программ реализующих различные алгоритмы моделирования, а также разработка разного рода программ обслуживания.

Математическое описание языка основано на представлении о «правильных текстах». Правильный текст определяется как последовательность речевых единиц, подчиняющаяся определённым закономерностям, другими словами, правильный текст – это предложение, построенное по строго определённым правилам. Множеством узлов этого предложения (П) служат слова, входящие в П. Среди узлов – один корень, не подчинённый никакому узлу. Нельзя, отправившись из какого-либо узла вдоль стрелок, вернуться в тот же узел. Узлы дерева подчинения – это вхождения слов в предложения. Формально для каждого (не слишком короткого) предложения можно построить много разных синтаксических структур любого из двух видов, но среди них либо одна или несколько являются правильными. Корнем правильного дерева подчинения служит обычно сказуемое.

Более совершенное представление синтаксической структуры предложения (требующее, однако, более сложного математического аппарата) дают системы синтаксических групп, в которые входят как словосочетания, так и синтаксические связи, причём не только между словами, но и между словосочетаниями. Системы синтаксических групп позволяют совмещать строгость формального описания строения предложения с гибкостью, присущей традиционным, неформальным описаниям. Деревья подчинения и системы составляющих являются предельными частными случаями систем синтаксических групп.

Другой раздел математической лингвистики, занимающий в ней центральное место - теория формальных грамматик, начало которой было положено работами Н. Хомского [1]. Она изучает способы описания закономерностей, характеризующих уже не отдельный текст, а всю совокупность правильных текстов того или иного языка.

Современные функционирующие системы МП обеспечивают лишь 40-55% синтактико-семантической правильности текста перевода, что подтверждает их все еще слабой формализованности [2], т.к. именно строгая математическая формальность языка может обеспечить высокую степень в точности перевода. Хотя известно, что формализация любого естественного языка относится к категории трудно решаемых проблем. Следует отметить, что особенно актуальны проблемы формализации узбекского языка и внедрения систем МП с включением узбекского языка в многоязычную ситуацию. В связи с этим проведение научных исследований в области формализации естественных языков, разработка и внедрение многоязычных систем МП требует необходимых теоретических выкладок. В настоящей работе изложены основные понятия (аксиомы) для формальных систем МП в многоязычной ситуации.

Определение 1. *Формальной системой* называется система, состоящая из множества специальных символов, множества понятий, баз слов/фраз и конечного множества математических моделей считающихся интерпретируемыми.

Определение 2. *Математическая модель естественного языка* – это есть способ формального описания его синтаксических и семантических конструкций. Основой синтаксических конструкций является вывод слово, а семантических конструкций правильный вывод фразы.

Утверждение 1. Одна математическая модель определяет одно или несколько синтаксических и/или семантических конструкций из грамматики естественного языка.

Утверждение 2. Каждая математическая модель является либо распознающей, либо порождающей в многоязычной системе МП.

Утверждение 3. Математическая модель является *распознающей*, если она характеризует язык **А** или является *порождающей* если характеризует язык **В**. При этом направлением МП считается **А** → **В**, а языки **А** и **В** принадлежат по классификации Н. Хомского [1] классу 0.

Определение 3. *Распознающая математическая модель* это анализ синтаксических и семантических конструкций выводящих висячее дерево предложений языка **А.**

Определение 4. *Порождающая математическая модель* это синтез синтаксических и семантических конструкций строящих дерево предложения языка **В.**

Утверждение 4. Каждая синтактико-семантические правила языков **А** и **В** имеют форму **a** → **b** без каких-либо ограничений на строки **a** и **b** в границах грамматики рассматриваемого языка.

Определение 5. Язык **А** есть множества форм **a** → **b**.

Определение 6. *Понятия Р* есть конечное множества словообразующих форм.

Определение 7. *Начальным символом* естественного языка **А** является любая буква из его алфавита - **Е**, называемым *терминальным символом*.

Определение 8. *Интерпретация* - это множество построения различных алгоритмов в соответствие с формами вывода **a** → **b**.

Определение 9. *Формальная грамматика G естественного языка* это есть G = **{Λ, Ф, П, Ψ}**, где

- **Λ** множество терминальных символов;
- **Ф** вспомогательное множество нетерминальных символов и фраз, с помощью которых определяются терминальные символы и понятия;
- **П** начальный символ, **П** = <математическая модель ЕЯ>;
- множество продукций **Ψ: χ→γ (χ≠γ, χ∈{Ф}, γ∈{Λ U Ф})**.

Целью формальной грамматики является определение с помощью правил вывода принадлежность слов, фраз и предложений к данному языку или наоборот строить слова, фразы и предложения в соответствии с правилами вывода этого языка. Таким образом, по сути, формальная грамматика представляет собой исчисление и для превращения его во множество алгоритмов позволяющих задать четкие правила вывода языка **А** внедряем в формальную грамматику математические модели.

Если $\varphi \to \psi$ - правило грамматики G и ω_1, ω_2 – цепочки из основных и вспомогательных символов, говорят, что цепочка $\omega_1\ \psi\ \omega_2$ непосредственно выводима в G из $\omega_1\ \varphi\ \omega_2$. Если ξ_0, ξ_1,… ξ_n… - цепочки и для каждого I = 1, … n цепочка ξ_1 непосредственно выводима из ξ_i - 1, говорят, что ξ_n выводима в G из ξ_0. Множество тех цепочек из основных символов, которые выводимы в G из её начального символа, называется языком, порождаемым грамматикой G и обозначается L(G). Если все правила G имеют вид η_1 A $\eta_2 \to \eta_1\ \omega\ \eta_2$, то G называется грамматикой составляющих (или непосредственно составляющих), сокращённо НС – грамматикой. Основные (терминальные) символы – это слова, вспомогательные (нетерминальные) – это грамматические категории (S – существительное, V - глагол, O – объект и т.п.). В НС-грамматике вывод предложения даёт для нас дерево составляющих, в котором каждая составляющая состоит из слов, «происходящих» от одного вспомогательного символа, так что для каждой составляющей указывается её грамматическая категория.

Определение 10. *Математическая модель слово* M_C в грамматике G это есть вывод формы вида **a** → **a**, либо $a_{i,j} \to (p_i a\ V\ ap_j)$, где $p_i \in P$, $i=\overline{1,n}$

Определение 11. *Математическая модель* M_Δ *предложения* в грамматике G это есть вывод **a→b**, где **a** содержит синтаксически правильную последовательность элементов a_i .

Определение 12. Язык **А** определяемый грамматикой G есть множества моделей трех типов $M_{\Delta k}$ $(k=\overline{1,3})$, в соответствии с типами предложений естественного языка.

Определение 13. Язык **А** называется *неоднозначным,* если он содержит хотя бы одну математическую модель любого типа для которой существуют более одной формы вывода **a→b**.

Определение 14. *Предложение* естественного языка – это одна из математических моделей любого типа.

Определение 15. *Технология МП* - это процесс достижения однозначности перевода в многоязычной ситуации в результате внедрения формальных систем.

Так как достижение однозначности происходит в разной степени функционирующих системах МП, то определим критерии по классификации технологий МП.

Определение 16. *Чистая технология МП* из языка **A** в язык **B** есть установление однозначности между грамматиками **A**(G) и **B**(G) в пределах 97-100%.

Чистая технология *МП* практически снимает вопрос постредактирования, возложив почти все проблемы на систему. Внедрение чистых технологий является серьезнейшей проблемой и скорее всего ее можно решить в недалеком будущем.

Определение 17. *Высокая технология МП* из языка **A** в язык **B** есть установление однозначности между грамматиками **A**(G) и **B**(G) в пределах 65-80%.

Внедрением высоких технологий МП практически занимаются многие исследователи, результаты работ должны появиться очень скоро.

Определение 18. *Средняя технология МП* из языка **A** в язык **B** есть установление однозначности между грамматиками **A**(G) и **B**(G) в пределах 40-55%.

К данной категории технологии МП можно отнести такие системы как ПРОМТ (Россия), SYSTRAN, Transparent Language (США), Lingvistica (Канада), Cross Language (Япония) [8].

Определение 19. *Низкая технология МП* из языка **A** в язык **B** есть установление однозначности между грамматиками **A**(G) и **B**(G) в пределах 25-35%.

Естественный человеческий язык с точки зрения математики представляет собой нечёткое или размытое множество – континуум. Нечёткость языка, в том числе значений слов, словосочетаний и других лингвистических единиц обусловливается особенностями восприятия и отражения объективной действительности в мозгу человека. Поэтому строение знака в математике и в естественном человеческом языке разное. Знак в математике – это двусторонняя сущность. Знак в языке многозначен и многопланов.

Потребность математики и информатики в языковедении и прикладной лингвистике связана с необходимостью построения алгоритмов, позволяющих быстро и эффективно извлекать и перерабатывать информацию, заключённую в научно-технических, деловых и художественных текстах, поток которых постоянно возрастает. Условием такой переработки является перевод информации, содержащейся в неформализованном виде в тексте, на формализованный искусственный язык. Если речь идёт об автоматизированной переработке текста, то таким искусственным языком является расширяемый входной язык математического моделирования естественного языка [3].

Более сложны лингвистические потребности робототехники и теории искусственного интеллекта. Обращаясь к опыту математической и прикладной лингвистики, исследователи ищут конструктивные решения применительно к формальному анализу нечётких объектов, к устранению многозначности языковых знаков и созданию алгоритмов высоких информационно-семантических уровней. Однако решение этих сложных задач невозможно без предварительной разработки методов системного описания и моделирования парадигматики и синтагматики языка.

Цель математической лингвистики как науки состоит в том, чтобы изложить элементы системного анализа языка и речи с помощью аппарата современной математики и элементарных математических правил, приложимых к лингвистике.

Наиболее простыми для формализации и компьютерного перевода являются тексты научного характера с чётко представленным синтаксисом во взаимосвязи с ограниченным числом морфологических категорий. Это объясняется тем, что термин, т.е. слово в специфически научном употреблении, максимально приближен к математическому знаку по своей сути, что позволяет формализовать конструкции и осуществить переход к синтаксическому анализу на основе представленных в базе данных морфологического характера.

Двусоставные предложения биноминальной структуры являются наиболее частотными и коммуникативно-значимыми в научных текстах.

В настоящее время ведутся исследования именно в направлении математического моделирования естественных языков. В качестве объектов выбраны узбекский, русский, английский, немецкий и турецкие языки. Разработана **технология моделируемого компьютерного переводчика** (МКП).

МКП состоит из трех ступенчатой архитектуры. На первой ступени модели проводится синтаксический и семантический анализ естественного языка. Здесь определяются все – префиксы, суффиксы, приставки, окончания, предлоги, морфемы, постфиксы, союзы, модальные слова, частицы, междометия, аффиксы и т.д. Слова исследуются по категориям, т.е. составляющим части предложения (числительное, существительное, прилагательное, местоимения, глагол, наречие). Создаются специальные базы данных по указанным категориям [6,7].

На второй ступени архитектуры выявляются все синтаксические и семантические связи построения слов и словосочетаний с построением логико-лингвистических моделей в рамках «сущность-связь».

Для описания математических моделей слов, словосочетаний и предложений был разработан расширяемый входной язык математического моделирования естественного языка [3].

На третьей ступени архитектуры описываются математические модели естественных языков [4,5,7].

Технология МКП требует наличия различных баз данных по естественному языку, например [7]. Кроме указанных требуется наличие базы слов по категориям естественных языков участвующих в переводе, а также предметных словарей, которые будут составлять базы данных со специальными атрибутами, например [8-11].

Осуществление перевода компьютером – сложная, но интересная научная задача. Основная ее сложность состоит в том, что естественные языки плохо поддаются формализации. Отсюда и невысокое качество получаемого с помощью систем МП текста. Однако идея машинного перевода уходит корнями далеко в прошлое.

В последнее время большое значение придается автоматизированным информационным технологиям. Свидетельством тому является обсуждение этого вопроса на различных международных форумах за последние 10 лет. Так, 12 декабря 2003 года в Женеве (Швейцария) состоялся Всемирный Саммит, посвященный проблеме построения Информационного Общества. Он проходил под лозунгом: "Построение Информационного Общества – глобальный вызов нового тысячелетия. Саммит принял два документа: Декларацию о принципах создания Информационного Общества и План работы по реализации этих принципов.

В Декларации формулируются принципы построения Информационного Общества с учетом социально-политических, правовых и гуманитарных аспектов. При этом подчеркивается центральная роль науки в развитии такого Общества и в развитии информационных и телекоммуникационных технологий.

В Плане конкретизируются пути построения открытого Информационного Общества. При этом указывается, что потенциал человеческих знаний и информационных и телекоммуникационных технологий следует направить на достижение задач развития, одобренных международным сообществом. Большое значение придается необходимости сохранения культурного многообразия и языковой самобытности народов, населяющих землю, и в этой связи подчеркивается важность исследований и разработок в области машинного перевода.

Литература

1. Мальков В. Формальные модели анализа и распознавания языковых структур/ Материалы международной конференции «Диалог-2007». М, 2007.

2. Хомский Н. Формальные свойства грамматик. «Кибернетический сборник», НС, вып. 2, 1966, стр. 121-230.

3. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. ҰзМУ хабарлари, № 1, 2009, 75-80 с.
4. Хакимов М.Х. Математические модели узбекского языка. ҰзМУ хабарлари, № 3, 2010, с.185-188
5. Хакимов М.Х. К моделям естественных языков для многоязычных ситуаций компьютерного перевода. Труды научной конференции «Проблемы современной математики» 22-23 апреля 2011 г., г. Карши, с.531-537
6. Хакимов М.Х. Абдурахманова Н. Семантические базы английского языка для многоязычной ситуации компьютерного перевода. Труды научной конференции «Проблемы современной математики» 22-23 апреля 2011 г., г. Карши, с.311-314
7. Хакимов М.Х. Семантические базы и математические модели русского языка для многоязычной ситуации компьютерного перевода. Проблемы информатики и энергетики, №2, 2011, с.57-65
8. Хакимов М.Х. База англо-русско-узбекских терминов и фраз по компьютерным знаниям. ГПВ РУ, РА №6, 2008, Свидетельство № BGU 00139
9. Хакимов М.Х. База англо-русско-узбекских терминов и фраз по химии. ГПВ РУ, РА №6, 2008, Свидетельство № BGU 00140
10. Хакимов М.Х. База русско-узбекских терминов и фраз по математике. ГПВ РУ, РА №6, 2008, Свидетельство № BGU 00141
11. Хакимов М.Х. База русско-узбекских терминов и фраз по таможне, геодезии, почве и агрохимии. ГПВ РУ, РА №4, 2009, Свидетельство № BGU 00179
12. www.promt.ru, www.systransoft.com, www.transparent.com, www.lingvistika.com, www.crosslanguage.co.jp/englisn

У.А. ТУКЕЕВ, С.З. САПАКОВА, А. МАРАТҚЫЗЫ, Қ.ӨТЕПОВА

Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан

ҚАЗАҚША-ОРЫСША МАШИНАЛЫҚ АУДАРМАСЫНЫҢ МӘЛІМЕТТЕР БАЗАСЫ ЖӘНЕ ОНЫҢ ҚҰРЫЛЫМЫ

1. Мәліметтер базасының құрылымы

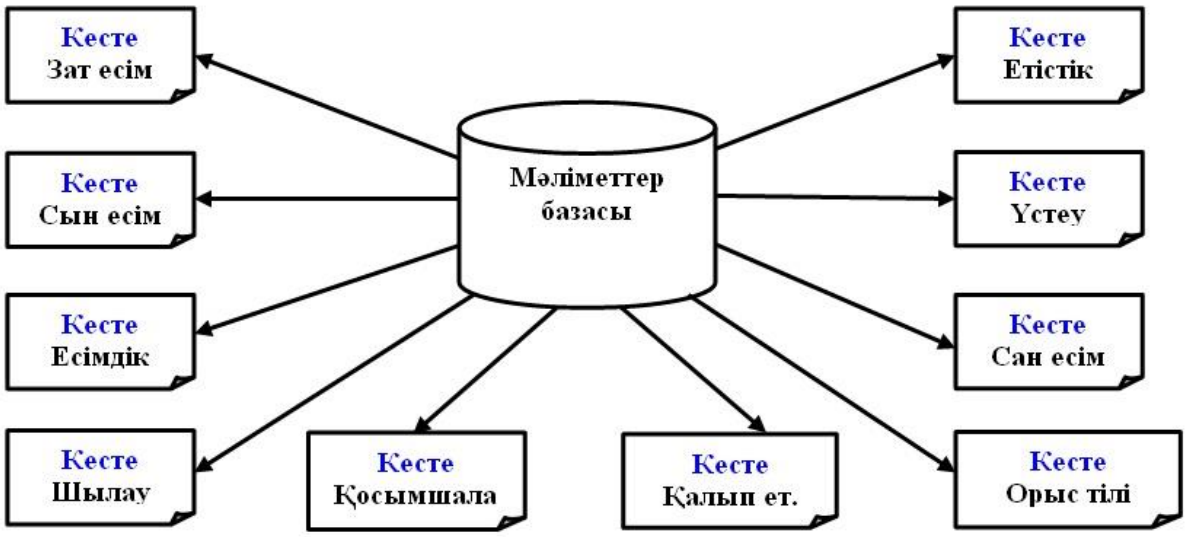
Мәліметтер қоры дегеніміз ақпаратты сақтауға және жинақтауға арналған ұйымдасқан құрылым. Ең алғаш мәліметтер қоры ұғымы жаңадан қалыптасқан кезде онда шындығында мәліметтер сақталатын. Бірақ қазіргі кездегі көптеген мәліметтер қоры басқару жүйелері өздерінің құрылымдарында тек мәліметтерді ғана емес, сонымен қатар олардың тұтынушымен және басқа да ақпараттық – программалық кешендермен қарым – қатынасының әдістерін де қамтиды. Сондықтан біз қазіргі заманғы мәліметтер қорында тек мәліметтер ғана емес, ақпараттар да сақтай аламыз.

Мәліметтер базасы деп деректердің электрондық сақтаушысын айтады. Оларға қатынас бір немесе бірнеше компьютерлер көмегімен іске асады. Әдетте деректер базасы деректерді сақтау үшін жасалады.

Мәліметтер базасы – ақпаратты сақтауды және мәліметтерге ыңғайлы, тез кіруді қамтамасыз етеді. Мәліметтер базасы белгілі бір ережелерге сай құрылған деректер жиынтығын құрайды.

Мәліметтер базасын басқару жүйесі деректер базасын құруға, толтыруға, жаңартуға, жоюға арналған программалық жабдық болып табылады.

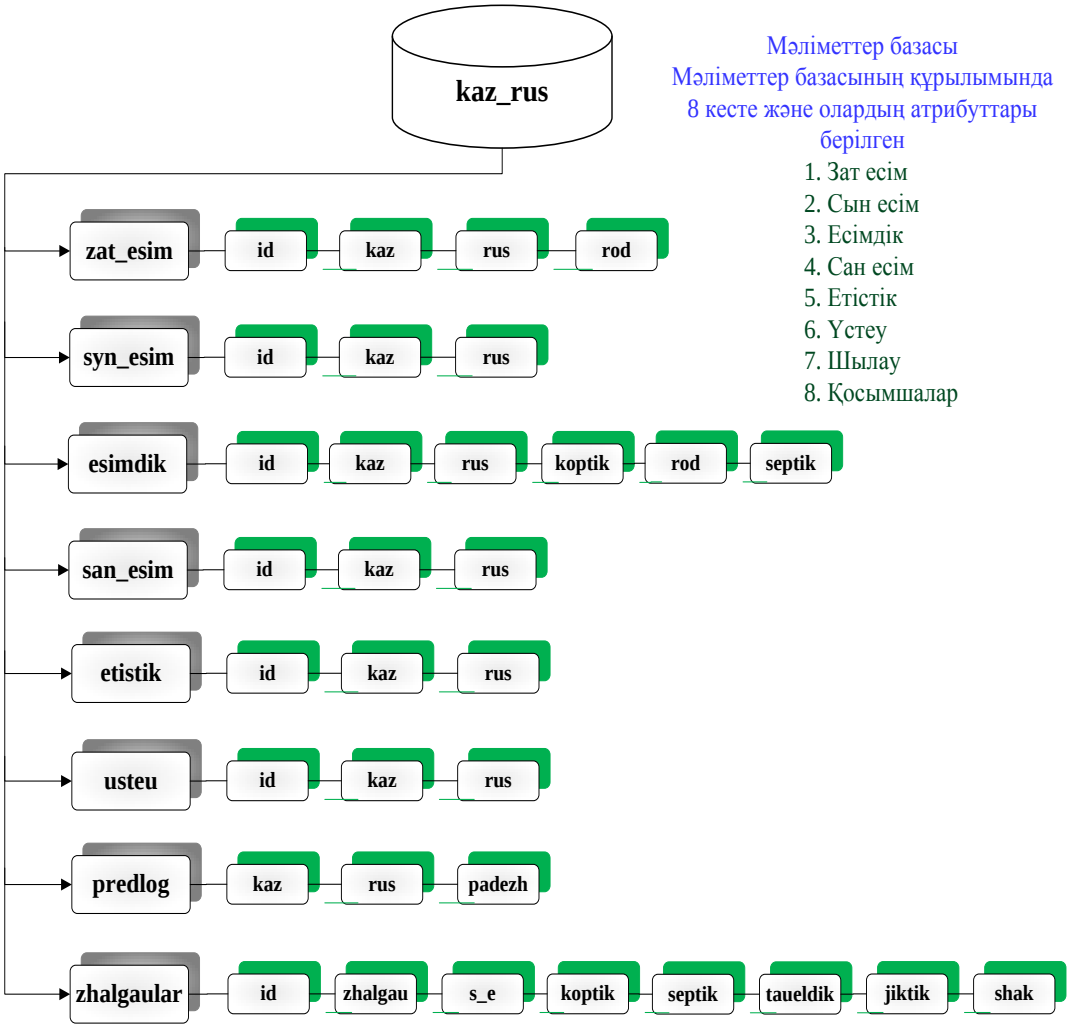
Бұл жобада SQLite арқылы Қ.Б. Бектаевтың «Үлкен сөздігі» бойынша толтырылған 12000 сөзі бар мәліметтер базасы және 753 қосымшаларды және олардың атрибуттарын қамтитын қосымшалар кестесі бар. Талдаулар соның негізінде жасалынады. Мәліметтер базасының жалпы құрылымы келесі суретте берілген (1.1-сурет):



Сурет 1.1. Мәліметтер базасы.

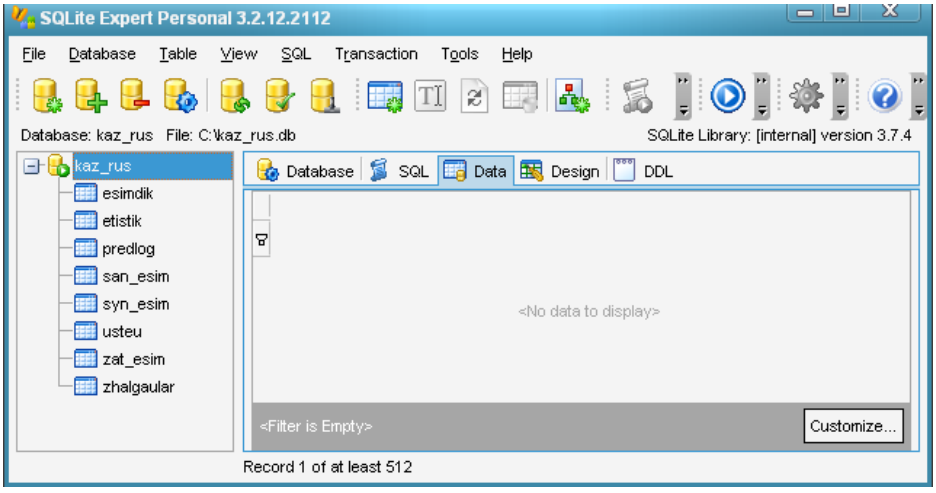
Мәліметтер базасын программаға қосу үшін dotconnectsqlite компонентін орнату қажет. Осыдан кейін sqlite форматындағы мәліметтер базасы дайын болса, онымен байланыстыратын кодты енгіземіз:

```
using System.Data.SqlClient; // Байланыс үшін қажетті директивалар
using System.Data.OleDb; //
using Devart.Data.SQLite; //
using System.IO; //
String mySelectQuery;
SQLiteCommand sqCommand;
SQLiteDataReader sqReader;
SQLiteConnection sqConnection = new SQLiteConnection("Data Source=c:\\zhalgaular.db;");
```

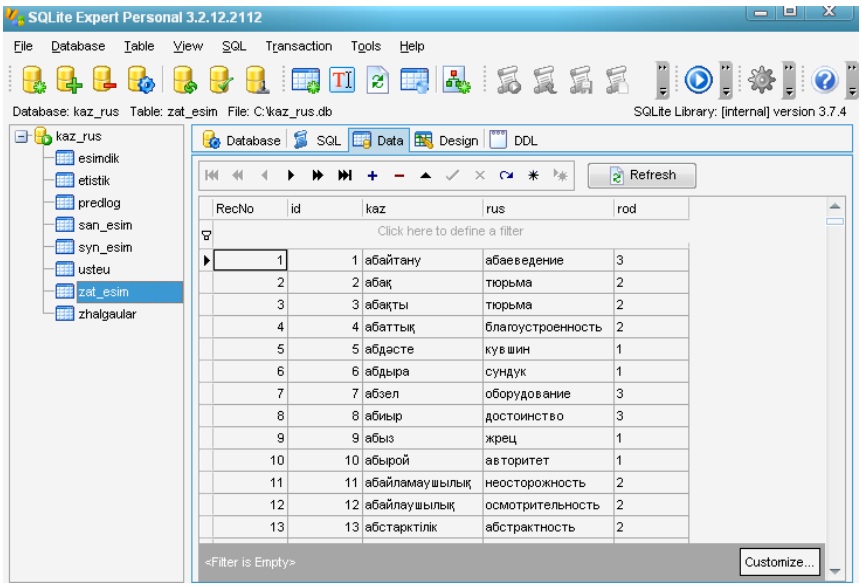


Сурет 1.2. Мәліметтер қорының моделі.

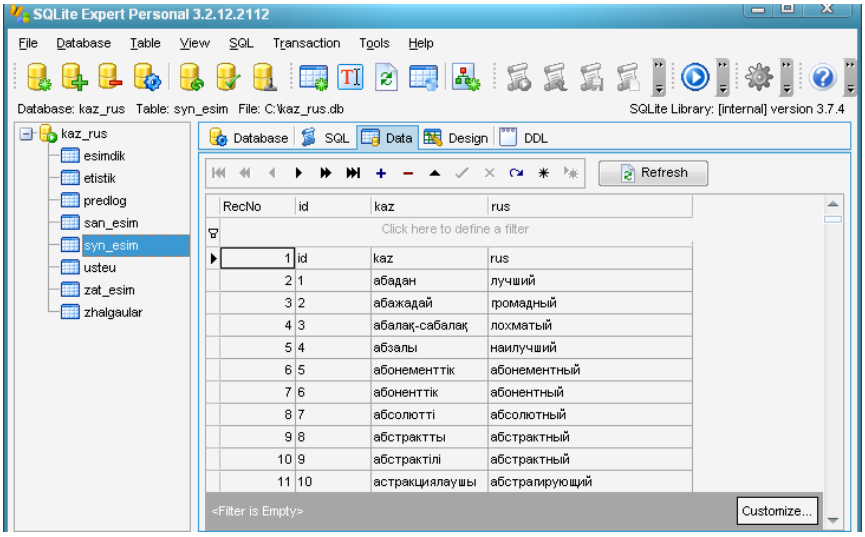
Бұл машиналық аударманың мәліметтер базасы SQLite программасында құрастырылған. Барлығы 8 кестеден тұрады (1.1.-сурет). Кестелер сөз таптарына байланысты және бір кесте қосымшаларға арналған. Олар: "**zat_esim**" (зат есім), "**syn_esim**" (сын есім), "**san_esim**" (сан есім), "**esimdik**" (есімдік), "**etistik**" (етістік), "**usteu**" (үстеу), "**predlog**" (шылау), "**zhalgaular**" (қосымшалар) (1.2 – 1.11-суреттер):



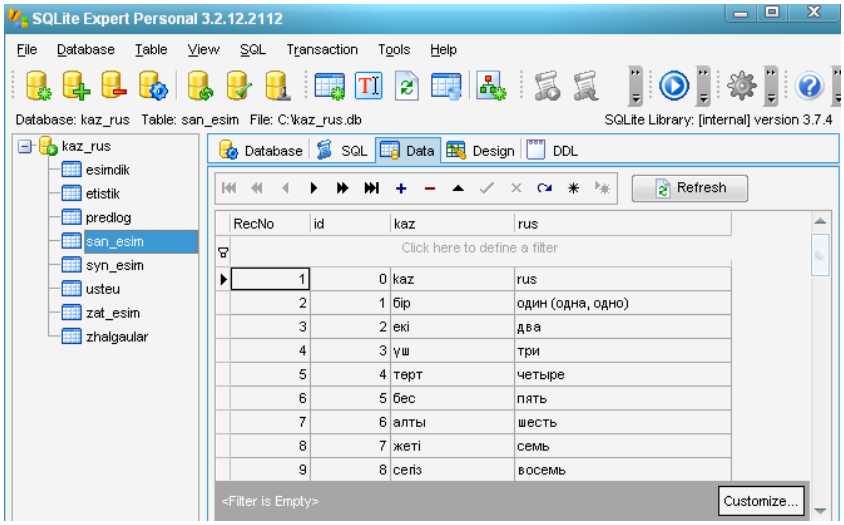
Сурет 1.3. SQLite программасының жалпы көрінісі.



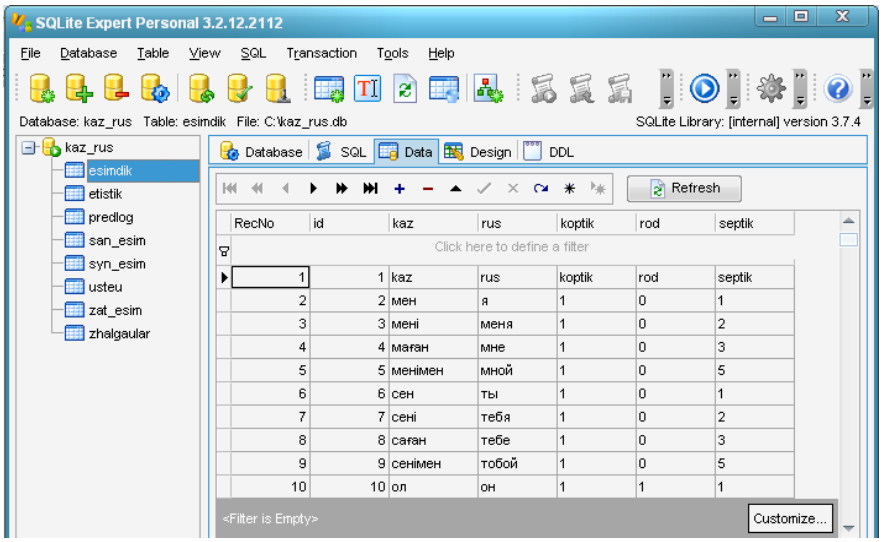
Сурет 1.4. Zat_esim кестесі.



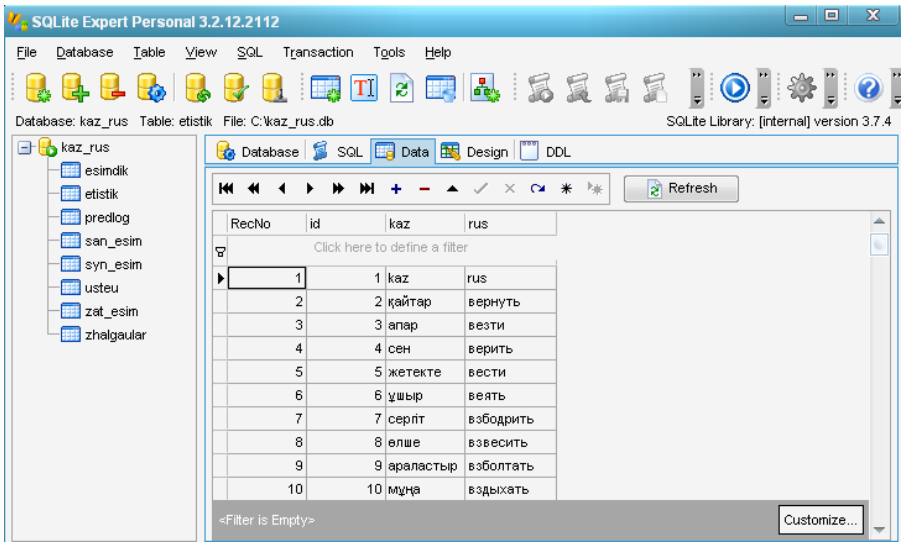
Сурет 1.5. Syn_esim кестесі.



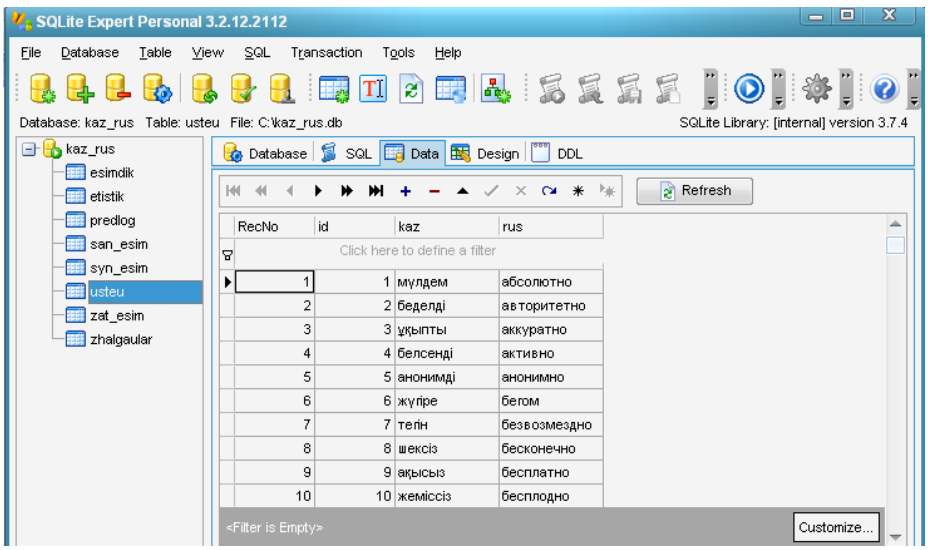
Сурет 1.6. San_esim кестесі.



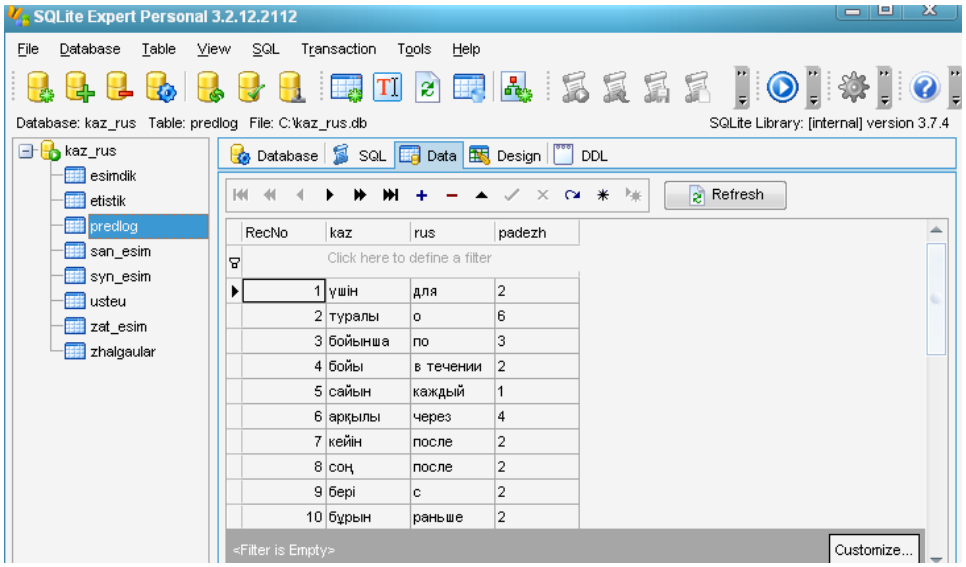
Сурет 1.7. Esimdik кестесі.



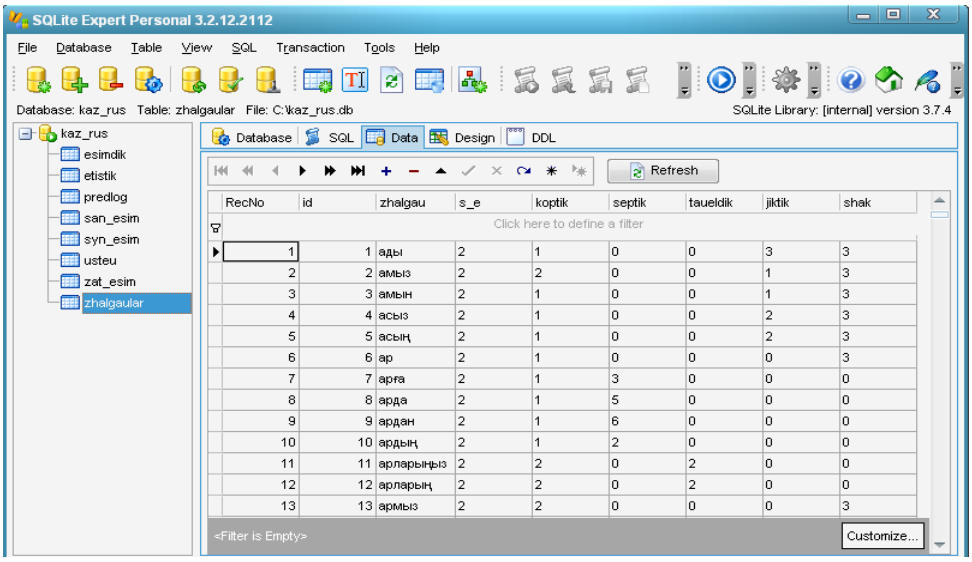
Сурет 1.8. Etistik кестесі.



Сурет 1.9. Usteu кестесі.



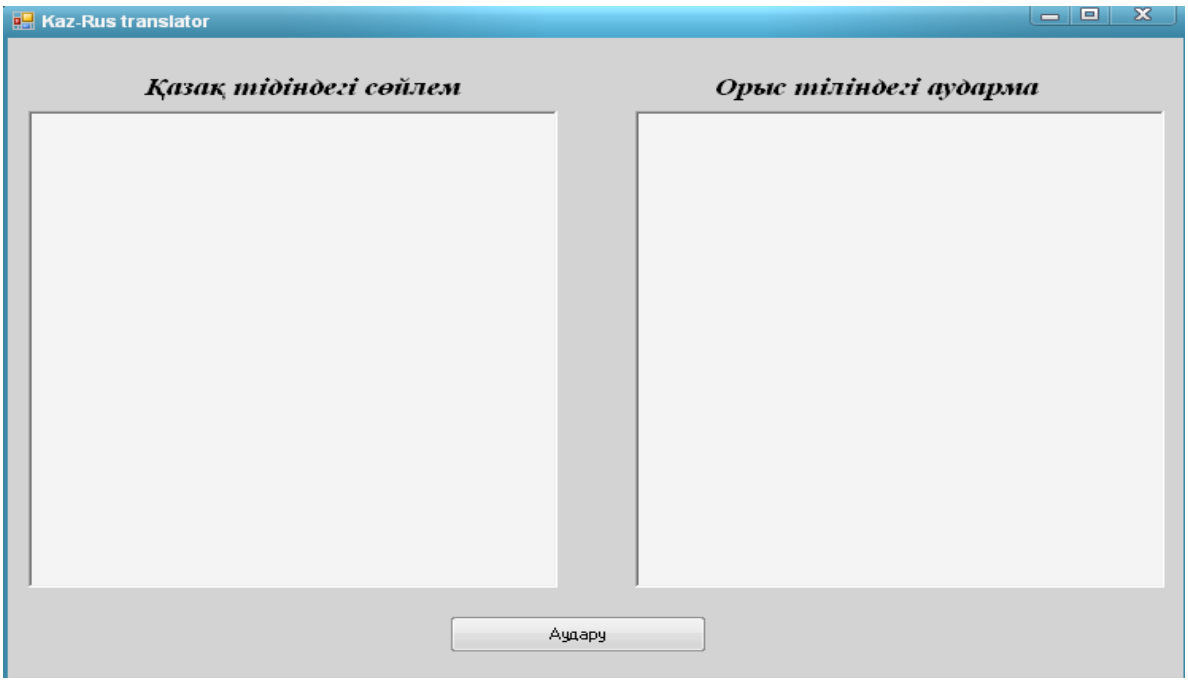
Сурет 1.10. Predlog кестесі.



Сурет 1.11. Zhalgaular кестесі.

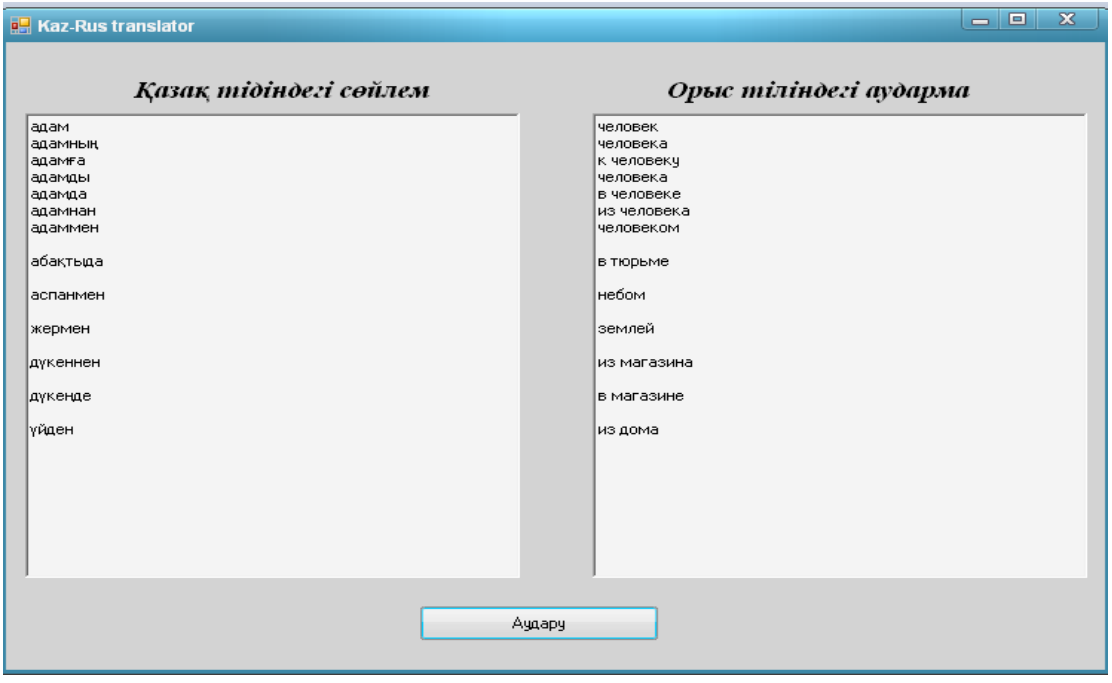
2. Мәліметтер базасы және программалық жабдықтау жұмыс істеу нәтижелері

Осы жұмыс барысында индекстік файлдар арқылы жұмыс жасауды дұрыс деп игердік. Себебі, олар жұмысты жеңілдетуге көп септігін тигізді. Қолданылған индекстік файлдар, олар: қазақ тілінде берілген сөзге жалғанатын қосымшаларға арналған кесте, сонымен қатар орыс тіліндегі аударма алынған кездегі оның жалғауларын анықтап, сәйкестендіруге пайдаланылатын кестелер. Мәліметте базасындағы кестелерге жүргізілетін операциялардан бөлек, бірнеше ережелер жазылды. Соның ішінде, қазақ және орыс тілінің септік жалғауларын сәйкестендіру, қазақ және орыс тілінің етістіктерін және олардың шағын, қай жақта, жекеше немесе көпше түрде берілгендігін анықтау сияқты т.б. бірнеше ережелер жазылды. Осы айтылған жұмыстарды қамтып, қазақша-орысша машиналық аударма программасын құрастырдық, ол келесі суретте көрсетілген (2.1-сурет):



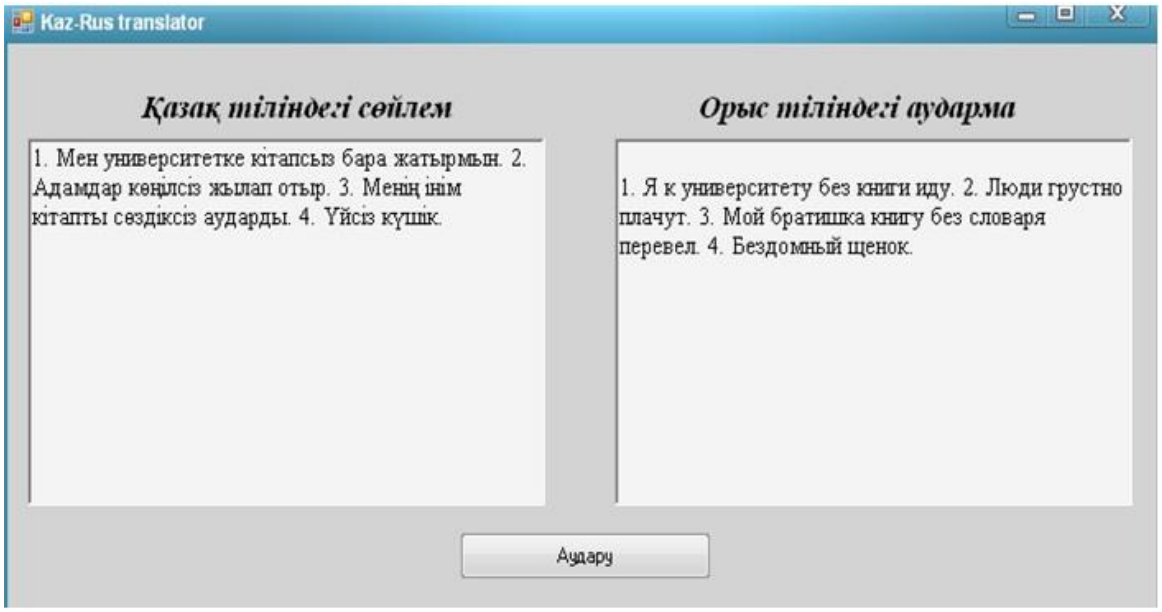
Сурет 2.1. Қазақша-орысша машиналық аударманың интерфейсі.

Машиналық аударма жасау кезінде қазақ және орыс тілдерінің грамматикасы ұқсас болғанымен, толық сәйкес емес екендігіне тағы да көз жеткіздік. Мысал ретінде септік жалғауларын айтсақ та, олар қазақ тілінде жетеу, ал орыс тілінде алтау. Бірақ, олардың сәйкестігі анықталып, қажетті ережелер жазылды. Нәтижеміз келесі суретте берілген (2.2-сурет):



Сурет 2.2. Сөздердің септелуі.

Қазақ тілінде берілген сөйлемнің орыс тіліндегі жақсы аудармасын алуға да қол жеткіздік (2.3-сурет):



Сурет 2.3. Сөйлем аудару.

Қорытындылай келе, машиналық аударма қазіргі кезде үлкен сұранысқа ие екені белгілі. Қазіргі таңда жұмыс істеп тұрған қазақ тілінен орыс тіліне аударатын demo_kaz_rus программасы сөйлемдердің құрылымын, сөздің мағынасын ескере отырып түсінікті аударма жасауда. Еліміздегі машиналық аудармалардың сапасын арттыру мақсатында атқарылып жатқан жобаның тиімділігі айқын.

Әдебиеттер

1. Бектаев, К.Б. Большой казахско - русский, русско-казахский словарь / Калдыбай Бектайұлы Бектаев.- Алматы: Алтын қазына, 2007.- 709 с.

У.А. ТӨКЕЕВ, С.З. САПАҚОВА

Әл-Фараби атындағы ҚазҰУ, Алматы, Қазақстан

ҚАЗАҚ ТІЛІНЕН ОРЫС ТІЛІНЕ МАШИНАЛЫҚ АУДАРМА

1. Қазақ тілді машиналық аудармашыларға қысқаша шолу

Қазіргі таңда қазақ тілінен өзге тілдерге аударатын программалар, онлайн-аудармалар баршылық, бірақ олардың жұмыс нәтижесі мардымды емес. Оның ең негізгі себебі қазақ тілінің грамматикасының басқа тілдің грамматикасына қарағанда анағұрлым күрделілігі, өзге тілдің грамматикасына ұқсамайтындығында. Еліміздегі қазіргі кезде кеңінен қолданылып жүрген sozdik.kz, soylem.kz, sanasoft.kz секілді онлайн аудармашылармен қатар «Ізет-тілмәш» қолданбалы программасында қазақ тілінен орыс тіліне аудару мүмкіндігі бар. Бірақ бұл аудармашы программалар енгізілген сөздерді аударғанымен сөйлемнің құрылымына, сөз мағынасына аса мән бермейтінін олардан алынған нәтижелерден көре аламыз. Айта кететін жайт осы айтылған машиналық аудармаларда сөйлемдер енгізіп, оларды аударатын болсақ ол сөздердің көп жағдайда орнын өзгертпей, басқа мағыналарын қарастырмайтынын көреміз,

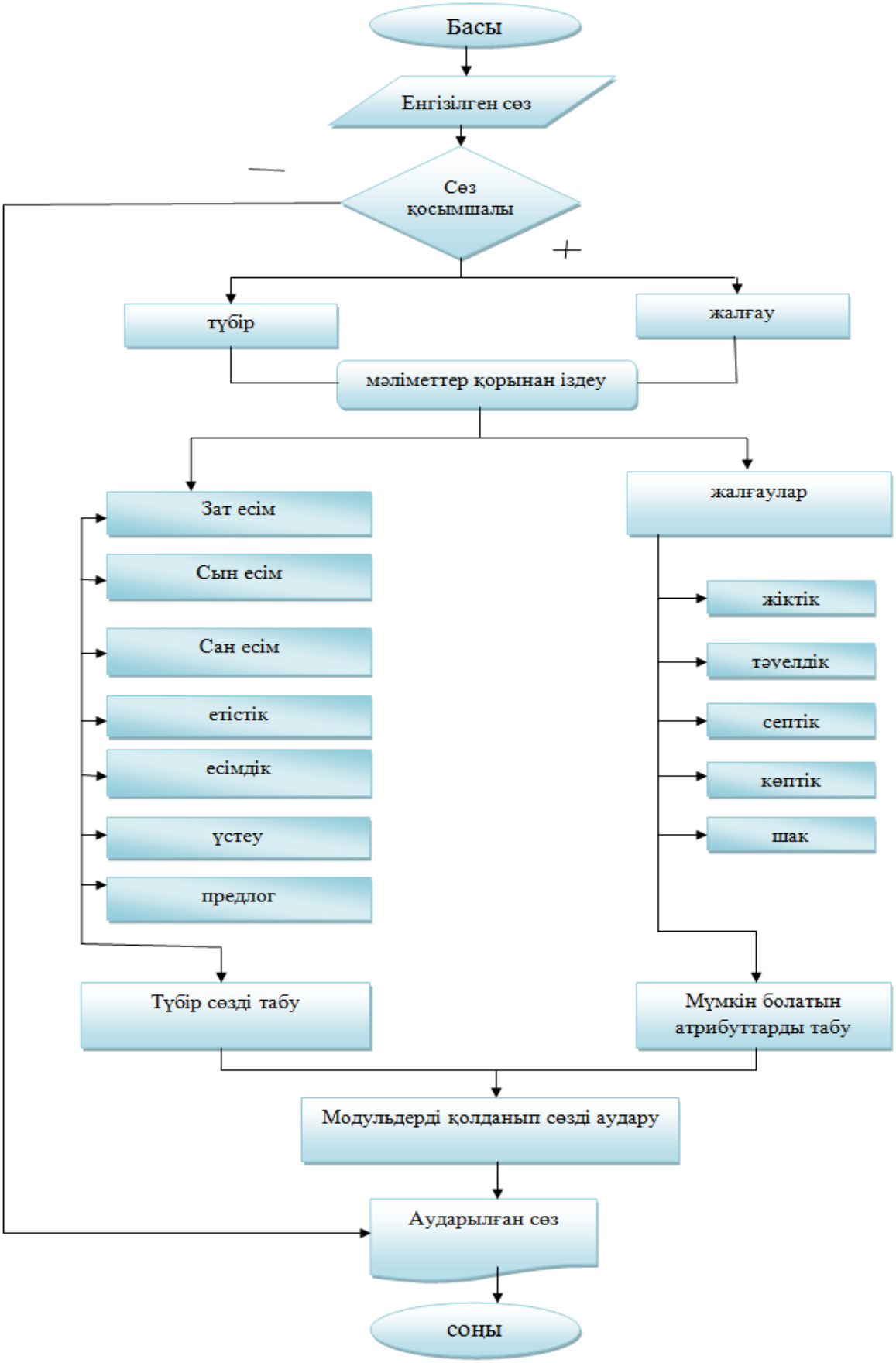
яғни түсініксіз аудармаларға грамматика толық қарастырылмағандықтан тап боламыз, оны дәлелдемесекте болады. Сонымен қатар, бұл бағытта атқарылып жатқан іс-шараларды да атап кететін болсақ, «ағылшын- қазақ» машиналық аударма бағытында Apertium программасы Микель L. Forcada (Испания) басшылығымен және әл-Фараби атындағы Қазақ Ұлттық Университетінің қолдауымен, free/ машиналық аударманың ашық кодты платформасы пайдаланылуда. Apertium –бұл машиналық аударма жүйесінің ашық кодасын құруға арналған құралдардың жиыны, әсіресе өзара байланысқан тілдер жұбы үшін ыңғайлы, оның құрамына ашық лингвистивкалық мәліметтерге арналған бірнеше сөздіктер, техникалық қызмет көрсету т.с.с. енетінін білеміз. Осы бағдарлама негізінде «қазақ- татар» тілдер бағытындағы ашық кодалы жүйе құрып, онымен қарқынды айналысып жатқан ғалымдарды: Ильнар Салимзянов, Джонатан Вашингтон және Фрэнсис Tyers атап кетуге болады.

Ұсынылып отырған жұмыс қазақ-орыс бағытында құрылған машиналық аударма жүйесінің негізгі жұмыс істеу принциптеріне, қазіргі таңда туындаған мәселерге тікелей байланысты. Жұмыс нәтижесінде шағын **«kaz-rus translator»** қолданбалы программасы жасалынды және одан әрі дамыту үстіндеміз, бұл программа Visual Studio 2010 және SQLite орталарында орындалды.

2. Қазақ тілінен орыс тіліне аудару барысындағы морфологиялық талдау сұлбасы

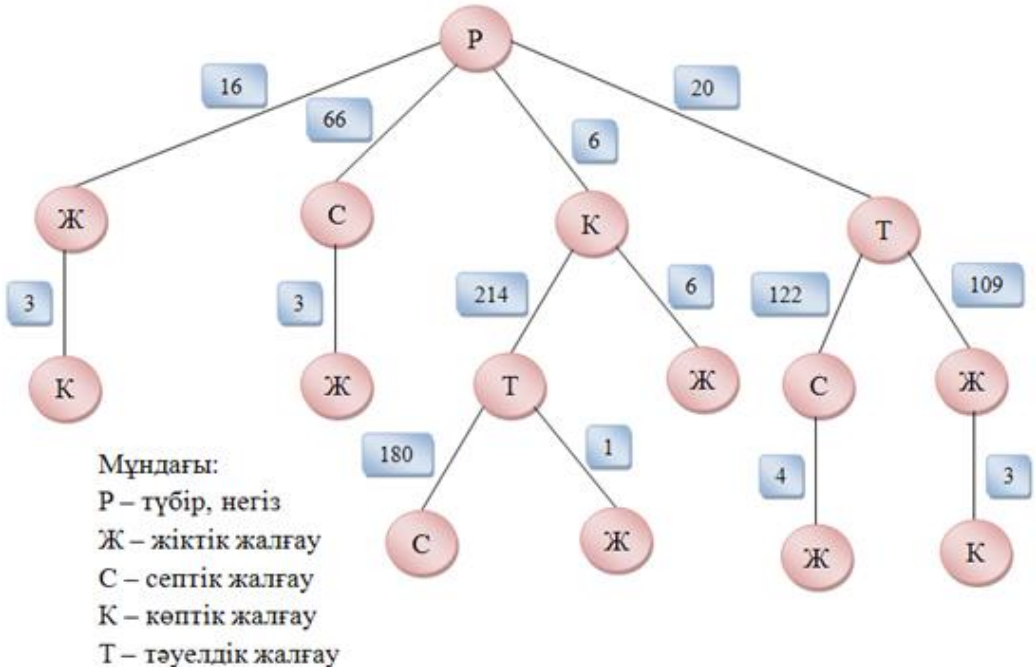
Қазіргі уақытта машиналық аударма барысында бірнеше бөліктерден тұратын күрделі жүйе қолданылады, мысалы:

- Морфологогиялық талдау – мәтіндегі сөздерді талдау
- Синтаксистік талдау – сөйлемдерді, грамматиканы және сөздер арасындағы байланыстарды талдау;
- Семантикалық талдау – белгілі бір пәндік аймаққа бағытталған деректер қоры негізінде әр сөйлемнің мағынасын талдау.
- Прагматикалық талдау- өзіндік мәліметтер қоры негізінде белігілі бір контекстің ауқымында сөйлемнің мағынасын талдау.



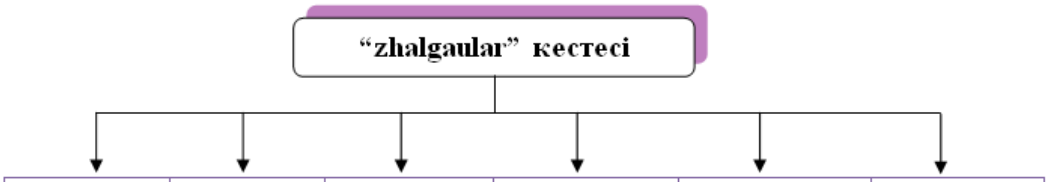
Сұлба 1. Қазақ тілінен орыс тіліне сөз аударудың блок-сұлбасы.

Бұл сұлбадан көретініміз морфологиялық талдауда маңызды рол атқаратын жалғаулардың МҚ бөлек кесте түрінде сақталуы. МҚ «жалғаулар» кестесінде кездесетін 753 жалғауды Бектаевтың сөздіктер кітабынан [1] енгіздік. Бектаевтың сөздігінде қазақ тілінде кездесетін барлық қосымшаларды қарастырылған, сол сөздік бойынша қосымшалардың жалғану реті жасалынды. Мысалы, бара-лар-ымыз-дың деген сөзді алсақ, қосымшалардың жалғану реті былай болады**: Р – К – Ж – Т (**түбір – көптік жалғау – жіктік жалғау – тәуелдік жалғау**)**. Осы қосымшалардың бүкіл жағдайын қарастырып, барлығын қосындыласақ 753 қосымша шығады (2-сұлба).



Сұлба 2. Қазақ тіліндегі жұрнақтардың жалғану реті.

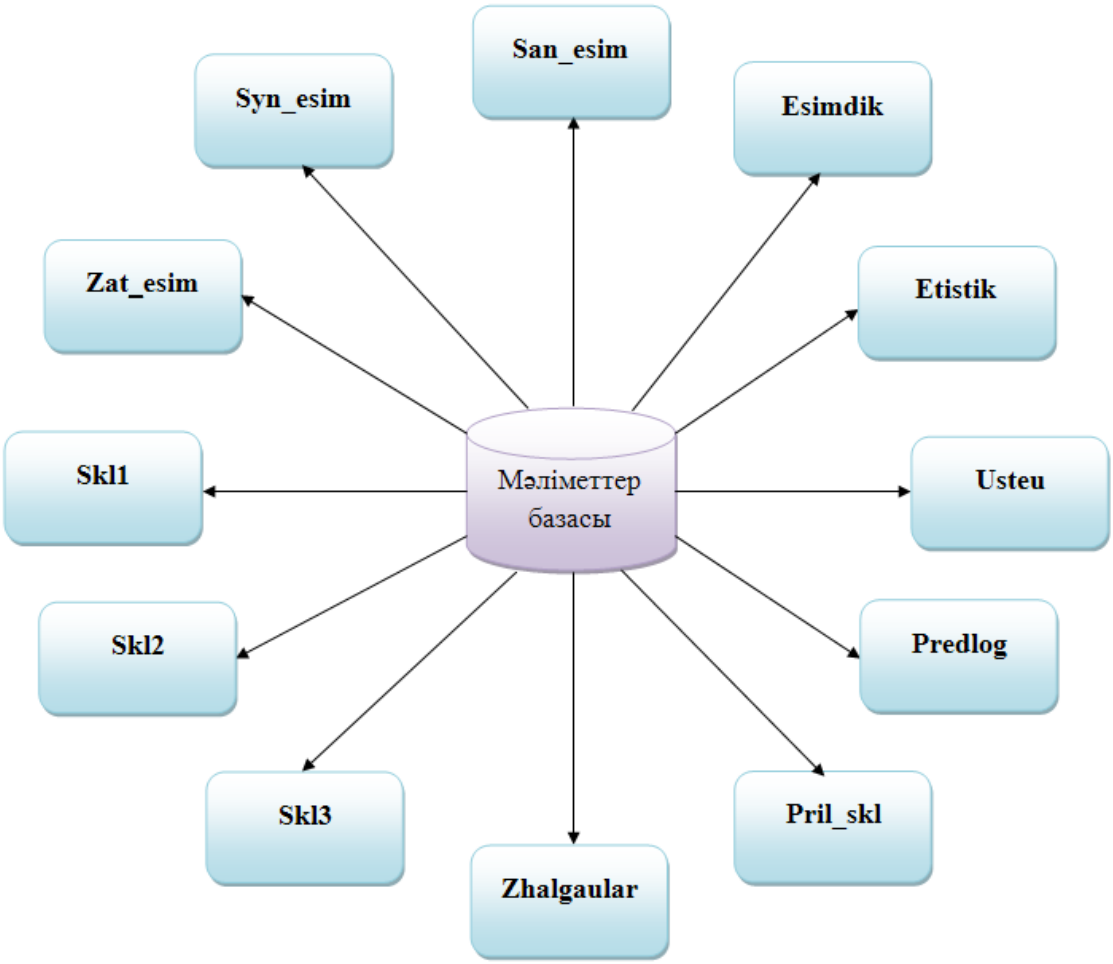
Қарастырылып отырған тілдер жұбында, әр тілдің өзіндік ерекшеліктері мен олардың арасындағы бірмәнді сәйкестік болмайтынын да айқын көреміз. Мысалы, қазақ тіліндегі әртүрлі сөз таптарының өзіне тиселі жалғаулары бар, ол басқа сөз табына жалғанбайды, сонымен қатар бірнеше сөз табына жалғанатындары да бар. Осы ерекшеліктерді ескере отырып, біз жалғаулар кестесіне келесі атрибуттарды пайдаландық



s_e	koptik	Septik	Taueldik	Jiktik	Shak
0-белгісіз	0-жоқ	0-жоқ	0-ешқандай	0-ешқандай	0-жоқ
1-есім сөз	1-жекеше	1-атау	1 – І-жақ	1 – І-жақ	1-өткен
2-етістік	2-көпше	2-ілік	2 – ІІ-жақ	2 – ІІ-жақ	2-осы
		3-барыс	3 – ІІІ-жақ	3 – ІІІ-жақ	3-келер
		4-табыс			
		5-жатыс			
		6-шығыс			
		7-көмектес			

Сұлба 3. Жалғаулар кестесінің атрибуттары.

Бұл жұмыста қарастырылып отырған Мәліметтер қорының құрылымы келесідей:

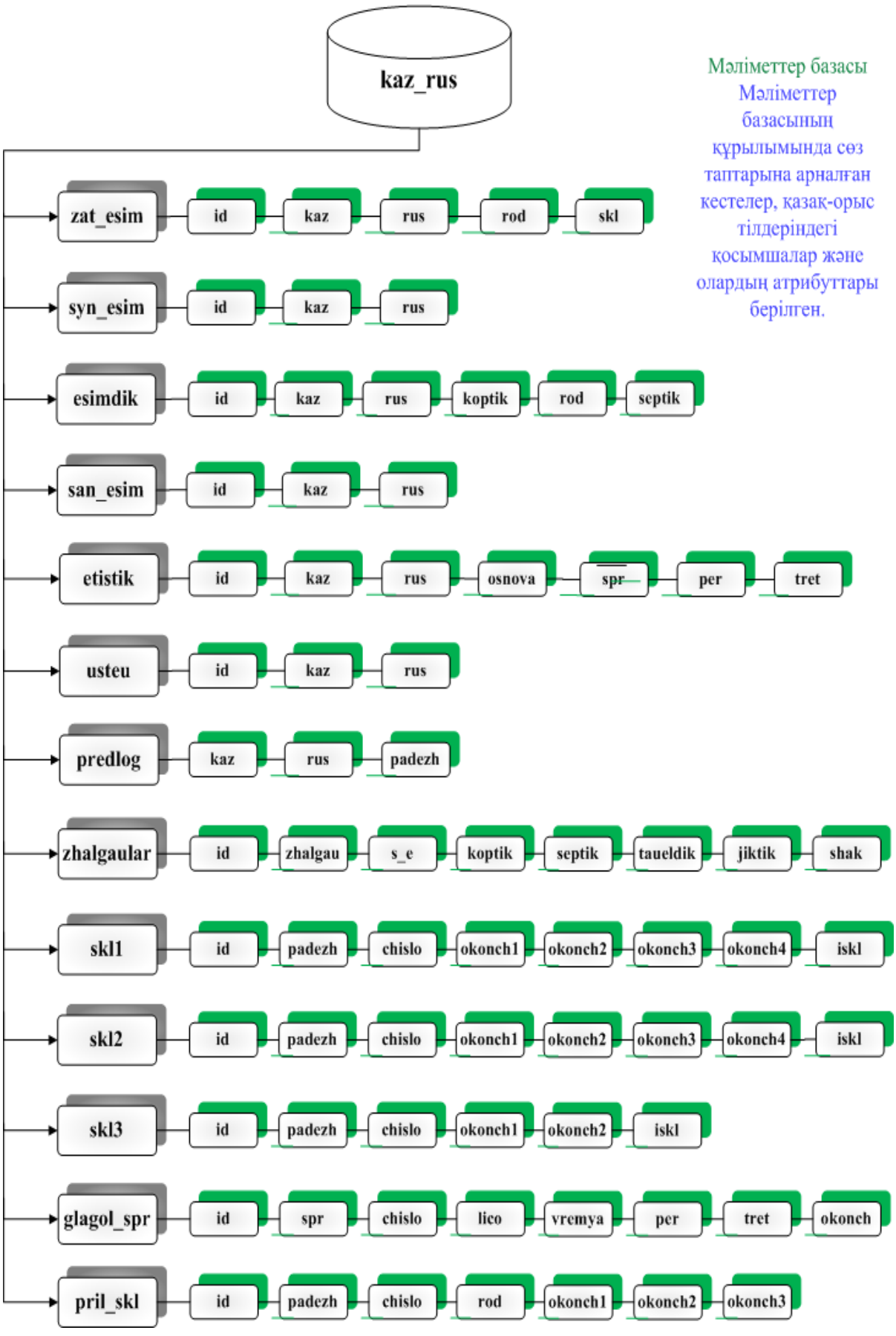


Сұлба 4. Мәліметтер қорының шағын сұлбасы.

Әрқайсысына келесідей жекеленген түсініктеме беріп кетуге болады:

- ✓ ***zat*** (зат есімнің қазақша-орысша аудармасы)
- ✓ ***syn*** (сын есімнің қазақша-орысша аудармасы)
- ✓ ***san_esim*** (сан есімнің қазақша-орысша аудармасы)
- ✓ ***etistik*** (етістіктің қазақша-орысша аудармасы)
- ✓ ***usteu*** (үстеудің қазақша-орысша аудармасы)
- ✓ ***esimdik*** (есімдіктің қазақша-орысша аудармасы)
- ✓ ***predlog*** (орыс тіліндегі предлог)
- ✓ ***zhalgaular*** (қазақ тіліндегі барлық мүмкін болатын жалғаулар)
- ✓ ***skl1*** (орыс тіліндегі склонение 1-дің жалғаулары)
- ✓ ***skl2*** (орыс тіліндегі склонение 2-дің жалғаулары)
- ✓ ***skl3*** (орыс тіліндегі склонение 3-дің жалғаулары)
- ✓ ***iya*** (орыс тілінде ия-ға бітетін зат есімдердің жалғаулары)
- ✓ ***iyi*** (орыс тілінде ий-ға бітетін зат есімдердің жалғаулары)
- ✓ ***pril_skl*** (орыс тіліндегі сын есімнің жалғаулары)
- ✓ ***glagol_spr*** (орыс тіліндегі етістіктің жалғаулары)

Осы МҚ кестелердің, әрқайсысының ерекшеліктері ескеріле отырып, келесі түрдегі атрибуттар тағайындалды.

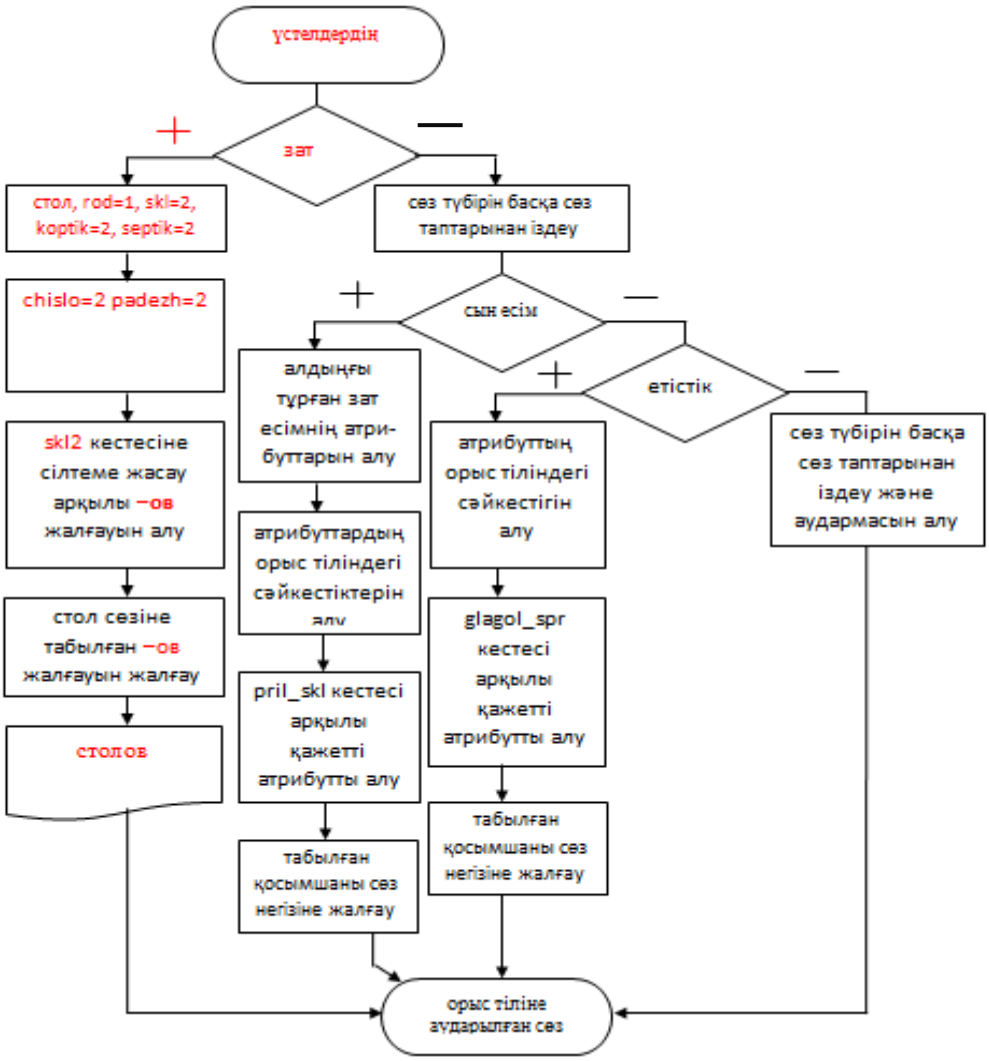


Сұлба 5. Сөз таптарының және көмекші кестелердің атрибуттары.

RecNo	id	zhalgau	s_e	koptik	septik	taueldik	jiktik	shak
		Click here to define a filter						
553	553	тен	1	1	6	0	0	0
554	554	тер	1	2	1	0	0	0
555	555	терге	1	2	3	0	0	0
556	556	терде	1	2	5	0	0	0
557	557	терден	1	2	6	0	0	0
558	558	терді	1	2	4	0	0	0

Сурет 1. Мәліметтер қорындағы «Жалғаулар» кестесі.

Мысал, «үстелдердің» көпше түрдегі сөзді аудару процесін қарастырсақ.
<үстелдердің>::=<үстел><дердің>
<үстел>::=<стол><rod=1, skl=2>
<дердің>::=<chislo=2, septik=2 >



Сұлба 7. Морфологиялық генератордың сұлбасы.

Сұраныс бойынша МҚ «ов» жалғауы алынады және сөздің соңына жалғанады. «үстелдердің» - «стол**ов**».

Бұл жұмыста бүкіл атқарылып жатқан жұмыстардың сипаттамасын, ішкі құрылымын көрсету мүмкін емес, сондықтан «зат есім» сөз табының маңызды жақтарын қарастырсақ.

Соның ішінде Септік жалғаулары жобада толық қарастырылған. Мысал ретінде **Табыс септігін** алсақ, ол орыс тіліндегі **винительный падежге** сәйкес.

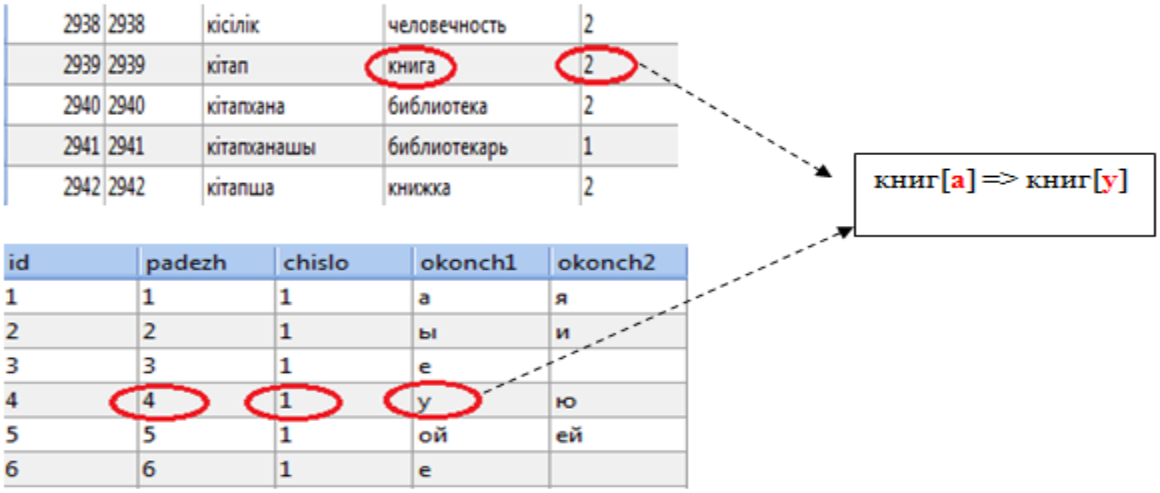
1-склонение. Егер енгізілген сөздің соңғы әріпі «а» және оның соңғы әрпінің алдындағы әрпі келесі жиыннан болса, pb[]={ 'б', 'в', 'г', 'д', 'ж', 'з', 'к', 'л', 'м', 'н', 'п', 'р', 'с', 'т', 'ф', 'ч', 'ш', 'щ', 'х', 'ц'}, онда аударылған сөздің соңына «у» жалғауы жалғанады, мысалы, көлік**ті** =>машин**у**.

–у немесе –ю жалғауларын жалғау ережелері:

МҚ келесі атрибуттары бар сұраныс түседі: padezh=4, chislo=1, rod=2.

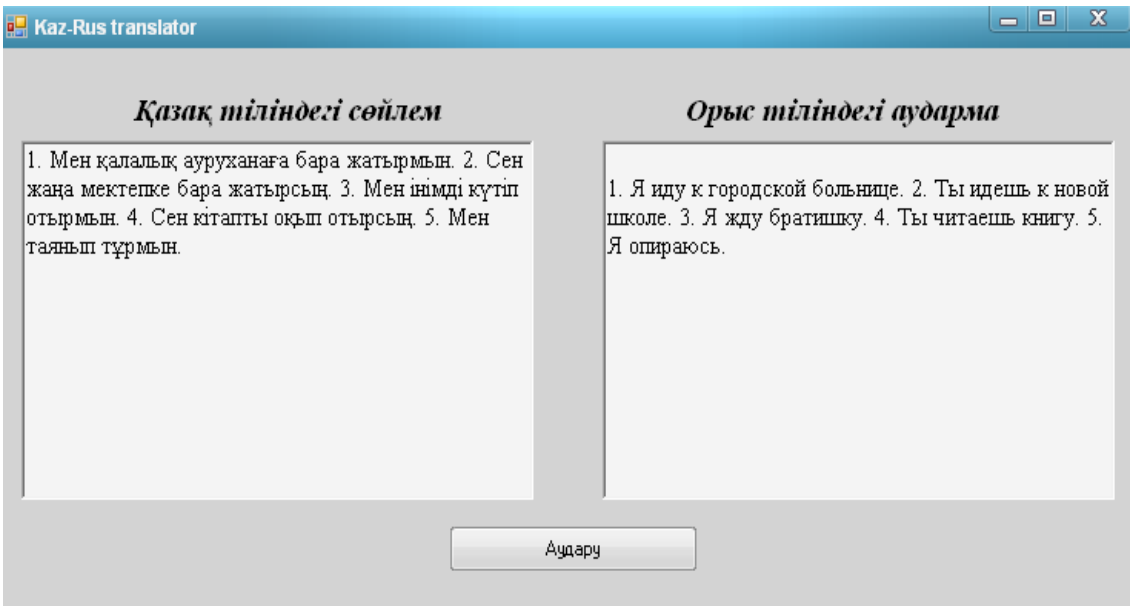
```
string[] string1 = new string[] { "я" };
   if (defineVariables.skl[defineVariables.i] == "1")
   {  if (padezh == "4")
{if
(string1.Contains(defineVariables.words_rus[defineVariables.i].Substring(defineVariables.words_r
us[defineVariables.i].Length - 2, 1)))
    { defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch2; }
    else   {   defineVariables.words_rus[defineVariables.i] =
   defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1;          }
```

Нәтижесі:



Сурет 2. Зат есімдер кестесі мен сұраныстар кестесінің нәтижесі.

Kaz-Rus translator жұмыс нәтижесін көрсететін болсақ келесі қарапайым сөйлемдерді аударады.



Сурет 3. Kaz-Rus translator жұмыс нәтижесі.

Қорыта айтқанда,

- қазақ тілінен орыс тіліне аударатын машиналық аудармалар салыстырылып, оларға талдау жасалды, яғни сапалы машиналық аударма алу үшін кеткен қателіктер зерттелді;
- машиналық аударманың негізі болып табылатын мәліметтер қорын толтыру нұсқалары қарастылып ең тиімді шешім алынды, кестелерге қажетті атрибуттар анықталды;
- 12000 сөз енгізілген мәліметтер қорына программа күрделілік деңгейін азайту үшін орыс тілінің жалғаулары қосылды;
- қазақ тілінен орыс тіліне машиналық аудармада кездесетін қиындықтар талқыланып, шешуге қажетті модульдер құрылды.

Әдебиеттер

1. Бектаев, К.Б. Большой казахско - русский, русско-казахский словарь / Калдыбай Бектайұлы Бектаев.- Алматы: Алтын қазына, 2007.- 709 с.

N.Z. ABDURAKHMONOVA

National University of Uzbekistan named after Mirzo, Tashkent, Uzbekistan

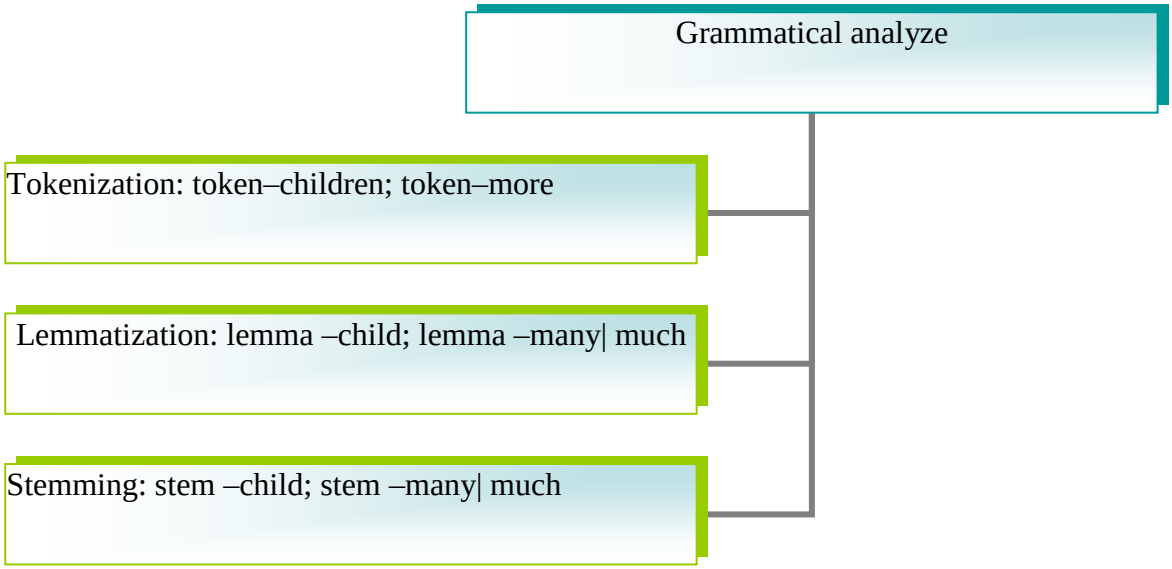
GRAMMATICAL ANALYZE IN MACHINE TRANSLATION BETWEEN ENGLISH AND UZBEK

Today human society characterizes high degree of activity in different fields such as economy, science, technology, culture etc. And it has caused to increase body of information that presents some difficulties between person to person or among the group of people. Computer has being considered one of the main approaches to ease opportunities of people since it was invented. So machine translation is used to exchange information communicative attitudes. Translation of the text is very complex creative process from a natural language into another one. We can see now variety forms of machine translation system; even they can recognize speech and translate orally. Most of them are multilingual translation programs. The Uzbek language is being developed rapidly

after our independence. Therefore our research has taken the first step to build of linguistic database of translation program.

It is always noticed to mainly grammatical analyze (morphological and syntax) of the lexemes in any system (retrieval of database, machine translation, automatic editor). So analyze is investigated as one of the base of linguistic approaches.

There are many methods such as tokenization, lemmetization, stemming which used in the automatic morphologic analysing process for the texts. Tokenization helps to detach the meaning of units of speech (token, wordform) separately.The previous forms of the words are determined in lemmatization process. One of the processes is stemming. The roots of the words are found by its assist. Three ways which are used for analyze, we may describe like a chart:



In some literatures above mentioned methods are used as terms which are special procedures and programs of creating corpus – database of computer software[1].

Our decent research has shown the necessity of morphoclassificator in the automatic process to translate from English texts into Uzbek. Morphological classification of the words might be taken as the main way to clarify part of speech. We admit that two languages belong to other language family.Some issues are demanded that to solve to input the linguistic database to computer software. We cite an example contrasting between adjective and adverb in English with equivalency in the Uzbek language: *It is a **good** impression –Bu ijobiy taassurot. He speaks English **well (badly)** –U ingliz tilida **yaxshi (yomon)** gapiradi*. Both of the words (***yaxshi, yomon***) are considered as adjective. But they are only analyzed as modifier in syntax not in morphology as adverb. Focus on the problem is these words are not existed in adverbial list in the process. As well as we can face to again other examples between adjective and noun: *I like to eat wooden bowl –Men yog`och kosani yoqtiraman.* In this place "yog`och"is not "*wood*" but "*wooden*".

Naturally, English and Uzbek are member of different type of language so their linguistic nature is diverse too. For instance some pronounces in English don't exist in Uzbek, and they are called as other categories: *few, little* words are used as adverb such as "kam, oz" in Uzbek. Such problems seem easy to solve at first. It depends on not only electron dictionary which are decoded in the languages, but also it is responsive to grammatical analyze in context. One is of the urgent request for any translation to save agreement between the form and sense of the text.

Transformation method is estimated as effective way to solve the problems. Four stages proceed in the transformation method: transposition, substitution, replenishment, and omission [2]. But other author presents only three ways: adjunction, substition, deletion. In addition to this, the base of the transformation process contains kernel structures, and it consists of simple sentences in syntax [3].

These main stages signifies in the process of analyze and synthesis. We analyze the bases on the examples abovementioned types of transformation method.

In the process of transposition it has been analyzed words and word combinations in the text.

1	***2***	***3***	***4***	***5***
I	***go***	***to***	***the***	***university***
1	***5***	***3***		***2***
Men	***universitet***	***-ga***	***-***	***boraman***

Substitution proceeded through in two ways: 1) concretization; 2) generalization. These are proceeded in morphologic and synacticacts. For example, "**and"**– **(va, hamda** in Uzbek**)** is as connectable conjunction used the following functions: bo`lsa, esa, biroq – *I shall go* ***and*** *you stay here* –Men ketaman, sen bo`lsang shu yerda qol.

It comes as infinitive (harakat nomi) in Uzbek in the compound units of verbs, as well: *try* ***and*** *do it* – buni qilishga harakat qiling, *come* ***and*** *see* –ko`rishga keling, *wait* ***and*** *see* –yashasak ko`ramiz [4]. Generalization may often occur mainly among the syntactic units.

In replenishment process some morphologic categories are added by the position of speech. For instance, I have a book –Men**da** kitob bor. The affix **–da** is added to personal pronoun. It seems very easy if it is done by human. But for the linguistic database of translation it should be clarified accurately. In the process of omission one or several units are deleted in the context: demonstrative pronoun *those –ular, o`shalar, ana o`shalar; these –bular, shular* are used for plural noun, but sometimes as singular in Uzbek. ***Those children are mine –O`sha bolalar meniki***. In English we may see the agreement of those and children (plural), but in Uzbek it is not normal. That's why affix **–lar (**o`shalar**)** is omitted in this situation.

There is some evidence to suggest that machine translation system has contextual and strong grammatical database. It is important to input morphological classification with equivalencies in two languages (English-Uzbek). We remind that Krosslexicon has morphological classificatory which holds 115 groups of declinable words in electron dictionary [5]. The morphoclassificatory can make the word forms even if they do not exist in the dictionary.

We can obtain good results in this field in case grammatical peculiarities of the text are considered true. There are many problems in morphological level in bilingual program. Especially, it should be done formalization and modeling of linguistic database in Uzbek.

Translation program consists of three stages: languageprocessor which contains analyser and synthesizer; linguistic model which contains of the knowledge of grammar and semantics; associative procedure which expresses linguistic translation operation that is connective between declarative and procedural parts[6]. We may observe that the module as analyze->transfer->synthesis is proceeded in many machine translation systems. Analizator should be provided linguistically in the process of morphological analyze. Generally speaking, grammatical base of translation program has done in scientific researches. According to them there are following types of the grammar: 1) chain grammar (цепочечная грамматика); 2) component grammar (грамматика составляющих); 3) dependency grammar (грамматика зависимостей); 4) context-free grammar (контекстно-свободных грамматика); 5) lexical-functional grammar (лексико-функциональная грамматика); 6) unificational grammar (унификационные грамматика). "Chain grammar consists of words that belong to the groups of the terms (article+noun+preposition) and compounding units such as (subject+predicate) functional elements of terms. The order of units of speeches are shown in this. Component grammar providesthe group of grammar elements, for example, group of noun phrase (noun, article, adjective and other modifications), prepositional group (preposition +part of speech). In dependency grammar each elements are dependable each others. The strategy of analyze is as top-downand the center of sentence is predicate (verb). A transformational method is used in contextual grammar and it has shown above. Unification grammar consist of four components: suite of unification, interpreter for grammar rules and

description of the words, directed graphs of possessing program and analyzer with helping graph-devices. Unification grammar identifies semantic valency with synactic valency and description of dictionary with grammar rules [7]".

It is truly estimated by V.Rojdestvenskiy that central problem of artificial intelligence is machine translation [8].Because facilities of languageare are appeared by the influence of linguistic and nonlinguistic factors. Machine translation is complex physiological process. That's why by using contemporary methods of the schoolars, we have to build well-built linguistic database of translation program.

Sum up, powerful linguistic and programming database characterize the quality of machine translation. All grammar rules of the text must be investigated depending on types of style of the texts. It will be better if contextual dictionaries are created in English-Uzbek translation program.

References

1. Захаров В.П., Богданова С.Ю.Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. С. 39.
2. Бархударов Л. С. Язык и перевод (Вопросы общей и частной теории перевода). М., «Междунар. отношения», 1975.С.190-191.
3. Бўронов Ж.Б. Инглиз ва ўзбек тиллари қиёсий грамматикаси "Ўқитувчи" Т., 1973. 40 –бет.
4. ABBYLingvo×5
5. Большаков И. А., Большакова Е. И. Автоматический морфоклассификатор русских именных групп. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2012) Выпуск 11 С.81
6. Марчук Ю.Н.Компьютерная лингвистика М., Восток.2006.С. 272.
7. Мамедова М.Г., Мамедова З.Ю.Машинный перевод: эволюция и основные аспекты моделирования. Баку: Изд. «ИНФОРМАСИЙАТЕХНОЛОЭИЙАЛАРЫ», 2005. С. 69-72 .
8. Рождественский Ю.В., Волков А.А., Марчук Ю.Н. Введение в прикладную филологию, МГУ, 1988. С. 116.

Н.З. АБДУРАХМОНОВА, М.Х. ХАКИМОВ

Национальный Университет Узбекистана им. Мирзо Улугбек,
Ташкент, Республика Узбекистан

ЛОГИКО-ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ СЛОВ И ПРЕДЛОЖЕНИЙ АНГЛИЙСКОГО ЯЗЫКА ДЛЯ МНОГОЯЗЫЧНЫХ СИТУАЦИЙ КОМПЬЮТЕРНОГО ПЕРЕВОДА

Функциональность слова естественного языка (ЕЯ) проявляется в его многозначности. В конкретных случаях каждое слово свое конкретное значение приобретает в фразах и/или в предложениях [3]. Признание функциональности слова приводит к семантической однозначности, за исключением некоторых конкретных случаев, вытыкаемых из ЕЯ. Функциональность слова приводит к двум принципиально различным подходам при построении логико-лингвистических моделей ЕЯ – либо разработать единую систему линейной обработки слов и предложений, либо рассматривать каждое слово и предложение как единичную структуру, в соответствии с которой оно обрабатывается. Как в [3], также и здесь используем первый подход, выполнение которой обеспечивает перевод из языка **A** в язык **B,** относящихся к классу 0 по классификации Н. Хомского [4], когда математическая модель является

распознающей характеризующей язык **А** или порождающей характеризующий язык **В** [5], в многоязычной ситуации машинного перевода.

В свете вышеизложенного были проведены исследования над английским языком (АЯ) [2], являющегося как одним из языков системы машинного перевода для многоязычной ситуации и построены логико-лингвистические модели составления слов различных частей предложения. На основе лексического анализа проведенных над АЯ определяем, что слова делятся на четыре типа составляющих – корень, аффиксы образующие слова, аффиксы образующие форму и аффиксы изменяющие слова. Согласно этого строим общую логико-лингвистическую модель образования слова АЯ:

⇓предлог ⊕⇓префикс ⊕ корень ⊕⇓аффиксы образующие слова ⊕⇓аффиксы образующие форму ⊕ ⇓аффиксы изменяющие слова

Здесь, знаки означают: ⊕ - операцию присоединения, ⇓ - операцию “подключения” или “не подключения” следующей за ней составляющей. Эти знаки являются операциями расширяемого входного языка формирующих математические модели естественного языка при многоязычных ситуациях машинного перевода [6].

Логико-лингвистические модели слов

Логико-лингвистические модели вывода имен существительных АЯ имеет тринадцать типов. Логико-лингвистические модели вывода имен существительных с примерами приведена в табл. 1.

Таблица 1

№	Логико-лингвистические модели существительных	Пример
1.	корень существительное ⊕⇓суффикс	work⊕er
2.	корень существительное ⊕ корень существительное	sky⊕rocket
3.	корень существительное ⊕ корень существительное ⊕суффикс	snow⊕ball⊕s
4.	корень существительное ⊕ предлог ⊕ корень существительное	mother⊕ in ⊕law
5.	префикс ⊕корень существительное ⊕суффикс	dis⊕abili⊕ty
6.	местоимение ⊕существительное	he⊕wolf
7.	корень прилагательное ⊕ суффикс	difficult⊕y
8.	корень прилагательное ⊕ корень существительное	blue⊕bell
9.	корень глагол ⊕корень существительное	pick⊕pocket
10	корень глагол ⊕ предлог	break⊕out
11	корень глагол ⊕ суффикс	assist⊕ant
12	корень глагол ⊕суффикс ⊕⇓суффикс	use⊕less⊕ness
13	корень глагол ⊕корень местоимение⊕not	forget⊕ me⊕ not
14	корень существительное ⊕ предлог ⊕ корень существительное	commander ⊕in ⊕chief

Логико-лингвистические модели вывода прилагательных АЯ имеет двадцать типов. Логико-лингвистические модели вывода прилагательных с примерами приведена в табл. 2.

Таблица 2

№	Логико-лингвистические модели прилагательных	Пример
1.	корень прилагательное ⊕корень существительное ⊕ суффикс	long ⊕legg ⊕ed
2.	корень прилагательное ⊕ корень прилагательное	deaf ⊕mute
3.	корень прилагательное ⊕причастие I	good ⊕looking
4.	корень прилагательное ⊕причастие II	clean ⊕shaven
5.	корень прилагательное ⊕ предлог	hard ⊕up (student)
6.	префикс ⊕ корень прилагательное	un ⊕interesting

7.	корень глагол ⊕ суффикс	impress ⊕ ive
8.	корень глагол ⊕ предлог	run ⊕ dawn
9.	корень существительное ⊕ суффикс	home ⊕ less
10	корень существительное ⊕ корень существительное ⊕ d	doll ⊕ face ⊕ d
11	корень существительное ⊕ корень прилагательное	snow ⊕ white
12	корень существительное ⊕ причастие I	life ⊕ giving
13	корень существительное ⊕ корень существительное ⊕ суффикс ⊕ d	lyn ⊕ ed ⊕ eye ⊕ d
14	корень наречие ⊕ причастие II	under ⊕ bred
15	корень наречие ⊕ причастие I	well ⊕ seeming
16	корень наречие ⊕ прилагательное	off ⊕ black
17	корень наречие ⊕ корень существительное ⊕ суффикс ⊕ d	over ⊕ people ⊕ d
18	корень наречие ⊕ предлог	well ⊕ off
19	корень числительное ⊕ корень существительное ⊕ ed	four ⊕ wheel ⊕ ed
20	корень местоимение ⊕ предлог	all ⊕ out

Логико-лингвистические модели вывода глагола АЯ имеет девять типов. Логико-лингвистические модели вывода глагола с примерами приведена в табл. 3.

Таблица 3

№	**Логико-лингвистические модели глагола**	**Пример**
1.	корень глагол ⊕ корень прилагательное	catch ⊕ cold
2.	корень глагол ⊕ корень существительное	have ⊕ lunch
3.	корень глагол ⊕ предлог ⊕ корень глагол	fall ⊕ in ⊕ love
4.	корень глагол ⊕ корень глагол	take ⊕ interest
5.	префикс ⊕ корень глагол	be ⊕ little
6.	корень существительное ⊕ корень существительное	day ⊕ dream
7.	корень существительное ⊕ суффикс	fantas ⊕ ize
8.	to be ⊕ корень прилагательное	to be ⊕ ill
9.	корень прилагательное ⊕ суффикс	red ⊕ en

Логико-лингвистические модели вывода наречий АЯ имеет два типа. Логико-лингвистические модели вывода наречий приведена с примерами в табл.4.

Таблица 4

№	**Логико-лингвистические модели наречия**	**Пример**
1.	корень прилагательное ⊕ суффикс	calm ⊕ ly
2.	корень существительное ⊕ суффикс	clock ⊕ wise

Логико-лингвистические модели вывода местоимения АЯ имеет восемь типов. Логико-лингвистические модели вывода местоимения с примерами приведена в табл. 5

Таблица 5

№	**Логико-лингвистические модели местоимения**	**Пример**
1.	корень местоимения (possessive P) ⊕ суффикс	your ⊕ s
2.	корень местоимения ⊕ существительное	any ⊕ thing
3.	корень местоимения ⊕ корень числительного	every ⊕ one
4.	корень местоимения (interrogative P) ⊕ ever	which ⊕ ever
5.	корень местоимения (interrogative P) ⊕ preposition{of}	which ⊕ of

6.	корень местоимения (possessive P) ⊕ self	my ⊕ self
7.	корень местоимения (possessive P) ⊕ selves	our ⊕ selves

Логико-лингвистические модели вывода числительных АЯ имеет три типа. Логико-лингвистические модели вывода числительных с примерами приведена в табл. 6.

Таблица 6

№	**Логико-лингвистические модели числительных**	**Пример**
1.	корень числительного ⊕ ⇓ суффикс	four ⊕ ty
2.	корень числительного ⊕ ⇓суффикс ⊕ ⇓корень числительного	four ⊕ ty ⊕ five
3.	числительное ⊕ ⇓союз ⊕ ⇓числительное	hundred ⊕and ⊕one

Логико-лингвистические модели предложений

На АЯ простые повествовательные предложения можно строить на основе логико-лингвистической модели:

подлежащее ⊕ сказуемое ⊕ слова по типам ⊕ второстепенные члены предложения

Конкретные логико-лингвистические модели вывода повествовательных предложений имеют двенадцать типов:

1. существительное ⊕глагол ⊕ ⇓наречие
2. артикль ⊕ ⇓прилагательное ⊕существительное ⊕⇓глагол ⊕⇓прилагательное
3. артикль ⊕ ⇓прилагательное ⊕существительное ⊕⇓глагол ⊕⇓наречие
4. местоимение ⊕глагол ⊕⇓существительное ⊕⇓прилагательное
5. существительное ⊕глагол ⊕⇓местоимение ⊕⇓существительное
6. существительное ⊕глагол ⊕⇓прилагательное ⊕⇓существительное ⊕⇓наречие ⊕ ⇓местоимение ⊕⇓глагол
7. местоимение ⊕глагол ⊕⇓числительное ⊕⇓существительное ⊕⇓союз ⊕⇓прилагательное ⊕существительное
8. ⇓артикль ⊕существительное ⊕глагол ⊕⇓прилагательное ⊕⇓союз ⊕⇓прилагательное
9. ⇓местоимение ⊕модальный глагол ⊕глагол ⊕⇓наречие
10. существительное ⊕модальный глагол ⊕глагол ⊕наречие
11. модальное слово ⊕⇓местоимение ⊕глагол ⊕⇓союз ⊕⇓местоимение
12. предлог ⊕местоимение ⊕существительное ⊕глагол ⊕⇓числительное ⊕существительное

На АЯ вопросительные предложения можно вывести на основе десяти типов логико-лингвистических моделей:

1. вспомогательный глагол ⊕существительное ⊕⇓глагол ⊕⇓существительное
2. вспомогательный глагол ⊕существительное ⊕⇓местоимение ⊕⇓глагол ⊕⇓существительное
3. вспомогательный глагол ⊕существительное ⊕⇓существительное
4. вспомогательный глагол ⊕местоимение ⊕⇓существительное
5. модальный глагол ⊕ ⇓существительное ⊕ глагол ⊕⇓наречие
6. модальный глагол ⊕⇓местоимение ⊕ глагол ⊕⇓наречие
7. вспомогательное слово ⊕ёрдамчи глагол ⊕⇓существительное ⊕⇓местоимение ⊕ глагол

8. вспомогательный глагол ⊕ существительное ⊕ ⇓ существительное ⊕ ⇓ союз ⊕ ⇓ существительное
9. вспомогательный глагол ⊕ местоимение ⊕ ⇓ существительное ⊕ ⇓ союз ⊕ ⇓ существительное
10. ⇓ существительное ⊕ ⇓ местоимение ⊕ глагол ⊕ ⇓ существительное ⊕ вспомогательный глагол ⊕ ⇓ местоимение

На АЯ восклицательные предложения можно вывести на основе следующих двенадцати типов логико-лингвистических моделей:

1. существительное ⊕ вспомогательный глагол ⊕ глагол
2. существительное ⊕ ⇓ существительное ⊕ вспомогательный глагол ⊕ ⇓ глагол
3. вспомогательный глагол ⊕ ⇓ местоимение ⊕ союз ⊕ глагол ⊕ ⇓ союз ⊕ ⇓ артикль ⊕ ⇓ существительное
4. глагол ⊕ ⇓ артикль ⊕ существительное
5. глагол ⊕ ⇓ артикль ⊕ ⇓ существительное ⊕ модальное слово
6. вспомогательное слово ⊕ ⇓ прилагательное ⊕ существительное ⊕ ⇓ местоимение ⊕ вспомогательный глагол ⊕ ⇓ глагол
7. вспомогательное слово ⊕ ⇓ прилагательное ⊕ ⇓ артикль ⊕ ⇓ существительное ⊕ вспомогательный глагол ⊕ ⇓ предлог ⊕ ⇓ существительное
8. вспомогательное слово ⊕ ⇓ наречие ⊕ ⇓ местоимение ⊕ глагол
9. вспомогательное слово ⊕ ⇓ прилагательное ⊕ ⇓ союз ⊕ ⇓ прилагательное ⊕ ёрдамчи глагол ⊕ ⇓ местоимение ⊕ ⇓ существительное
10. вспомогательное слово ⊕ ⇓ артикль ⊕ ⇓ прилагательное ⊕ существительное
11. глагол ⊕ ⇓ артикль ⊕ ⇓ существительное ⊕ ⇓ существительное
12. ⇓ артикль ⊕ существительное ⊕ ⇓ предлог ⊕ ⇓ артикль ⊕ существительное ⊕ ⇓ существительное

Отрицательные предложения можно вывести на основе шести типов логико-лингвистических моделей:

1. местоимение ⊕ вспомогательный глагол ⊕ ундалма ⊕ глагол ⊕ ⇓ предлог ⊕ ⇓ артикль ⊕ ⇓ существительное ⊕ ⇓ предлог ⊕ ⇓ артикль ⊕ ⇓ местоимение
2. существительное ⊕ вспомогательный глагол ⊕ ундалма ⊕ глагол ⊕ ⇓ предлог ⊕ ⇓ артикль ⊕ ⇓ существительное ⊕ ⇓ предлог ⊕ ⇓ артикль ⊕ ⇓ местоимение
3. местоимение ⊕ вспомогательный глагол ⊕ ⇓ частица ⊕ ⇓ глагол ⊕ ⇓ местоимение
4. существительное ⊕ вспомогательный глагол ⊕ ⇓ частица ⊕ ⇓ глагол ⊕ ⇓ местоимение
5. местоимение ⊕ модальный глагол ⊕ обращение ⊕ глагол ⊕ предлог ⊕ местоимение
6. существительное ⊕ модальный глагол ⊕ обращение ⊕ глагол ⊕ предлог ⊕ местоимение

Литература

1. Абдурахмонова Н.З. Машина таржимасининг лингвистик асослари. “Академнашр”, Тошкент, 2012.
2. Кобрина Н.А., Корнеева Е.А., Оссовская М.И., Гузеева К. А. Грамматика английского языка. Морфология. Синтаксис. Санкт-Петербург, Издательство «Союз», 2008.
3. Хакимов М.Х. К моделям естественных языков для многоязычных ситуаций компьютерного перевода. Труды научной конференции «Проблемы современной математики» 22-23 апреля 2011 г., г. Карши, с.531-537

4. Хомский Н. Формальные свойства грамматик. «Кибернетический сборник», НС, вып. 2, 1966, стр. 121-230.

5. Хакимов М.Х. Формальные системы машинного перевода в многоязычной ситуации. Материалы республиканской научной конференции «Современные проблемы математики, механики и информационных технологий», НУУз, Институт Математики и ИТ АН РУз, Т, 2008, с.297-301

6. Хакимов М.Х. Расширяемый входной язык математического моделирования естественного языка для многоязычной ситуации машинного перевода. Ташкент, «ЎзМУ хабарлари», № 1, 2009, с. 75-80

М.А. БОЛАТБЕК, А. МАРАТҚЫЗЫ, Л.Р. МҰСАЕВА

Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

ҚАЗАҚША-ОРЫСША МАШИНАЛЫҚ АУДАРМАДАҒЫ МОРФОЛОГИЯЛЫҚ ЖӘНЕ СИНТАКСИСТІК АНАЛИЗ МЕН СИНТЕЗ АЛГОРИМТДЕРІ

Бұл мақала машиналық аударманың жұмыс істеуі барысында орындалатын морфологиялық және синтаксистік анализ бен синтез алгоритмдерінің орындалу ережелері тақырыбында жазылған. Қазақша-орысша машиналық аударманың жұмыс істеу үлгісіне сәйкес алдымен қазақ тілінде енгізілген сөзге түрлі талдаулар (морфологиялық, синтаксистік) жасалып, олардың шығыс мәліметтері бойынша енгізілген сөз түбірі орыс тіліне аударылып, өзгертілетін болады. Сөзге жасалатын морфологиялық талдау дегеніміз – қазақ тілінде енгізілген сөздің түбірі мен қосымшасын ажырату. Біздің ұсынған алгоритміміз бойынша морфологиялық талдау екі бағыттан жүргізіледі, яғни деректер базасынан сол жақтан түбір ізделінетін болса, сол мезетте оң бағыттан сөз қосымшасына іздеу жүргізіліп жатады. Мысалы, кіріске «мектептер» сөзі келіп түсетін болса, ол сөздің деректер қорынан ізделуін келесі түрде көрсетуге болады (1-сурет) :

Ізделінетін түбір **Ізделінетін қосымшалар**

М	**е**	**к**	**т**	**е**	**п**	**т**	**е**	**р**			
М	**е**	**к**	**т**	**е**	**п**	**т**	**е**				**р**
М	**е**	**к**	**т**	**е**	**п**	**т**				**е**	**р**
М	**е**	**к**	**т**	**е**	**п**				**т**	**е**	**р**
М	**е**	**к**	**т**	**е**				**п**	**т**	**е**	**р**
М	**е**	**к**	**т**				**е**	**п**	**т**	**е**	**р**
М	**е**	**к**				**т**	**е**	**п**	**т**	**е**	**р**
М	**е**				**к**	**т**	**е**	**п**	**т**	**е**	**р**

М			**е**	**к**	**т**	**е**	**п**	**т**	**е**	**р**
		м	**е**	**к**	**т**	**е**	**п**	**т**	**е**	**р**

Сурет 1. Түбір мен қосымшаны іздеу алгоритмі.

Суреттен көріп тұрғанымыздай, алдымен деректер базасынан «мектептер» сөзіне іздеу жүргізіледі. Деректер базасында аталған сөз болмағандықтан іздеу жұмысы әрі қарай жалғасады. Келесі қадамда деректер базасынан «мектепте» деген түбір мен «р» қосымшасына іздеу жүргізіледі. Деректер базасында «р» қосымшасы болғанымен, «мектепте» деген түбір жоқ, сол себепті іздеу жұмысы тағы да жалғасады. Осылайша түбір сөздер мен қосымшалар кестесінен қажетті әрі сәйкес нұсқа табылған жағдайда («мектеп» түбірі мен «тер» қосымшасы) іздеу жұмысы тоқтатылады.

Түбір сөз – мектеп, зат есімдер кестесінде орналасқан (2-сурет):

RecNo	id	kaz	rus	rod	koptik	skl
Click here to define a filter						
3606	3608	мекен-жай	адрес	1	адреса	2
3607	3609	мексикалықтар	мексиканцы	0	мексиканцы	<null>
3608	3610	мектеп	школа	2	0	1
3609	3611	мектеп-интернат	школа-интернат	2	школы-интернаты	1
3610	3612	мемлекет	государство	3	0	2
3611	3613	мемлекеттілік	государственность	2	0	3

Сурет 2. Деректер базасынан табылған сөз түбірі.

Ал «тер» жалғауы қосымшалар кестесінде орналасқан (3-сурет):

RecNo	id	zhalgau	s_e	koptik	septik	taueldik	jiktik	shak
Click here to define a filter								
553	553	тен	1	1	6	0	0	0
554	554	тер	1	2	1	0	0	0
555	555	терге	1	2	3	0	0	0
556	556	терде	1	2	5	0	0	0
557	557	терден	1	2	6	0	0	0
558	558	терді	1	2	4	0	0	0

Сурет 3. Деректер базасынан табылған сөз қосымшасы.

Іздеу жұмысы тоқтатылғаннан кейін сөз түбірі мен қосымшасының атрибуттары анықталады. Зат есімдер кестесінде rod, koptik, skl атрибуттары бар. Rod атрибуты сөз аудармасының орыс тіліндегі қай «родқа» тиістілігін анықтау үшін тағайындалды (0 - сөз аудармасы үш «родқа» да тиісті емес, 1- сөз аудармасы «мужской родқа» тиісті, 2 – сөз аудармасы «женский родқа», 3 – сөз аудармасы «средний родқа» тиісті). Koptik атрибуты көптік формасы ережеге бағынбайтын, ерекше жағдайлы сөздерге арналған. Егер сөздің орыс тіліндегі аудармасы көптік формасын қабылдау барысында ережеге бағынбайтын болса, оның дұрыс көптік формасын аталған бағанға орналастырдық (мысалы, адамдар – ереже бойынша **человеки** болуы тиіс, бұл дұрыс болмағандықтан, сөздің көпше түрі **люди** сөзін

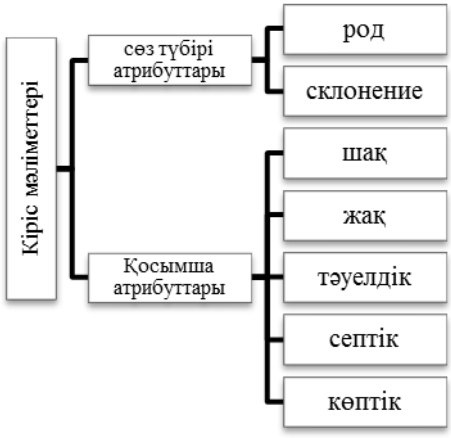
енгіздік. Ережеге сәйкес көпше түрге енетін сөздердің аталған атрибуты 0 мәнін қабылдайды.

Сөз қосымшасының атрибуттарын сипаттайтын болсақ, суреттен көріп тұрғанымыздай, бұл кесте жазбаларына s_e, koptik, taueldik, jiktik, septik және shak атрибуттары тағайындалған. Олардың мәндерін төмендегі кестеден көруге болады (1-кесте):

Кесте 1. Сөз қосымшасының атрибуттары.

Атрибут	S_e	koptik	jiktik	taueldik	Shak	septik
0	-	Көптік жалғауы жалғанбай-ды	Жіктік жалғауы жалғанбай-ды	Тәуелдік жалғауы жалғанбай-ды	Шақтық жұрнақ жалған-байды	Септік жалғауы жалған-байды
1	Зат есімге жалғанады	Жекеше түр	1-жақ	1-жақ	Өткен шақ	Атау септік
2	Етістікке жалғанады	Көпше түр	2-жақ	2-жақ	Осы шақ	Ілік септік
3	-	-	3-жақ	3-жақ	Келер шақ	Барыс септік
4	-	-	-	-	-	Табыс септік
5	-	-	-	-	-	Жатыс септік
6	-	-	-	-	-	Шығыс септік
7	-	-	-	-	-	Көмектес септік

Қазақ тілінде енгізілген сөйлемге жоғарыдағы қадамдар бойынша анализ жасалғаннан кейін оларды орыс тіліне аудару мақсатында морфологиялық және синтаксистік синтездер жасалады. Морфологиялық синтез енгізілген сөздің орыс тіліндегі аудармасы түбірінің атрибуттары мен қазақ тіліндегі қосымшалар атрибуттары негізінде жасалады. Морфологиялық синтез барысында қазақ тілінде анықталған сөз қосымшаларының (көптік, септік, жіктік және тәуелдік) орыс тіліндегі сәйкестіктерін орнату арқылы сөз түбірінің аудармасы өзгертіледі. Морфологиялық синтездің кіріс мәліметтері: морфологиялық талдау нәтижесінде алынған сөз түбірі мен қосымшалардың атрибуттары (4-сурет):



Сурет 4. Морфологиялық синтездің кіріс мәліметтері.

Шығыс мәліметі: морфологиялық үлгі мен алгоритмдерге сәйкес орыс тіліне аударылған сөз.

Мысал ретінде септік жалғаулары жалғанған қазақ тіліндегі зат есімдерді орыс тіліне аудару алгоритмдерін көрсетейік:

Ілік септігі. 1-склонение. Ілік септігі орыс тіліндегі «родительный падежге» сәйкес келеді. Егер сөз аудармасының соңғы әрпі «а», ал соңғы әріптің алдындағы әріп pb[] = { 'б', 'в', 'д', 'з', 'л', 'м', 'н', 'п', 'р', 'с', 'т', 'ф', ‘ц’} әріптерінің біріне тең болса, сөз аудармасына «ы» қосымшасы жалғанады, мысалы көліктің =>машины. Егер сөз аудармасының соңғы әрпі «я» болса, онда аударманың соңғы әрпі «ю» болып өзгертіледі, мысалы жердің=>земли.

2-склонение. Егер сөз аудармасы «мужской родқа» тиісті болса, онда «родительный падежде» сөз соңына «а» қосымшасы жалғанады, мысалы үй**дің**=>дом**а**. Егер сөз аудармасы «мужской родқа» тиісті болса және соңғы әріп «ь» болса, онда сөздің соңғы әрпі «я» әрпімен алмастырылады, мысалы жылқы**ның** =>кон**я**. Егер сөз аудармасы «средний род» категориясына тиісті болса, онда морфологиялық синтез жасау барысында сөздің соңғы әрпі «а» жалғауымен алмастырылады, мысалы күн**нің** =>солнц**а.**

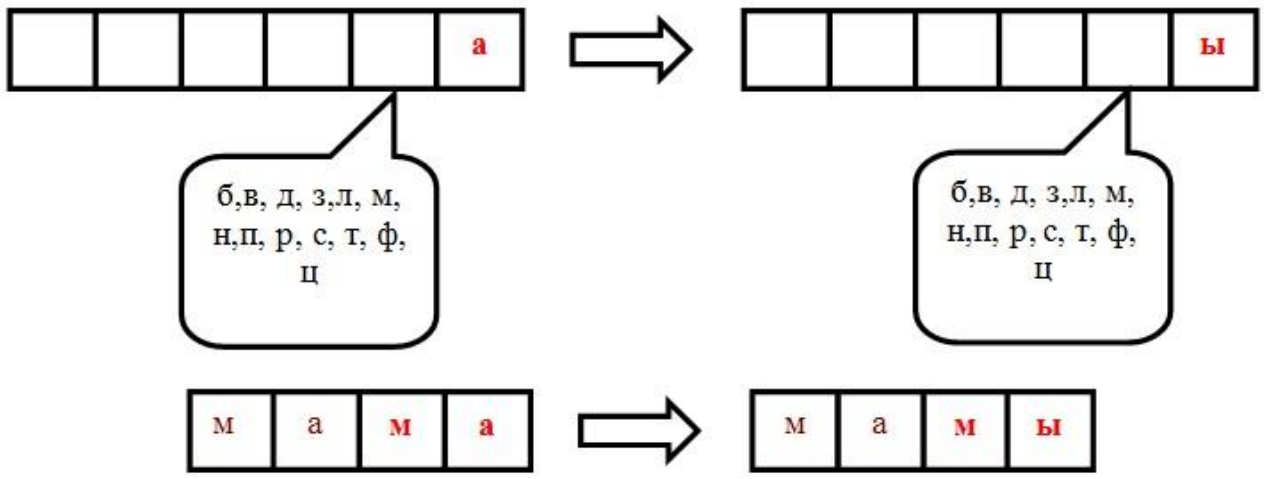
3-склонение. Егер сөз аудармасы «женский род» категориясына тиісті болса және оның соңғы әрпі «ь» болса, онда «родительный падежде» ол сөз аудармасының соңғы әрпі «и» жалғауымен алмастырылады, жол**дың** =>пут**и**

–ы және – и қосымшаларын таңдау әдісі (бірінші склонение мысалы):

Деректер базасындағы padezh=2, chislo=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

defineVariables.skl – орыс тіліндегі сөздің склонениесі;
defineVariables.words_rus – орыс тіліндегі сөз;
padezh – орыс тіліндегі сөздің септігі;

```
string[] string1 = new string[] {"б", "в", "д", "з", "л", "м", "н", "п", "р", "с", "т", "ф", "ч",  "ц" };
   if (defineVariables.skl[defineVariables.i] == "1")
   {  if (padezh == "2")
{if(string1.Contains(defineVariables.words_rus[defineVariables.i].Substring(defineVariables.
words_rus[defineVariables.i].Length - 2, 1)))
 {                    defineVariables.words_rus[defineVariables.i]                    =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1; }
else if (commonFunctions.lastLetter(defineVariables.words_rus[defineVariables.i], 2) ==
"ка"){   defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 2) + isk1;
 }   else{defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch2;}
```



Барыс септігі. Орыс тіліндегі «дательный падежге» сәйкес келеді. 1-склонение. Егер сөз аудармасы «дательный падежге» тиісті болса, сөз аудармасына «е» қосымшасы жалғанады және сөз алдына «к» предлогы жалғанады, мысалы көлік**ке** =>**к** машин**е.**

2-склонение. Егер сөз аудармасы «мужской род» немесе «средний родқа» тиісті болса, онда аударманың қосымшасы «у» қосымшасымен алмастырылады және сөз алдына «к» предлогы жалғанады, мысалы сынып**қа** =>**к** класс**у**, күн**ге** =>солнц**у.** Егер с-з аудармасы «мужской родқа» тиісті болса және оның соңғы әрпі «ь» болса, онда сөзге «и» қосымшасы жалғанады.

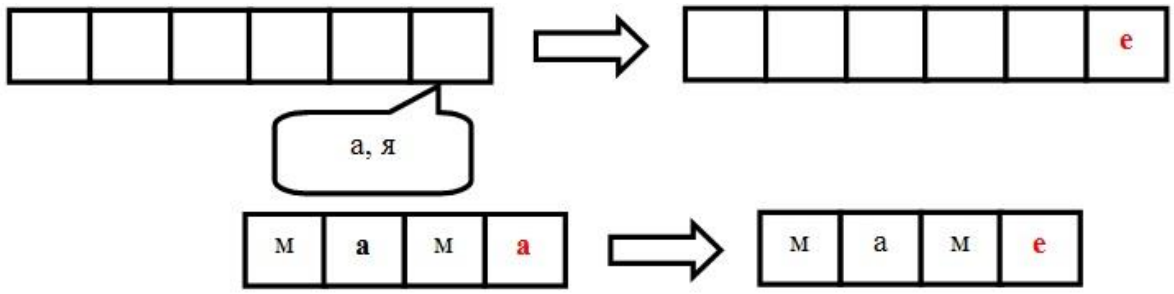
3-склонение. Егер сөз аудармасы «женский род» категориясына тиісті болса және оның соңғы әрпі «ь» болса, онда қосымша «и» жалғауымен алмастырылады және сөз алдына «к» предлогы жалғанады, мысалы жол**ға** =>**к** пут**и.**

-е қосымшасын жалғау әдісі:

Деректер базасындағы padezh=3, chislo=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

```
if (padezh == "3")
  {        defineVariables.words_rus[defineVariables.i]  =  "к  "  +
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1;  }
```

Мысалы:



Табыс септігі. Орыс тіліндегі «винительный падежге» сәйкес келеді. 1-склонение. Егер сөз аудармасы «винительный падежде» тұрса және қосымша «а», ал соңғы әріптің алдындағы әріп pb[]={ 'б', 'в', 'г', 'д', 'ж', 'з', 'к', 'л', 'м', 'н', 'п', 'р', 'с', 'т', 'ф', 'ч', 'ш', 'щ', 'х', ‘ц’} әріптерінің біріне тең болса, сөз аудармасына «у» жалғауы жалғанады, мысалы көлік**ті** =>машин**у**. Егер сөздің соңғы әрпі «я» болса, онда аудару барысында сөздің қосымшасы «ю» болып өзгереді, мысалы жерд**і** =>земл**ю.**

2-склонение. Егер сөз аудармасы мужской родқа тиісті болып, оның соңғы әрпі «ь» болса және сөз аудармасы средний родқа тиісті болып, оның соңғы екі әрпі «ие», «ре», «ле» болса, сөз қосымшасы «я»-ға ауысады, қалған жағдайларда «а» жалғауы жалғанады, мысалы терезені =>окна, теңізді =>моря.

3-склонение. Егер сөз аудармасы орыс тіліндегі үшінші склонениеге сәйкес келсе, онда «винительный падежде» сөз түбірінің өзі шығарылады, мысалы жолд**ы** =>путь[5].

-у және –ю қосымшаларын жалғау әдісі:

Деректер базасындағы padezh=4, chislo=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

```
      string[] string1 = new string[] { "я" };
         if (defineVariables.skl[defineVariables.i] == "1")
         {  if (padezh == "4")
      {if(string1.Contains(defineVariables.words_rus[defineVariables.i].Substring(defineVariables.
words_rus[defineVariables.i].Length - 2, 1)))
       {                    defineVariables.words_rus[defineVariables.i]                =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch2; }
```

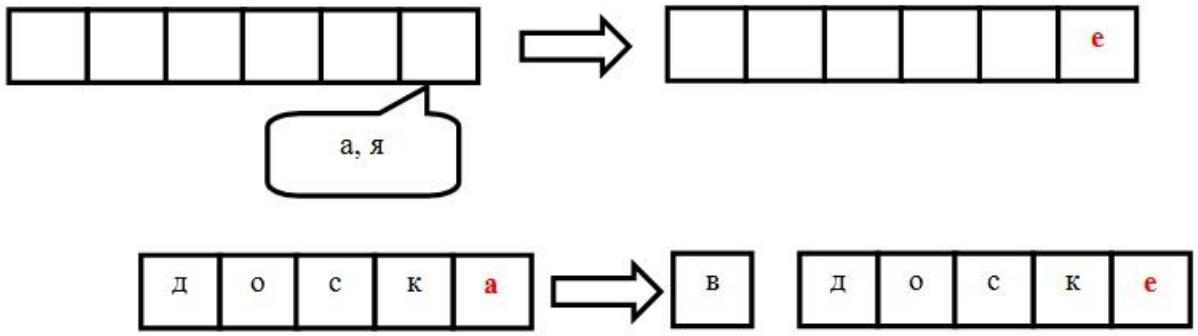
else { defineVariables.words_rus[defineVariables.i] = defineVariables.words_rus[defineVariables.i].Substring(0, defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1; }

Жатыс септігі. Орыс тіліндегі «предложный падежге» сәйкес келеді және аударманың алдына «на» немесе «в» предлогтарының бірі қойылуы тиіс. Егер сөз «мужской род», «женский род» немесе «средний родқа» тиісті болса, онда сөз түбірінің аудармасына «е» жалғауы жалғанады да, сөз алдына «в» немесе «на» предлогы қойылады. Мысалы көлік**те** =>**в** машин**е**, үйд**е**=>**в** дом**е**.

-е қосымшасын және «в» предлогын жалғау әдісі:

Деректер базасындағы padezh=6, chislo=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

if (padezh == "6")
{ defineVariables.words_rus[defineVariables.i] = "в "+ defineVariables.words_rus[defineVariables.i].Substring(0,defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1; }



Шығыс септігі. Орыс тіліне аударылу барысында «родительный падежге» сәйкес аударылады, тек қана аударманың алдына «из» немесе «от» предлогы жалғанады.

1-склонение. Егер сөз аудармасының соңғы әрпі «а», ал соңғы әріптің алдындағы әріп pb[] = { 'б', 'в', 'д', 'з', 'л', 'м', 'н', 'п', 'р', 'с', 'т', 'ф', ‘ц’} әріптерінің біріне тең болса, сөз аудармасына «ы» қосымшасы жалғанады және «из/от» предлогы қойылады, мысалы көліктен => из машин**ы**. Егер сөз аудармасының соңғы әрпі «я» болса, онда аударманың соңғы әрпі «ю» болып өзгертіледі және «из/от» предлогы қойылады, мысалы автоматтандырудан=> из автоматизаци**и.**

2-склонение. Егер сөз аудармасы «мужской родқа» тиісті болса, онда сөз соңына «а» қосымшасы жалғанады және сөз алдына «из/от» предлогы қойылады, мысалы үйд**ен**=>из дом**а**. Егер сөз аудармасы «мужской родқа» тиісті болса және соңғы әріп «ь» болса, онда сөздің соңғы әрпі «я» әрпімен алмастырылады және «из/от» предлогы қойылады, мысалы ажыратқышт**ан** => из выключател**я**. Егер сөз аудармасы «средний род» категориясына тиісті болса, онда морфологиялық синтез жасау барысында сөздің соңғы әрпі «а» жалғауымен алмастырылады және «из/от» предлогы қойылады, мысалы терезе**ден** => из окн**а.**

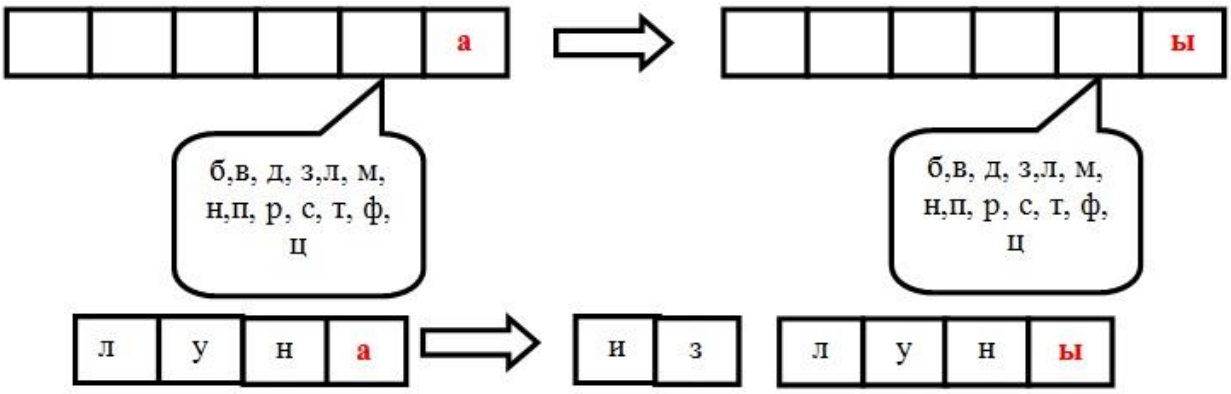
3-склонение. Егер сөз аудармасы «женский род» категориясына тиісті болса және оның соңғы әрпі «ь» болса, онда сөз аудармасының соңғы әрпі «и» жалғауымен алмастырылады және «из/от» предлогы қойылады, жолд**ан** =>из пут**и.**

-ы немесе -и қосымшасы мен «из» предлогын таңдау әдісі:

Деректер базасындағы padezh=2, koptik=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

if (padezh == "2")
{if(string1.Contains(defineVariables.words_rus[defineVariables.i].Substring(defineVariables.words_rus[defineVariables.i].Length - 2, 1)))

{defineVariables.words_rus[defineVariables.i]
=defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1; }
else { defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,defineVariables.words_rus[defineVariables.i].Length - 1) + okonch2; }
if (defineVariables.septik[defineVariables.i] == "6")
{defineVariables.words_rus[defineVariables.i]="из"+defineVariables.words_rus[defineVariables.i]; } }



Көмектес септік. Орыс тіліндегі «творительный падежге» сәйкес келеді.

1-склонение. Егер орыс тіліндегі сөз «творительный падеж» формасында тұрса және оның соңғы әрпі «а», ал соңғы әріптің алдындағы әріп pb[]={ 'б', 'в', 'г', 'д', 'ж', 'з', 'к', 'л', 'м', 'н', 'п', 'р', 'с', 'т', 'ф', 'ч', 'ш', 'щ', 'х'} әріптерінің біріне тең болса , онда сөзге «ой»жалғауы жалғанады, мысалы көлік**пен** =>машин**ой**. Егер сөздің соңғы әрпі «ь» болса, онда морфологиялық синтез барысында сөзге «ей» жалғауы жалғанады, мысалы жер**мен** =>земл**ей** (8-сурет)

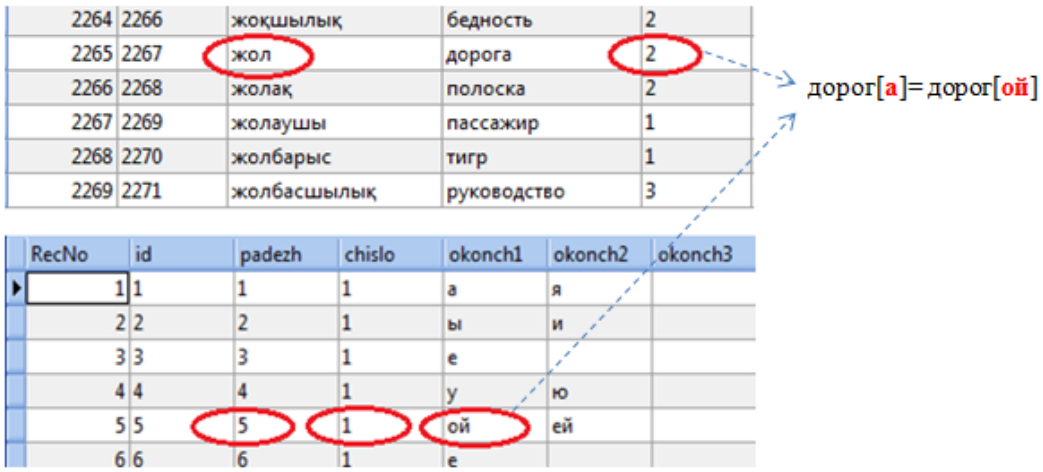
2- склонение. Егер сөз аудармасы «мужской родқа» тиісті болса, синтез жүргізу барысында сөз түбіріне «ом» жалғауы жалғанады, мысалы үй**мен** =>дом**ом**. Егер сөз аудармасы «мужской родқа» тиісті болса және оның соңғы әрпі «ь» болса немесе сөз аудармасы «средний родқа» тиісті болса, онда сөз түбіріне «ем» жалғауы жалғанады, мысалы күн**мен** =>солнц**ем**.

3-склонение. Егер сөз аудармасы орыс тіліндегі үшінші склонениеге сәйкес келсе, онда бұл септікте сөз соңына «ю» жалғауы жалғанады, мысалы жол**мен** =>путь**ю**.

«ой» немесе «ей» қосымшаларын жалғау әдісі:

Деректер базасындағы padezh=5, chislo=1, rod=2 болатын орыс тіліндегі жалғауға сұраныс жасалады:

if (padezh == "5")
{if(defineVariables.words_rus[defineVariables.i].Substring(defineVariables.words_rus[defineVariables.i].Length - 1, 1) == "я")
{defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,
defineVariables.words_rus[defineVariables.i].Length - 1) + okonch2;
} else
{defineVariables.words_rus[defineVariables.i] =
defineVariables.words_rus[defineVariables.i].Substring(0,defineVariables.words_rus[defineVariables.i].Length - 1) + okonch1; }}



Сурет 5. Көмектес септігінің мысалы.

Жоғарыдағы мысалдарда орыс тіліндегі бірінші «склонениеге» тиісті сөздердің септелуі келтірілген, өзге «склонение» сөздерінің де өздеріне тән аударылу ережелері бар және олар сол бойынша аударылады.

Қазақ тілінде енгізілген сөзге морфологиялық анализ бен синтез жасалғаннан кейін ол сөздердің қай сөйлем мүшесі болатындағын анықтау қажет болады, яғни сөйлемге синтаксистік талдау жасалынады. Синтаксистік талдау – сөйлемді, оның мүшелерін, түрлерін, құрмалас сөйлемнің құрамындағы жай сейлемдерді анықтау. Синтаксистік талдауда сөйлем мүшелері келесі ережелер бойынша анықталады:

Енгізілген сөз бастауыш болады, егер берілген сөз зат есім немесе есімдік болып және ол атау септігінде тұрса;

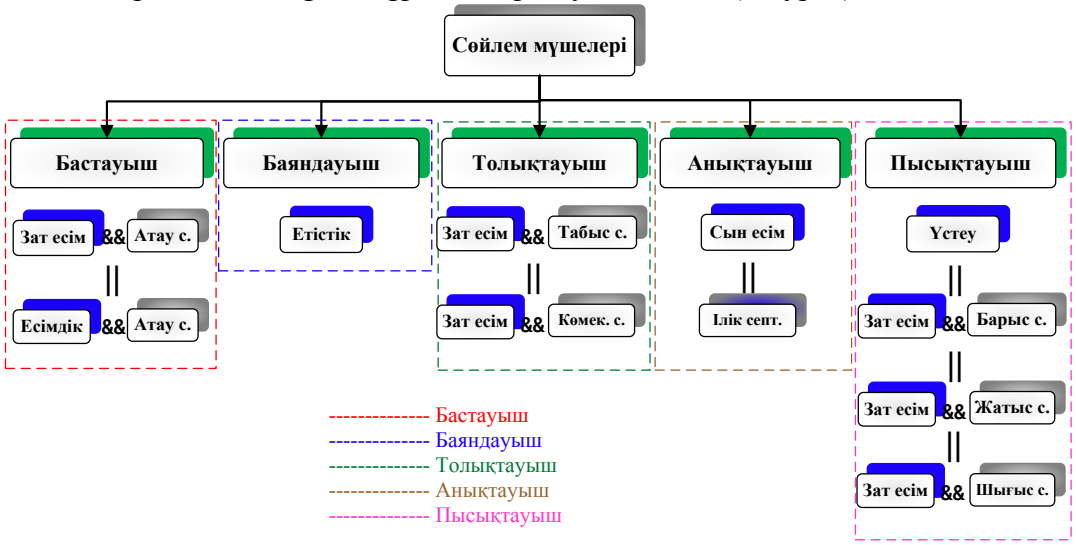
Енгізілген сөз баяндауыш болады, егер енгізілген сөз етістік болса;

Енгізілген сөз толықтауыш болады, егер берілген сөз зат есім болып және ол табыс немесе көмектес септігінде берілсе;

Енгізілген сөз анықтауыш болады, егер берілген сөз ілік септігінде тұрса немесе сын есім болса;

Енгізілген сөз пысықтауыш болады, егер сөз табы үстеу болса немесе берілген сөз зат есім болып және ол барыс, жатыс, шығыс септіктерінің бірінде тұрса;

Осы аталғандарды келесі сұлба түрінде көрсетуге болады (6-сурет):



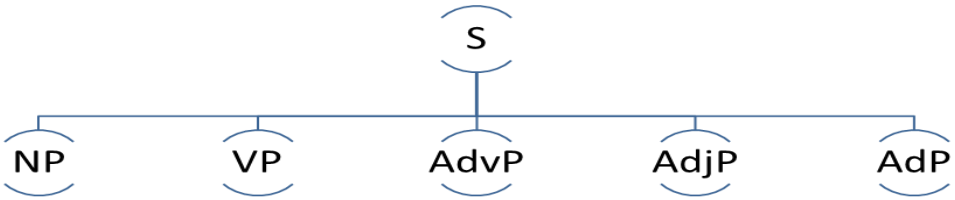
Сурет 6. Сөйлем мүшелерінің анықталуы.

Қазақ тіліндегі сөйлем мүшелері анықталғаннан кейін, оларды орыс тіліндегі сөйлем мүшелеріне сәйкестендіру керек болады. Қазақ (орыс) тілдеріндегі сөйлем мүшелеріне келесі атрибуттарды тағайындадық:

1 – бастауыш (подлежащее);
2 – баяндауыш (сказуемое);
3 – толықтауыш (дополнение);
4 – анықтауыш (определение);
5 – пысықтауыш (обстоятельство);

Осылайша синтаксистік анализдің жұмысы аяқталады да, келесі қадам синтаксистік синтезге беріледі. Синтаксистік синтез – синтаксистік талдау жасалған сөйлемдегі сөйлем мүшелерін дұрыс ретпен орналастыру. Қазақша-орысша машиналық аударманың үлгілерін құрастыру тарауында айтылғандай, екі тілдегі сөйлем мүшелері нөмірленген болатын. Тілдегі сөйлемді құрауыштар арқылы формальді грамматика түрінде көрсететін болсақ (7-сурет):

- сөйлем (sentence-S)
- бастауыш тобы (noun phrase-NP)
- баяндауыш тобы (verb phrase-VP)
- пысықтауыш тобы (adverbial phrase-Adv.P)
- анықтауыш тобы (adjectival phrase-Adj.P)
- толықтауыш тобы (addition phrase-AdP)



Сурет 7. Сөйлем құрауыштары.

Тілдегі сөйлем құрылымын келесі формальді грамматика түрінде көрсетуге болады:

<сөйлем>::= <бастауыш><баяндауыш><толықтауыш><анықтауыш><пысықтауыш>

<бастауыш> ::= <атау септігіндегі зат есім> |<есімдік >

<баяндауыш> ::= <етістік> |<күрделі етістік >

<толықтауыш> ::= <табыс септігіндегі зат есім> |<табыс септігіндегі есімдік>|<көмектес септігіндегі зат есім>

<анықтауыш> ::= <сын есім> |<ілік септігіндегі зат есім>

<пысықтауыш> ::= <үстеу>|<барыс септігіндегі зат есім>|<жатыс септігіндегі зат есім>|<шығыс септігіндегі зат есім>

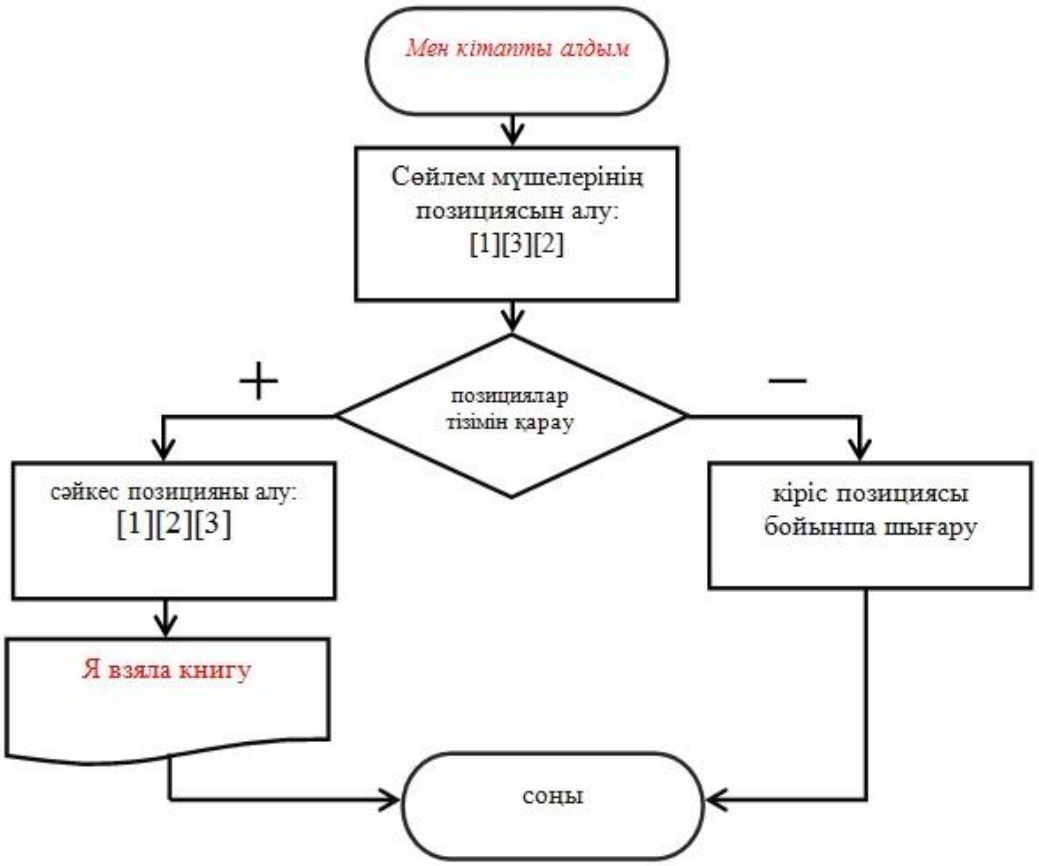
Сурет 8. Синтаксистік синтездің кіріс және шығыс мәліметтері.

Синтаксистік синтез барысында белгілі бір құрылымдағы қазақ тілінде енгізілген сөйлемнің орыс тіліндегі дұрыс құрылымы тағайындалады, яғни бұл жерде синтаксистік анализ барысында тағайындалған сөйлем мүшелерінің атрибуттары (нөмірлері) пайдаланылады. Сөйлем мүшелерінің сөйлемдегі орындарының реті келесі сәйкестікте болады, теңдіктің сол жағында қазақ тіліндегі сөйлем құрылымы, ал теңдіктің оң жағында қазақ тіліндегі сөйлем құрылымына сәйкес келетін орыс тіліндегі сөйлем құрылымы орналасқан:

[1][2]=[1][2]
[3][2]=[3][2]
[1][3][2]=[1][2][3]
[1][4][2]=[1][2][4]
[1][5][2]=[1][2][5]
[3][1][2]=[3][2][1]
[1][5][4][3][2]=[1][5][2][4][3]
[1][4][3][5][2]=[1][4][3][5][2]
[1][3][4][5][2]=[1][2][3][4][5]
[1][4][5][3][2]=[1][2][4][5][3]
[1][4][3][5][2]=[1][4][3][5][2]
[1][4][3][4][5][2]=[1][4][3][4][5][2]

Бұл жерде қазақ тіліндегі алты сөзге дейінгі жиі кездесетін сөйлем құрылымдары жинақталған, егер машина енгізілген құрылымды бұл тізімнен таппаса, онда шығысындағы сөйлем кіріске келіп түскен позиция бойынша шығарылады (орыс тілінде сөйлем мүшелерінің реті орнықты болуы шарт емес). Қазақ тіліндегі сөйлемге синтаксистік талдау жасалып, олардың сөйлемдегі орналасу реттері анықталғаннан кейін жоғарыда келтірілген сәйкестік бойынша орыс тіліндегі сөйлем құрастырылады. Мысалы, қазақ тіліндегі [1][3][2] құрылымына сәйкес келетін «Мен кітапты алдым» сөйлемі енгізілетін болса, ондағы сөздердің қай сөз таптарына жататындықтары анықталады. Ендігі кезекте ол сөз таптарының қай сөйлем мүшесіне сәйкес келетіндіктері жоғарыда келтірілген формальді грамматика арқылы анықталады (9-сурет).

Кіріс: «Мен кітапты алдым», құрылымы [1][3][2]



9-сурет. Синтаксистік синтез алгоритмі

Қазақ тілінде енгізілген сөзге жоғарыда келтірілген талдаулар (морфологиялық, синтаксистік) және оларға сәйкес синтездер (морфологиялық және синтаксистік) жасалу арқылы мәтінді орыс тіліне аудару орындалды. Аталған алгоритмдерді қолданбас бұрын оларға ұзақ зерттеу жұмыстары жүргізіліп, олардың ішіндегі тиімділері таңдалып алынды. Зерттеу жұмысы барысында Баскаков Н.А., Хасенова А.К., Исенгалиева В.А., Кордабаев Т.Р. авторларының бірігіп жазған «Қазақ және орыс тілдерінің салыстырмалы грамматикасы» кітабы қолданылды. Осы алгоритмдер қазақша-орысша машиналық генераторының программасын жазу барысында C# ортасында жүзеге асырылды. Жұмыс нәтижесі болып табылатын қазақша-орысша машиналық аударма программасы әзірге қарапайым сөйлемдерді сапалы түрде аудара алады. Аталған өнімді аудармашы көмегіне жүгінген кез-келген қолданушы пайдалана алады, ол үшін интернет желісінің болуы жеткілікті.

Қазіргі таңда аударма ісі еш жүйеде толықтай автоматтандырылмаған. Оның негізгі себебі – аударма процесін компьютер көмегімен үлгілеу, яғни адам аудару барысында түйсікке жүгінетін болса, машина ойлау қабілетіне ие бола алмағандықтан, оның мүмкіншіліктері де төмен болады. Осы себепті аталған жобаны келешекте дамыту ісі жоспарлануда. Алға қойған жоспар бойынша жоба келесі бағыттарда жетілдірілуі тиіс:

- қазақ тіліндегі сұраулы және лепті сөйлемдерді аудару;
- құрмалас сөйлемдерді аудару ерекшеліктерін зерттеу;
- қазақ тіліндегі сан есімдердің аударылу ерекшеліктерін зерттеп, оларды жобаға енгізу;
- мәліметтер қорын жаңа сөздермен толықтыру;
- аударма ісін пәндік аймаққа бөліп қарастыру.

М.АБАҚАН, С.КЫЗЫРКАНОВА

Әл-Фараби атындағы ҚазҰУ, Алматы, Қазақстан

ОРЫС ТІЛІНДЕГІ ПРЕДЛОГТАРДЫҢ КӨПМАҒЫНАЛЫЛЫҒЫНА БАЙЛАНЫСТЫ ҚАЗАҚ ТІЛІНЕ АУДАРЫЛУ ЕРЕКШЕЛІКТЕРІ

Заманауи технология дамыған кезеңде қазақ тілінен орыс тіліне, орыс тілінен қазақ тіліне, немесе ағылшын тілінен қазақ тіліне компьютермен тура аударма жасау үлкен сұранысқа ие болып отыр. Компьютерлік аударма жасауға байланысты көптеген ғылыми зерттеулері бар. Алайда бір тілден екінші тілге аударма жасау тыңғылықты еңбек пен терең білімді қажет етеді.

Сөздердің көпмағыналылығы көптеген жылдар бойы шешімін іздеп жүрген мәселелердің бірі болып табылады. Өткен ғасырдың 40-жылдарынан бастап жеке мәселе ретінде қарастырыла бастаған сөздердің лексикалық, синтаксистік көпмағыналылығы бүгінгі күні өзінің зерттеу аясын біршама кеңейткен. Қазіргі таңда бұл мәселе әлемдік тәжірибеде көпмағыналылық мәселесі сан қырынан танылып, зерттеліп келеді.

Қазақ тіліндегі кез келген сөз контексте мағыналық өзгеріске ұшырап, әртүрлі мағынада жұмсалуы мүмкін. Көпмағыналық тек лексикалық атауларға ғана емес, сондай-ақ, грамматикалық тұлғаларға да тән. Осы мәселені шешу барысында қазақ тілінің ішкі заңдылықтарының негізгісі саналатын сингармония заңдылығын негізге алуды жөн көрдік. Бұл мәселені шешу мақсатында орыс тіліндегі предлогтардың қазақ тіліне аударылуы және олардың көпмағыналық сипаты қарастырылады. Орыс тілінде предлогтардың алатын орны ерекше. Олар жеке тұрып өзңнше лексикалық мән бере алмайды. Предлогтар зат есіммен, сан есіммен, сын есіммен, үстеумен тіркесте тұра алады және байланыса келе, әрекет пен қимылдың орнын, бағыты мен уақытын білдіреді. Сан есім мен сын есім, үстеу тек зат есім орнына жұмсалғанда немесе зат есім қызметін атқарғанда ғана предлогтармен тіркеске түсе

алады. Бір сөзге бірнеше предлог жалғана отырып әр қайсысы әр түрлі мағына үстейді. Мысалға “через” предлогын қарастырайық. “Через” предлогі тіркеске түсіп “кейін, соң, арқылы” дегенді білдіре алады:

- Через секунду - бір секундтан кейін;
- Через мост - көпір арқылы ;

Машиналық аудармада көпмағыналылықтың предлогтарға байланысты мәселесі осы жерден көрініс табады. Қазақ тілінде предлог кездеспейді, алайда орыс тіліндегі предлогтарды мағыналық ерекшеліктеріне қарай әдеби тілде және ауызекі сөйлеуде аударылу жүйесі қалыптасқан. Орыс тіліндегі предлогтарды аударудағы кейбір мәселелерге тоқталайық. Біздің негізгі мақсатымыз предлогтардың қазақ тіліне аударылу жүйесіне назар аударып, олардың мағыналық ерекшеліктеріне, әсіресе, предлогтардың мәтіндегі көпмағыналық сипатына ғылыми сипаттама жасау.

Орыс тілінде кез келген предлог өзімен байланысты сөзбен белгілі бір ретпен, ішкі заңдылықпен байланысады. Байланысқа түскен сөздің тұлғасына, “род”-қа, жекеше не көпше түрде келуіне байланысты белгілі бір заңдылықты, реттілікті анықтауға болады. Осы орайда предлогтардың контексте келу жолдары мен олардың мағыналық құрылымын анықтау, аударма жасау маңызды. Ал қазақ тіліне бұл предлогтар септік жалғауларымен алмастырылып аударылады. Қазақ тіліндегі кеңістік септіктері барыс табыс, жатыс шығыс және көмектес септіктері тілдің тарихи дамуына бірінің орнына бірі жұмсалған. Септік жалғауларының арасындағы семантикалық жақындық орыс тіліндегі предлогтарды аударғанда ерекше көрінеді. Мысалы, “на” предлогы “үстіне” немесе “үстінде” мағынасында келуі мүмкін, ал “за” предлогы әрекеттің мақсатын немесе заттық мағынадағы есімдердің орнын білдіруі мүмкін. Екіншіден, мағыналық жақындық дегеніміздің өзі олардың мағыналық құрылымы бірдей деген сөз емес, арасында семантикалық айырмашылық бар. Бір ғана предлог екі септікпен аударылғанда екі түрлі мәнге ие болуы ықтимал. Компьютерлік аударма жасағанда предлогтардың аударылуындағы көпмағыналық пен мағыналық құрылымды ескерудің маңызы ерекше болатыны осыдан. Орыс тіліндегі көптеген предлогтар зат есіммен байланысып келіп, әрекет пен қимылдың орнын, бағыты мен уақытын білдіреді. ғалымдардың көрсетуіне қарағанда, предлогтар қазіргі орыс тілінде әрекеттің уақытын, мезгілін, себебін, мақсатын да білдіру үшін жұмсалады.

Орыс тіліндегі барлық предлогтарды қазақша контексте келгендегі аударылу жолдарын анықтауға толық мүмкіндік бар. Мысалы, “*на столе”* (егер *на* предлогі және мужской род жекеше түрдегі, жалғауы *е* болатын болса) тіркесін “үстелде” (жатыс септігіндегі зат есім) немесе “үстел үстінде” деп алсақ, ал “*на стол”* (егер *на* предлогі және мужской род жекеше түрдегі, жалғаусыз болса) тіркесін “үстелге” (барыс септігіндегі зат есім) немесе “үстел үстіне” деп барлық предлогтармен кездесуі мүмкін жағдайларды қарастыру арқылы жүйелеуге болады.

Мысал		1 нұсқа	2 нұсқа
На стол+е	На + сущ(ед\м.р) + е	зат+ жат(да.де..)	Зат+тәуел+үстінде
На книг+е	На + сущ(ед\ж.р) + е	зат+ жат(да.де..)	Зат+тәуел+үстінде
На окн+е	На + сущ(ед\с.р) + е	зат+ жат(да.де..)	Зат+тәуел+үстінде
На книг+у	На + сущ(ед\ж.р) + у	зат+бар (ға.ге..)	Зат+тәуел+үстіне
На стол	На + сущ(ед\м.р)	зат+бар (ға.ге..)	Зат+тәуел+үстіне
На книг+ах	На + сущ(мн\ж.р) + ах	зат+көп+ жат(да.де..)	Зат+ көп +тәуел+үстінде
На стол+ах	На + сущ(мн\м.р) + ах	зат+көп+ жат(да.де..)	Зат+ көп +тәуел+үстінде
На окн+ах(ях)	На + сущ(мн\с.р) + ах(ях)	зат+көп+ жат(да.де..)	Зат+ көп +тәуел+үстінде
На книг+и	На + сущ(мн\ж.р) +	зат+көп+ бар	Зат+ көп

	и	(ға.ге..)	+тәуел+үстіне
На стол+ы	На + сущ(мн\м.р) + и	зат+көп+ бар (ға.ге..)	Зат+ көп +тәуел+үстіне
По книге	По +сущ(ед\ж.р)+е	Зат+шығыс(мен ,бен..)	Зат+ бойынша
По столу	По +сущ(ед\м.р)+у	Зат+шығыс(мен ,бен..)	Зат+ бойынша
По окну	По +сущ(ед\с.р)+у	Зат+шығыс(мен ,бен..)	Зат+ бойынша
по книгами	По +сущ(мн\ж.р)+ами(ями)	Зат+копт+шығыс(мен,бен..)	Зат+ копт+бойынша
По столами	По +сущ(мн\м.р)+ами(ями)	Зат+копт+шығыс(мен,бен..)	Зат+ копт+бойынша
По окнам	По +сущ(мн\с.р)+ам(ям)	Зат+копт+шығыс(мен,бен..)	Зат+копт+ бойынша

Қорыта келгенде тіліндегі предлогтардың көпмағыналығын ескеру керек. Компьютерлік аударма жасауда предлогтардың көпмағыналығын анықтау маңызды. Біз зерттеуімізде көп мағыналы осындай тұлғалардың аударылуындағы модельдерді анықтап ұсынамыз. Болашақта өзге де тілдік тұлғалардың көпмағыналығы зерттеуіміздің нысаны болмақ.

С. ҚҰЛМАНОВ, А.БАЙМЕНШИН

Мемлекеттік тілді дамыту институты, Алматы, Қазақстан

АВТОМАТТЫ АУДАРМА ЖҮЙЕСІНДЕ ПАЙДАЛАНЫЛАТЫН MOSES БАҒДАРЛАМАСЫ ТУРАЛЫ

Қазіргі жаһандану заманында техника мен технологиялық инновацияның дамуына байланысты өндірісті, жалпы қоғам салаларын жаппай автоматтандыру ісі қарқын алып келеді. Осы орайда тіл білімінің лексикография саласында да тілді компьютерлендіру бағытында екі және көптілді автоматты сөздіктер құрастыру, яғни машиналық аударма ісі кең етек алып келеді.

Профессор А.Жұбанов «машиналық аударманың «өмірге келуіне», біріншіден, XX ғасырдың екінші жартысынан бастап әр елдерде (континенттерде) бірнеше тілдегі ақпарат ағымының қарқындап өсуі, екіншіден, ғылыми-техникалық прогресс үшін оларды меңгеру қажеттігіне қатысты әлеуметтік себептер негіз болды», – дей келе [1, 71], машиналық (автоматты) аударманың тарихына, оның түрлері мен құрылымдарына кеңінен сипаттама береді.

Қазақ лексикографиясында машиналық (автоматты) аударма ісі енді ғана қолға алынып, негізінен екітілді сөздіктер құрастыру ісі (ЭЕМ-ді қоспағанда) XXI ғасырдың басында басталды десек қателеспейміз. Мұндай сөздіктердің қатарында ең алдымен Ш.Құрманбайұлының «Қазақша-орысша, орысша-қазақша терминдер сөздігі (бекітілген терминдер)» автоматтандырылған сөздігін атауға болады [2]. Автор сөздіктің алғысөзінде бұл сөздіктің басқа сөздіктерін айырмашылықтарын көрсете отырып, сөздікті құрастыруға негіз болған бес факторды көрсетеді. Сөздікке бұдан бұрынғы басылымдарға енбеген 2002-2004 жылдары бекітілген 1681 термин енгізілген.

Қазақ автоматты сөздіктерінің келесі бір түрі 31 томдық салалық сөздіктің материалдарына негізделген [3]. Мұнда 25 сала қамтылған. Пайдаланушы әр саланың тұсын басып, қажетті сөздің қазақша немесе орысша нұсқасын іздеп таба алады.

Автоматты сөздіктердің көп қолданылатын тағы бір түрлері – «Мемлекеттік қызметшілерге арналған орысша-қазақша, қазақша-орысша сөздік» [4], «Сөз көмек» және интернет арқылы енуге болатын «www.sozdik.kz» сайты. Бұл сөздіктерде бағдарламаға енгізілген сөздер мен сөз тіркестерінің қазақша немесе орысша баламасын табуға болады. Әрине бұл сөздіктердің негізінде дәстүрлі сөздіктердің материалдарына сүйенгендігін байқау қиын емес. Қажетті сөздерді дәстүрлі кітап түріндегі сөздіктен іздеп жатқаннан гөрі бұл сөздіктердің пайдаланушының уақытын үнемдеуде пайдасы мол. Дегенмен, көріп отырғанымыздай, бұл сөздіктер тек берілген сөздердің (терминдердің) ғана баламасын табуға арналған. Екітілді немесе көптілді сөздіктер негізінен сөздерді, сөз тіркестерін, сондай-ақ сөйлемдерді аударуға бағдарлануға тиіс. Осындай сөздіктердің қатарына қазақ тілінен орыс тіліне, орыс тілінен қазақ тіліне сөздерді, сөз тіркестерін, сөйлемдерді, тіпті мәтіндерді аударатын «Тілмаш» және «Sana Soft» екітілді аударма сөздіктерін жатқызуға болады. Алайда бағдарламалық базаға қазақ тілінің барлық ерекшеліктері толықтай енгізілмегендіктен, қазақ тілінің лексикалық бірліктері толықтай қамтылмаған, грамматикалық жүйесі дұрыс анықталмайды. Мұндай олқылық осы өнімдерді әзірлеушілердің автоматты сөздік жасаудың теориясын толық меңгермегендігінен және бағдарламашылар мен тілшілердің тығыз байланыста жұмыс істемегендігінен болса керек.

Қазіргі кезде іскағаздарын мемлекеттік тілде жүргізуді автоматтандыру ісінде де бастамалар кездеседі. Мысалы, іскағаз үлгілерін автоматтаты түрде өңдеуге арналған ҚР Мәдениет министрлігі Тіл комитетінің тапсырысымен «Мемлекеттік тілді дамыту институты» ЖШС дайындаған «Орысша-қазақша іскағаз үлгілерінің электронды бағдарламасы» біздің жобамызға көп септігін тигізді [5]. Бұл бағдарлама орыс және қазақ тілдеріндегі іскағаздар үлгілерін автоматты түрде табуға арналған. Біз осы жұмыстарды әрі қарай жалғастырып, бағдарламаны жетілдіріп, іскағаздар үлгілерінің орысша-қазақша және қазақша-орысша автоматты сөздігін шығаруды қолға алып отырмыз. Бұл сөздіктің негізгі роботы ретінде Moses бағдарламасы қолданылады. Мақалада осы бағдарламаға қысқаша сипаттама беруді көздедік.

Moses бағдарламасы кез келген тілден аударма жасау моделін автоматты түрде дайындауға мүмкіндік беретін машиналық аударманы статистикалық жолмен жүзеге асыруға негізделген. Бағдарламаны қолдануға қажет нәрсе – аударылған мәтіндердің (параллель корпус) жиынтығы. Бағдарламаның тиімді іздеу алгоритмі көптеген нұсқалардың ішінен барынша ықтимал баламаны тез табады.

Moses бағдарламасында аударуға «үйрету» процесі параллель деректер негізінде жүргізіледі және екі тілдегі мәтінді сәйкес аудару үшін сөздердің coocurrences-і пайдаланылады. Бұл сәйкестіктер бір тілдегі сөздің екінші тілдегі ең жуық баламасын бірізділік негізінде табуға, сондай-ақ машиналық аудару кезіндегі синтаксистік иерархияны пайдалануға негізделеді.

Moses екі негізгі компоненттен тұрады: даярлық құбырөткізгіші (трубопровод подготовки) және декодер. Құбырөткізгішті даярлау, шынында, бастапқы деректерді (параллель және түсіндірме) қабылдап, оларды машиналық аударма моделіне айналдыратын құрал-саймандар жиынтығы болып табылады.

Мұнда енгізілген деректер сөзбе-сөз аударманы алу немесе қажетіне қарай иерархиялық ережелерді орындау үшін қолданылады да, осы ережелер бойынша алынған статистика ықтималдықты бағалау үшін қолданылады. Аударма жүйесінің маңызды бөлігі тілдік модель, яғни тілдік деректерді қолдану арқылы құрылған статистикалық модель болып табылады.

Машиналық аудармада аударма жақсы шығуы үшін әртүрлі статистикалық модельдер бір-біріне қарама-қарсы қойылатын баптау тәсілі маңызды рөл атқарады. Moses бағдарламасында төмендегідей ең танымал баптау алгортимдері пайдаланылады:

1) http :/ / www.statmt.org/moses/?n=Moses.LinksToCorpora
2) http :/ / mokk.bme.hu / ресурсы / hunalign /
3) http :/ / code.google.com [6].

Moses декодері модульдік қағидат бойынша жазылады және пайдаланушыға кодсыздандыру процесін төмендегідей тәсілдермен өңдеуге мүмкіндік береді:

? Кіру: Бұл аударма процесін қалай орындау қажеттігін сипаттайтын XML-элементі бар аннотация немесе желінің торы немесе «шытырманы» сияқты күрделі құрылым (мысалы, сөзді тану) болуы мүмкін.

? Үлгінің аудармасы: Бұл сөзбе-сөз немесе иерархиялық (синтаксистік) ережелерді аудару болып табылады.

? Алгоритмді расшифровкалау: кодсыздандыруда іздеу барысындағы «сәйкестіктердің» өте көп болуы қиындық тудыратындықтан, Moses мұндай іздеу үшін stackbased, талдау (разбор) графигі және т.б. сияқты әртүрлі бірнеше стратегиялық тәсілдерді қолданады.

? Тілдік Модель: Moses бағдарламасында SRILM, KenLM, IRSTLM, RandLM сияқты әртүрлі бірнеше тілдік модельдер құрал-саймандарын пайдалануға болады.

? Moses серверлер: декодерге арналған XML-RPC интерфейсін қамтамасыз етеді.

? Веб-трансляция: Moses веб-беттерді аудару үшін пайдаланылатын скриптер жиынтығы.

? Құрал-саймандарды талдау: Moses шығыстарын талдауға және визуалдауға арналған сценарийлер.

Moses бағдарламасында машиналық аударманы адам редакциялауы үшін FirstPass ретінде пост-редакциялау жүргізіледі. Бұл аударманың уақытын (тиісінше жалпы құнын) азайтуы мүмкін. Автоматтандырылған аудармада SMT қолданылуы да мүмкін, алайда қазіргі кезде (2012 жылдың сәуірінен бастап) әрі қарай тереңдей зерттелу үстінде, жуырда EC, Casmacat10 және MateCat11 жобалары іске қосылды.

Moses арқылы дайындалған әзірлеменің негізгі платформасы Linux Moses болып табылады. Алайда Moses басқа платформалармен де жұмыс істейді. Мысалы, Moses бағдарламасы Windows-те Cygwin арқылы жұмыс істеуі, Moses әзірлеушілері OSX қондырғысын да пайдалануы мүмкін.

Moses маузер және бірлескен автор (2009) ұсынған ауқымды лексика моделін пайдаланады, алайда әрбір тұтас сөзді үйрету мүмкіндігі баяу жүреді.

Moses бағдарламасында жалпы алғанда көптеген грамматикалық ережелер қамтылған. Алайда иерархиялық жүктеуге арналған ережелер кестесі жадында декодер өте баяу жұмыс істейді және жедел (оперативті) жадты көп пайдаланады. Мұндай ережелердің кейбіреулері үшін CKY арқылы іске асырылған кодсыздау алгоритмі оңтайлы болып табылмайды. Сондықтан бөлініп алынған модельдерді іздеуге арналған алгоритмдерді пайдалану немесе балама нұсқаларды зерттеуге аса көңіл бөлу керек.

Автоматты аударма жасау бағдарламаларына **жаңа қызметтерді қосу қазіргі кезде 2009 жылғы машиналық аударма марафонында** efforts арқасында оңайлады. Алайда бұл әліге дейін күрделі іс болып қалып отыр. Сонымен қатар интерфейс TranslationOption қажет болғанда Hypothesis-ті талап етеді.

Жуырда **RandLM, IRSTLM көпағындылығын (многопоточность) кеңейту**ге арналған жобалар іске қосылды. Мысалы, **сіздің компьютеріңіз көпядролы болса,** Multi-Threading өте пайдалы [7].

Мемлекеттік тілді дамыту институты қолға алған «Ісқағаздары үлгілерінің орысша-қазақша, қазақша-орысша автоматтандырылған сөздігін» жасауда негізге алынған Moses бағдарламасының жұмыс істеу принципі қысқаша айтқанда осындай. Алайда Moses бағдарламасы жалпы алғанда көпағынды болғандықтан, автоматты аударуға қатысты қызметтердің біразын атқарғанымен, аударма жасалатын тілдердің құрылымдық ерекшеліктеріне байланысты әлі де жетілдіруді талап етеді.

Әдебиеттер

1. *Жұбанов А.* Автоматты (машиналық аударма) // Аударматану. –Алматы: «Тіл» оқу-әдістемелік орталығы, 2008. –70-93-беттер.

2. *Құрманбайұлы Ш.* Қазақша-орысша, орысша-қазақша терминдер сөздігі (бекітілген терминдер). –Алматы: «Сөздік-Словарь», 2004.

3. *Шарипбаев А.А., Тренкеншу В.П.* Көпсалалы қазақша-орысша-қазақша сөздік. –Астана, 2004.

4. Русско-казахский словарь для государственных служащих. –Астана: «Алтынсофт Астана», 2008.

5. *Қапалбеков Б.С., Құсбекова Б.Ғ., Байменшин А.М., Әбділдаева М.Б.* Орысша-қазақша іскағаз үлгілерінің электронды бағдарламасы. – Алматы: Мемлекеттік тілді дамыту институты, 2010.

6. *Philipp Koehn*. Statical mashine translation. Cambridge University Press, 2009.

7. www.baseage.com

A. SUNDETOVA[1], M.L.FORCADA[2], A. SHORMAKOVA[1], A. AITKULOVA[1].

[1] Information Systems Chair, Al-Farabi Kazakh National University, Al-Farabi av., 71, 050040 Almaty, Kazakhstan, and

[2]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant, Spain

STRUCTURAL TRANSFER RULES FOR ENGLISH-TO-KAZAKH MACHINE TRANSLATION IN THE FREE/OPEN-SOURCE PLATFORM APERTIUM

Introduction

Translating natural text from English to a Turkic language such as Kazakh faces important challenges:

On the one hand, the complex agglutinative morphology of Turkic languages is very different from that of a fusional, morphologically not too complex language like English; an immediate effect is the fact that correspondences can seldom be modelled as word-for-word translations. Even if Turkic language morphology shows clear morphotactics (ordering of morphemes), its morphophonology shows complex phonological changes to due to interactions between neighboring morphemes (vowel harmony, sonorization, etc.) many of which are explicitly represented in writing.

On the other hand, there are many differences between the syntax of Turkic languages and English. Just to name a few: subject–object–verb order (compare subject–verb–object in English), use of postpositions (compare prepositions in English), head-final syntax with modifiers and specifiers always preceding the modified/specified (normally following in English), overt case marking allowing for a rather free ordering of arguments (versus a more fixed order in English), lack of definite articles (extensively used in English), verbal-noun-centered structures where English uses modal verbs (*must, have to, want to*) or verbal-noun or verbal-adjective-centered constructions where English has subordinate clauses using finite verbs with relatives or subordinating conjunctions (*the book which I read, the place where I saw him, before he came*), lack of a parallel of the English verb *have*, as used for possession, etc. For an account (in Russian) of syntax differences between English and Kazakh, see Печерских & Амангельдина (2012).

When sufficiently large sentence-aligned parallel corpora are available (for instance, as in the case of English to Turkish, see, for example, Tyers and Alperen 2010), statistical machine translation (Koehn 2010) may be used to attempt translation from English into a Turkic language

(in fact, statistical machine translation is currently offered by Google for two Turkic languages, Azeri and Turkish). However, in the case of Kazakh, it would be very hard to put together the necessary amount of sentence-aligned parallel text, and rule-based machine translation, in which experts write up dictionaries and grammatical rules that are applied by an engine, emerges as a clear solution; in fact, existing commercial systems for English to Kazakh (Sanasoft[7], Trident[8]) all appear to be rule-based.

We are currently engaged in building a free/open-source rule-based machine translation system from English to Kazakh, and we are using the Apertium free/open-source machine translation platform (Forcada et al. 2011, http://www.apertium.org) for various reasons. On the one hand, the platform already contains free/open-source English morphological dictionaries and, what is more important, Kazakh morphological dictionaries (Salimzyanov et al. 2013) which take care of all of the morphotactics and morphophonology and provide a basic vocabulary; this allows us to concentrate our work in two fronts: building the lexical transfer part, that is, a bilingual dictionary (already underway) and building structural transfer rules (grammatical rules for translation), which will be the subject of this paper. On the other hand, building free/open-source dictionaries and rules for English to Kazakh means that they will be freely available,[9] for instance, to build translation systems for other Turkic languages; this gives a strategic value to our work, as most of the structural transfer rules will be ready for use with other Turkic languages with little modification or no modification at all.[10]

The paper, which describes work in progress in the Apertium English-to-Kazakh structural transfer, is organized as follows: Section 323 describes the free/open-source rule-based machine translation platform, focusing on structural transfer. Section 0 describes the structural transfer rules currently available to tackle the main syntactic divergences between English and Kazakh; section 0 describes some successful structural translations and some limitations, and, finally, section 0 gives concluding remarks and outlines future work.

The Apertium platform

Apertium (Forcada et al. 2011, http://www.apertium.org) is a free/open-source rule-based machine translation (MT) platform that was launched in 2005 by the Universitat d'Alacant. Though it was initially aimed at translating between closely related languages, it was later extended to be able to deal with unrelated languages. All of the components of the platform (MT engine, developer's tools, and linguistic data for an increasing number of language pairs) are licensed under the free/open-source GNU General Public License (GPL, versions 2 and 3) and are available to everyone interested in the website.

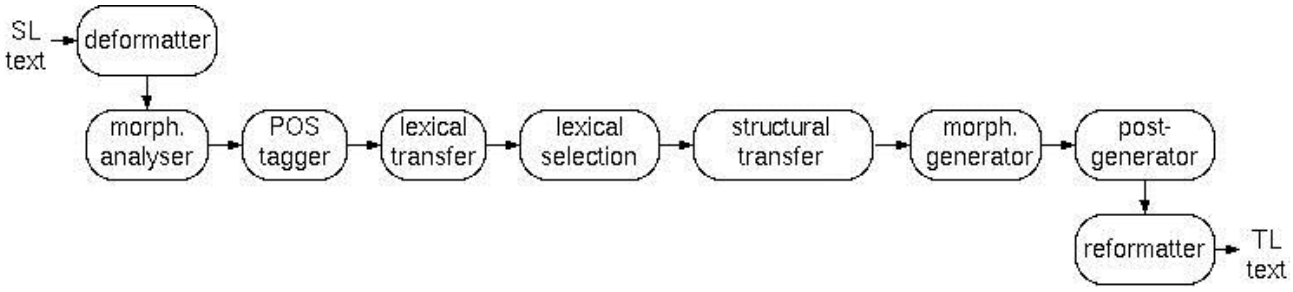


Figure 1: A sketch of the Apertium workflow

7 http://www.sanasoft.kz/c/ru/node/47 (in Russian) http://www.sanasoft.kz/c/kk/node/53 (in Kazakh).

8 http://www.translate.ua/us/on-line; also through http://itranslate4.eu/en/

9 They already are: see a snapshot at: https://svn.code.sf.net/p/apertium/svn/incubator/apertium-eng-kaz/

10 The Apertium project has a particularly active sub-project for Turkic languages (http://wiki.apertium.org/wiki/Turkic_languages), which has its own mailing list, https://lists.sourceforge.net/lists/listinfo/apertium-stuff.

Apertium-based MT systems are transfer systems implemented as text pipelines (see Figure 1) consisting of the following modules:

1. A deformatter that separates the text to be translated from the formatting tags. Formatting tags are encapsulated as "superblanks" that are placed between words in such a way that the remaining modules see them as regular blanks (for instance, tags in the HTML text I see the sky are encapsulated as I see []the sky[] and everything in square brackets is treated just as regular blanks).
2. A morphological analyser, yielding, for each surface form (SF), for each lexical unit as it appears in the text, a lexical form (LF) composed of: lemma (dictionary or citation form), lexical category (or "part-of-speech"), and inflection information. For instance, the English SF *books* would yield two LFs: *book*, noun, plural, as in *I have bought some books*) or *book*, verb, present tense, 3rd person, as in *He books a ticket*). The morphological analyser executes a finite-state transducer generated by compiling a morphological dictionary for the source language (SL).
3. A constraint-grammar (Karlsson 2005) module based on CG3[11] is used to discard some LFs using simple rules based on context (this module is not depicted in the figure).
4. A part-of-speech tagger based on hidden Markov models (Cutting et al. 1992) selects one of the remaining LFs. The statistical models may be supervisedly trained on an annotated SL monolingual text corpus, or trained in an unsupervised way, either on an unannotated monolingual SL corpus or using two unrelated, unannotated source language and target language corpora (as in Sánchez-Martínez et al. 2008). The Apertium part-of-speech tagger can also read linguistically-motivated constraints (much more rudimentary than constraint grammar rules in the previous module) that forbid specific sequences of two LFs.
5. A lexical transfer module adds, to each source language LF (SL LF), one or more corresponding target language LFs (TL LFs). This module executes a finite-state transducer generated by compiling a bilingual SL–TL dictionary.
6. An (optional) lexical selection module (currently not active in the English→Kazakh system) reads in rules that allow for the selection of one of the TL LFs according to context. When this module is absent, the TL LF given as default in the dictionaries is used.
7. A structural transfer module processes the stream of SL LF–TL LF pairs produced by the lexical transfer module and transforms it into a new sequence of TL LFs; a more detailed description is found in section 0 as this is the main subject of this paper.
8. A morphological generator takes the sequence of TL LFs and generates a corresponding sequence of TL SFs. The morphological generator executes a finite-state transducer generated by compiling a morphological dictionary for the TL.
9. A post-generator takes care of some minor orthographical operations such as apostrophations and contractions in the target language (this module is not used for English to Kazakh).
10. Finally, the deformatter opens the square-bracketed superblanks and places the formatting tags back into the text so that its format is preserved.

Structural transfer in Apertium

The structural transfer module in Apertium processes the stream of source-language lexical form – target-language lexical form pairs (SL LF–TL LF pairs) and transforms it into a new sequence of TL LFs after a series of structural transfer operations specified in a set of rules: reordering, elimination or insertion of TL LFs, agreement, etc. Structural transfer rules have a pattern–action form: when a specific (finite-length) pattern of SL LFs is detected, an action builds and generates the corresponding sequence of TL LFs. Rules are applied in a greedy, left-to-right, longest-match fashion. There are two main modalities of structural transfer. The first one (used for related languages) generates the TL LF sequence in a single step. The second one (used in the English–Kazakh system described in this paper) uses three stages to improve the granularity of structural transfer rules (each one has its own rules file):

11 http://beta.visl.sdu.dk/cg3.html

- A first round of transformations ("chunker") detects SL LF patterns and generates the corresponding sequences of TL LFs grouped in chunks representing simple constituents such as noun phrases, prepositional phrases, etc. These chunks bear tags that may be used for inter-chunk processing.
- The second round ("interchunk") reads patterns of chunks and produces a new sequence of chunks. This is the module where one can attempt to perform some longer-range reordering operations, inter-chunk agreement, case selection, etc.
- The third round ("postchunk") transfers chunk-level tags to the lexical forms they contain and whose lexical-form-level tags are linked (through a referencing systems) to chunk-level tags (for instance, case determined for a noun phrase is transferred to the main noun), and removes all grouping information to generate the desired sequence of TL LFs.

English-to-Kazakh structural transfer

This section describes the current structural transfer in Apertium-eng-kaz (revision 46018, 26.07.2013). English to Kazakh chunker rules (file apertium-eng-kaz.eng-kaz.t1x) are described in detail in section 0. English-to-Kazakh inter-chunk rules (file apertium-eng-kaz.eng-kaz.t2x) are described in detail in section 0. The English-to-Kazakh system has an additional clean-up stage that takes care of the fact that Kazakh morphotactics, as defined in the Kazakh morphological dictionary, contains optional morphemes: for instance, there is no singular morpheme or no-possessive morpheme, but these are generated in the previous three steps. They are eliminated here. Rules for cleanup have the same form as transfer rules for a related language pair (file apertium-eng-kaz.eng-kaz.t4x).

The English-to-Kazakh chunker

As regards the first round of structural transfer (the "chunker", rules written in the apertium-eng-kaz.eng-kaz.t1x file), rules have been written to address some of the the main local morphosyntactic divergences between the languages involved. The prototype is able to perform the local operations necessary to adequately process short noun phrases, adjective phrases, verb phrases and adpositional phrases (that is, prepositional phrases in English and postpositional phrases in Kazakh).

Chunking rules, of which there are currently 60, identify six kinds of chunks and translate them into Kazakh as much as possible, leaving some minor operations to be performed in later stages of structural transfer (for instance, the case of noun phrases). Table Table shows a description of the kinds of chunks found, with examples, and the chunk-level tags associated to them. The translation of English noun phrases and prepositional phrases are given below as examples.

Table 2: Chunk-level tags currently associated to each type of chunk. Those marked with an asterisk correspond to optional morphemes that will need special treatment in a cleanup module (see text in section 0).

Phrase	**Description**	**Examples of English chunks detected**	**Chunk-level tags**
NP	Noun phrase	noun determiner–noun numeral–noun adjective–noun determiner–adjective–noun	number*, person, possessor*, case
VP	Verb phrase	finite_verb, do–not–finite_verb have–verb_participle be–verb_gerund must–verb have_to–verb want_to–verb	number, person, tense/conditionality, possessive*, negation*

Phrase	Description	Examples of English chunks detected	Chunk-level tags
PP	Postpositional phrase (except genitive phrases ending in "-{N}{I}ң")[12]	preposition–noun preposition–determiner–noun preposition–numeral–noun preposition–adjective–noun preposition–determiner–adjective–noun	number, person, possessor*, and case
GenP	Genitive postpositional phrase ending in "-{N}{I}ң"	(same as PP, with "of" as preposition)	number, person, possessor*, and case
AdjP	Adjectival phrase (except superlatives with "ең")	adjective "more"–adjective	(none)
SupP	Superlative phrase (adjectival phrase with "ең")	adjective_"-est" the–"most"–adjective	possessor*,[13] case

Translation of noun phrases

Consider the following example: the chunker identifies the English sequence *the large book* (determiner–adjective–noun) as a noun-phrase chunk. It translates into Kazakh, and assigns it four chunk-level tags: number (set to singular), person (set to 3rd), possessor (to be determined, as the noun *кітап* ('book') could receive a 3rd-person possesive ending (кітабы) later if the context were, for instance, *the large book of animals, аңдардың үлкен кітаб**ы***), and case (to be determined as it could be, for instance, accusative in *I saw the large book, Мен үлкен кітап**ты** көрдім*).

Translation of prepositional phrases

On encountering an English prepositional phrase, which has to be rendered in Kazakh as a postpositional phrase, there are three possible outcomes:

(1) The prepositional phrase results in a simple postpositional phrase using the locative "-{D}{A}",[14] ablative "-{D}{A}н", etc., but not the genitive "-{N}{I}ң":

[PP [P *in*] [NP *the beautiful garden*]] → [PP [NP *әдемі бақша*] [P -*да*]]

(2) The prepositional phrase results in a simple postpositional phrase using the genitive -NIң, which will be marked GenP:

[PP [P *of*] [NP *the beautiful garden*]] → [GenP [NP *әдемі бақша*] [P -*ның*]]

(3) The prepositional phrase results in a complex postpositional phrase based around a noun such as *аст*, *үст*, etc.:

[PP [P *under*] [NP *the garden*]] → [PP [NP [GenP [NP *бақша*][P -*ның*]] [NP *астын*]] [P -*да*]]

In all three cases, the possessor tag of the chunk, which corresponds to the main noun in the PP or the GenP has to be left open to being determined in later transfer operations (consider, for instance, the case that the PP *in the beautiful garden* is part of a larger structure, *in the beautiful garden of the city*, *қаланың әдемі бақша**сын**да*, in which the noun *бақша* 'garden' receives a possessive ending).

Translation of verb phrases

The mapping of English verb tenses onto Kazakh is not completely straightforward and is treated in the chunker. Just to give a few examples, present simple and future are rendered using the same

12 Upper case letters in braces (such as {N}) represent hypothetical archiphonemes (actually archigraphemes) that are realized as phonemes (actually graphemes) after morphophonological rules have been applied. For instance, in the genitive ending "-{N}{I}ң", the archiphoneme {N} may be realized as т, д, or н and the archiphoneme {I} may be realized as і or ы depending on the previous phonological context. This is perfromed during morphological generation (see section 0).

13 Treated as a noun phrase with an implied noun ("the largest [book]")

14 {D} can be *д* or *т*, and {A} can be *е* or *а*, depending on the phonological context.

tense in Kazakh (*I play → Мен ойнаймын; I will play → Мен ойнаймын*); tenses expressing continued activity, such as the English present continuous or past continuous (*I am playing, I was playing*), have to be detected and mapped onto sets of two lexical units (*Мен ойнап жатырмын, Мен ойнап отырдым*) where the main verb is found in the *-п* participle form (*ойнап*), and a suitable finite form (*жатырмын, отырдым*) of an auxiliary verb (*жатыр, отыр*) is used to express number and person agreement[15] (see Table Table) for details.

Table 3 Examples of tense mapping operations performed at the chunk level

English tense	Example	Morphemes	Equivalent tense in Kazakh	Translation
Present Simple	*I play*	*ойна* + *й*(*а* or *е*) <aorist> + <person>	Ауыспалы осы шақ (changing present simple)	*Мен ойнаймын*
Present Continuous	*I am playing*	*ойна* + *п* (*ып* or *ін*) <perfect participle> + *жатыр* + <present> + <person>	Нақ осы шақ (now present tense)	*Мен ойнап жатырмын*
Past Continuous	*I was playing*	*ойна* + *п* (*ып* or *ін*) <perfect participle> + *отыр* + *д{I}* <past> + <person>	Бұрынғы өткен шақ (past continuous)	*Мен ойнап отырдым*

Verb-phrase chunks are also used to prepare translations not using a finite verb (but a nominal or adjectival structure, often based on non-finite forms of verbs instead). For instance, for obligatory English modal constructs (*have to, must, need to, should*) verb phrases made up of three lexical units have to be generated, with a verbal noun, an adjective roughly meaning "necessary" (*керек*) or "proper" (*жөн*), and a form of the copula (absent in present tense); the subject receives the genitive or dative case: *I have to go → Менің баруым керек*, *I need to go → Маған бару керек*, etc.; see these and other modal construction examples in Table Table.

Table 4 Translation of some English modal verbs

Construction	Example	Morphemes	Translation	Gloss
Must have to	*I must go, I have to go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-у* <gerund> + *-{I}м* <1st person possessive> + *керек* <adjective>	*Менің баруым керек*	My going necessary [is]
Should	*I should go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-{G}{A}н* <past gerund> + *-{I}м* <1st person possessive> + жөн <adjective>	*Менің барғаным жөн*	My going proper [is]
Need to	*I need to go*	*Мен* + *-{G}{A}[н]* <dative> + *бар* + *-у* <gerund> + *керек* <adjective>	*Маған бару керек*	To me, going necessary [is]
Want to	*I want to go*	*Мен* + *-{N}{I}ң* <genitive> + *бар* + *-{G}{I}* + *-{I}м* <1st person possessive> + *кел* + *-{E}д{I}* <past, 3rd person>	*Менің барғым келеді*	My going will come

15 Actually, Kazakh language uses four auxiliary verbs: *жатыр* ('lie', used when the activity takes a long time), *отыр* ('sit', used when the activity appears to be done in a sitting position), *тұр* ('stand', when the takes a short time), and *жүр* (when the activity repeats regularly). Choosing the most adequate auxiliary verb is hard without a semantic analysis, which is not easily available in Apertium. Our current choice (an approximation) is *жатыр* ('lie') for the present continuous and *отыр* ('sit') for the past continuous.

Finally, as negative constructions in English contain more words than their corresponding affirmative words, or may even use an auxiliary verb (as in *do not*, *did not*), they have to be separately detected as verb chunks to generate the appropriate Kazakh negative forms (*I play → мен ойнаймын; I do not play → мен ойна***ма***ймын*. For examples of other negative constructs, see Table Table).

Table 5 Translation of some negative constructions.

Construction	Example	Morphemes	Translation	Note
Present Continuous (negative)	*I am not playing*	*ойна* + *п* (*ып* or *ип*) <perfect participle> + *жатқан жоқ* + <person>	*Мен ойнап жатқан жоқпын*	In present auxiliary verbs (жатыр/отыр) do not have a synthetic negative form.
Can (negative)	*I can not play*	*Ойна* + *-{E}* <imperfect participle> + *ал* + *ма* <negative> + *й* <aorist> + *мын* <1st person>	*Мен ойнай алмаймын*	

Verb phrases (VP) are marked at the chunk level with person and number, both to be determined and linked via references to the appropriate morphemes in the appropriate verb lexical forms. The chunk-level person and number to be determined will be rewritten by the appropriate 2nd-level (interchunk) transfer rules, and will be propagated to lexical forms at the 3rd-level transfer stage (postchunk).

Other indicators that have to be made available at the chunk level are negation (for negative verbs) and conditional (which will be handled as a tense). For instance, negation can be easily determined at the chunking level when the English VP chunk contains *not*, as in *I don't play → мен ойна***ма***ймын*, but may need to be determined at the interchunk level in sentences having a non-negative VP but a negative word like those starting with *еш-*, like *I write nothing → мен ешнәрсе жаз***ба***ймын*, which requires a negative form of the verb (*-ба-* in the example).

Translation of adjectival phrases

In Kazakh noun phrases, adjectives come before nouns and do not show any agreement with nouns. Adjectives can also appear in separate adjective phrases. Here are some examples:

(4) The adjective alone, marked AdjP:

[AdjP *beautiful*] → [AdjP *әдемі*]

(5) Comparative adjective phrases (English *more* + adjective, or adjective-*[e]r*); the Kazakh translation chooses the comparative suffix "-{I}p{A}{K}":

[AdjP *more beautiful]* → [AdjP *әдем***ірек**]

(6) For superlative adjective phrases "*the most* + adjective" or "adjective-[*e*]*st*", translation is built using "ең" + adjective:

[SupP *the largest]* → [SupP **ең** *әдемі*]

[SupP *the most beautiful]* → [SupP **ең** *үлкен*]

As noted in §0, superlative adjective phrases have some properties of noun phrases (such as receiving possessive morphemes when modified by a genitive phrase: *the most beautiful of people → адамдардың ең әдемі***сі**); one could say that they are treated as NPs with an implied noun.

English-to-Kazakh inter-chunk processing

The second round of structural transfer (the "interchunk" rules written in the apertium-eng-kaz.eng-kaz.t2x file) is currently performed by a proof-of-concept set of 18 rules, representative of following operations:

- Inter-chunk agreement (for instance, number and person agreement between subject noun phrase and verb phrase): features to be agreed here are left undefined by the chunker; those that are not defined at the interchunk phase are left for the post-chunk phases.

- Assigning case to noun phrases (which are generated without case by the chunker): for instance, accusative case for objects (*I see the sky → Мен аспан***ды** *көремін*), genitive case for obligatory constructs (*I have to go → мен***ің** *баруым керек*), dative case for the verb *to need* (*I need a book →Ма***ған** *кітап керек*), locative case for possession (*I have a book → Мен***де** *кітап бар*), etc.
- Reordering: placing of object before verb (*I[1] see[2] the sky[3] → Мен[1] аспанды[3] көремін[2]*), placing of prepositional pharses before the verb (*They[1] played[2] on top of the tree [3] → Олар[1] ағаштың үстінде[3] ойнады [2]*), etc.

The set of rules has to be extended, as many combinations of the above phenomena are still not covered (for instance, there is no rule to obtain the right word order in *I have to go to the university →Менің университетке баруым керек*).[16]

Some results, problems and limitations

The system described is not much more than a proof-of-concept system that still needs to be extended to reasonably cover all transfer operations needed. Therefore, evaluating the output of the system using customary evaluation measures such as BLEU (Papineni et al. 2002) is still out of the question.

Instead, tables Table and Table show how our current prototype performs for some representative structures covered by the transfer rules currently available (some of them discussed above). As has been said above, are already at least two MT systems that translate from English to Kazakh: Sanasoft's and Trident's, both of which can be used online (see Introduction for details); therefore, we will briefly compare our results to those obtained by the commercial systems.

Table 6 Example machine translation output for some simple phrases and sentences.

Structure/problems	**English**	**Kazakh (Apertium)**	**Kazakh (Sanasoft)**	**Kazakh (Trident)**
Noun phrases	*your two beautiful gardens*	*сіздің екі әдемі бақшаңыз*	*Сенің екі әдемі бақ***тарың**.	**сендер** екі тамаша бақша**лар**
Prepositional phrases	*in the big city*	*үлкен қалада*	*Үлкен* **қала**	*үлкен қалада*
Possessives	the chief of the city	*қаланың басшысы*	*қаланың көсемі*	**Бас қала**
	On top of the tree of the garden of the city	*қала бақшасының ағашының үстінде*	**Зырылдауық ағаш бақ қала**	**В алқындыр-қаланың бақшасының ағашының**
Adjective phrases	bigger	*үлкенірек*	*Үлкен*	**үлкен**
Modal verbs	I have to go	*Менің баруым керек*	**Мен барып жатырмын жүрмін**	*Маған бару* **have**
	I can drive	*Мен жүргізе аламын*	*Мен* **болып жатырмын** *жүргізіп* **жатырмын**	*Мен жүру* **білемін**

16 As chunks detected by the chunker are finite-length and inter-chunk rules also process finite-length chunk sequences, it has to be noted that there will always be a limit to the scope of reordering or agreement rules.

Concluding remarks and future work

The current prototype already successfully solves many cases of noun-phrase, verb-phrase, prepositional-phrase, and adjectival-phrase translation (some actually better than the available commercial systems), and contains a reasonable vocabulary for testing purposes, which nevertheless still needs extending for real-world applications.

The following tasks have to be performed in order to have a working machine translation system:

- Completing the coverage of structural transfer rules and monolingual and bilingual vocabularies so that the system produces a translation for at least 90% of the English words and performs the basic operations to identify and process correctly short constituents (1–6 words).
- Releasing the resulting stable system as apertium-eng-kaz and disseminate it to the interested parties to obtain feedback about its functioning. We can reasonably expect this system to work better than the existing commercial systems in most aspects.

As a longer-range objective, and when a reasonably complete prototype is available, we will tackle another interesting goal: the use of feedback from human input (for instance, in an interactive machine translation system that provides completions to what the translator is typing).

Table 7: Example machine translation output for some simple phrases and examples.

English	Kazakh (Apertium)	Kazakh (Sanasoft)	Kazakh (Trident)
I see the blue sky.	*Мен көк аспанды көремін*	*Мен**де** көк аспан көріп жатырмын*	*Мен көгілдір аспанды көремін.*
You go to school	*Сіз мектепке барасыз*	*Сіз мектепке бар**дың***	*сендер үйрету барасыңдар*
A book has been given to you	*кітап сізге беріліп болған*	*Кітап* **барып жатыр** сізге берсін	Кітап **жібер-** сендерге болды
I can go to the three big shop	*Мен үш үлкен дүкенге бара аламын*	*Мен* **three** *үлкен магазинге* **болып** *жатырмын жүрмін*	*Мен* **үшке** *деген бару үлкен дүкен білемін*
The most beautiful of garden is opened	*бақшаның ең әдемісі ашылады*	**Көпшілік** *әдемі бақ ашық* **бар**	*Ең тамаша бақш**адан** болады* **ашыл-**
I see my car	*Мен менің жеңіл автокөлігімді көремін*	*Мен менің автомобилім көріп жатырмын*	*Мен өзінің автомобильсын көремін*
The famous doctor of the city is going to hospital	*қаланың танымал дәрігері емханаға барып жатыр*	*Атақты дәрігер қала ауруханаға барады*	*қаланың* **атайы** *докторы ауруханаға деген жиналады*
She eats chocolates with sugar	*Ол шоколадтарды қантпен жейді*	*Ол* **eats chocolates** *қант*	*Ол шоколадтарды қантпен жейді*

Acknowledgements: MLF thanks the Kazakh state program for the attraction of foreign scholars and Prof. Ualsher Tukeyev for supporting his visit to the Kazakh National University, where part of this work was carried out. We also thank Prof. Tukeyev for his valuable input.

References

1. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. 1992. “A practical part-of-speech tagger”. *Proceedings of the 3rd Conference on Applied Natural Language Processing* (ANLP’92), p. 133-140.

2. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A. Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M. 2011. “Apertium: a free/open-source platform for rule-based machine translation”. Machine Translation 25(2)127-144.

3. Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

4. Koehn, P. (2010) *Statistical Machine Translation*, Cambridge University Press.

5. Печерских, Т. Ф., Амангельдина, Г. А. (2012) “Особенности перевода разносистемных языков (на примере английского и казахского языков)”, Молодой ученый. №3, 259–261 [http://www.moluch.ru/archive/38/4406/]

6. Papineni, K., Roukos, S., Ward. T.. amd Zhu, W.J. (2002) “BLEU: a method for automatic evaluation of machine translation” *In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, Philadelphia (ACL 2002)*, pp.311–318.

7. Salimzyanov, I., Washington, J.N., Tyers, F.M. “A free/open-source Kazakh-Tatar machine translation”. Proceedings of MT Summit XIV (Nice, France, 4–6 September 2013), accepted.

8. Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2008). “Using target-language information to train part-of-speech taggers for machine translation”. *Machine Translation*, 22(1-2)29–66.

9. Tyers, F. M. and Alperen, M. S. (2010) “SETimes: A parallel corpus of Balkan languages”. Proceedings of the MultiLR Workshop at the Language Resources and Evaluation Conference at LREC2010, 49–53 [http://www.lrec-conf.org/proceedings/lrec2010/workshops/W22.pdf].

У. КАМАНУР, Б.З АНДАСОВА, Б.М БАЙГУШЕВА

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

ҚАЗАҚ-АҒЫЛШЫН-ҚЫТАЙ ДЫБЫСТЫҚ СӨЗДІГІН ӘЗІРЛЕУ

Қазақ тілі Қазақстан Республикасының мемлекеттік тілі ретінде елдің ішкі қарым-қатынасында ғана емес, сонымен бірге халықаралық байланыстарда да қолданылуы керек. Қазақстан Республикасы дүниежүзілік қауымдастыққа біріге отырып мемлекеттік тілдің даму деңгейін оның қарқынды зерттелуі мен қазақ тілін оқытудың әдістері мен құралдарын жасау негізінде дүниежүзілік деңгейге жеткізуді қамтамасыз ету керек. Сондықтан да қазіргі таңда тілді меңгеруде, көптілді қарым-қатынас дамыған заманда көптілді дыбыстық сөздікті құру технологиясын жасау өзекті болып отыр.

Бүгінгі таңда сөйлеу технологияларын қолдану арқылы дыбыстық сөздіктерді жасау қарқынды түрде дамып келеді. Бірақ олар негізінен ағылшын, орыс, қытай және т.б. тілдерге бағытталған. Қазақ тіліне арналған жарамды сөйлеу технологияларымен жасалған дыбыстық сөздіктер жоқтығы қазақ тілін бұл заманауи жетістіктерден тыс қалдырады.

Осы мақсатта ақпараттық және байланыс технологиялары саласы бойынша қазақ тіліндегі терминдер мен олардың ағылшын, орыс және қытай тілдеріндегі аудармаларының 4000 бірлік көлеміндегі базасы, қазақша сөйлеуді синтездеуге арналған дифон базасы құрастырылып, көп тілді дыбыстық сөздіктің онтологиясы, сонымен қатар, сөздікке енгізу үшін сөздерді тану және дыбыстық сөздіктегі сөздерді синтездеу алгоритмдері мен программалары жасалды.

Қазіргі кезде біздің елімізде бірнеше электрондық сөздіктер бар. Бірақ олардың ішінде әлі күнге дейін жаңа электрондық сөздіктерді жасауға немесе бар электрондық сөздіктерді өңдеуге мүмкіндік беретін компьютерлік программа жоқ. Белгілі бір салада сөздік жасаушылар өзі жасаған сөздіктің электрондық үлгісін алу үшін және оны кеңінен таратуға мүмкіндік беретін тиісті программа жоқ. Сонымен қатар осындай сөздіктердің ақырғы

пайдаланушылары өздерінің компьютерлерінде әртүрлі электрондық сөздіктерді орнатады. Сондықтан жаңа электрондық сөздіктерді жасау, өңдеу және оларды кеңінен таратуға мүмкіндік беретін программа NetBeans IDE ортасында жасалды.

NetBeans IDE — Java, JavaFX, Ruby, Python, PHP, javascript, C++ программалау тілдерінде қосымшаларды өңдеудің (IDE) еркін интегралданған ортасы. NetBeans ортасында программа өңдеу, сәтті инсталляциялау және NetBeans ортасының өзінде жұмыс жасау үшін алдын ала версиясына сәйкес Sun JDK немесе J2EE SDK орнатылуы керек. NetBeans өңдеу ортасы J2SE және J2EE платформалары үшін өңдеуді қолдады. 6.0 версиясынан бастап Netbeans J2ME, C++ (тек g++) мобильді платформалары және қосымша компоненттерді орнатусыз PHP үшін өңдеуді қолдайды [1].

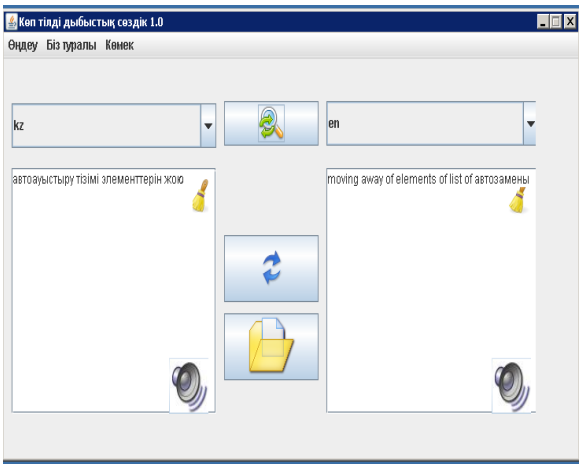
Бұл программа екі бөліктен тұрады: 1) сөздікке жаңадан сөздер қосу және 2) сөздіктен іздеуге арналған. Мұнда сөздік жеке файл ретінде болады. Оны сөздіктер базасына қосуға және алуға болады.

Сөздікке жаңадан сөздер қосу модулі. Дыбыстық сөздікте ақпараттық және байланыс технологиялары бойынша қазақ тіліндегі терминдер мен олардың ағылшын, орыс және қытай тілдеріндегі аудармаларының 4000 бірлік көлеміндегі деректер қоры бар, оларды толықтыруға болады.

Сөздіктен іздеу модулі. Программаның бастапқы беті төмендегі 1-суретте келтірілген.

Сурет 1. Программаның бастапқы беті.

Программаның сөздікпен жұмыс істеуге арналған жұмыстық беті төмендегі 2-суретте берілген

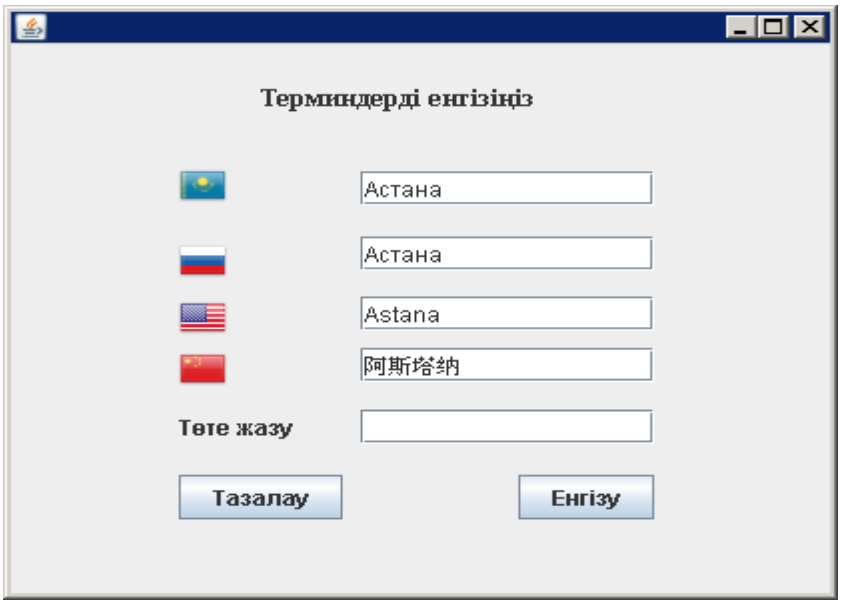


Сурет 2. Программаның сөздікпен жұмыс істеуге арналған жұмыстық терезесі

Программа терезесінің оң жақтағы бөлігінде қазақ, ағылшын, қытай немесе төте жазудағы терминді енгізіп қандай тілде екенін көрсетіп керек, содан кейін аударылатын тілді таңдап, аудару батырмасын басқан кезде, терезенің оң жақтағы бөлігінен аударылған

сөз көрсетіледі және оң жақ бұрыштағы дыбысты білдіретін суреті бар батырманы басқанда аударылған сөз дыбысталады. Бұл батырманыың тағы бір ерекшелігі – оқылуы күрделі сөздерді (көлемі шектелмеген) осы ұяшыққа енгізіп, батырманы басқанда енгізген сөздер дыбысталады. Егер енгізген термин сөздік қордан табылмаса, онда сәйкес нәтижелер табылмады деген хабарламаны шығарылады.

Программаның сөздікпен жұмыс істеуге арналған жұмыстық терезесінде «Өңдеу» және «Көмек», «Біз туралы» батырмалары орналасқан. Мұндағы «Өңдеу» батырмасы «Қосу» деген компоненттен тұрады, жаңа сөздікті жасау үшін мәзірдегі «Өңдеу ⇨ Қосу» команданы орындау немесе экрандағы қалтаның суретін басу керек. Нәтижесінде сөздікке сөз қосу терезесі шығады (3–ші сурет), мұнда қазақ тіліндегі терминдер мен олардың ағылшын, орыс және қытай тілдеріндегі, төте жазудағы аударамаларын өрістерге толтыру керек. «Тазалау» деген батырманы басқанда, өрістегі енгізілген мәліметтер жойылады.



Сурет 3. Жаңа сөздікті жасау үшін

«Енгізу» батырмасын басқаннан кейін «сөздік қорға сәтті жүктелді» деген хабарлама шығып, жаңа сөздер сөздік қорына қосылатын болады.

Автор туралы деректерді «Сөздік ⇨ Автор» мәзір арқылы көре аласыз.

Көмек мәзірінде сөздікті қалай қолдану керек екендігі жазылған.

Қазіргі таңда көп тілді дыбыстық сөздіктер ғылым мен білім, техника, өнеркәсіп салаларының бәріне де қатысты, қолданым аясы өте кең саналатыны белгілі.

Магистірлік зерттеу нәтижесінде көп тілді дыбыстық сөздіктерді жасаудың теориялық негіздерімен таныстым, көп тілді дыбыстық сөздіктің түрлері, қолданысы, тарихы ерекшеліктерін талдап, дыбыстық сөздікті жасау технологияларымен, әдістерімен үйрендім;

1) көп тілді дыбыстық сөздіктің онтологиясын құрлды;

2) ақпараттық және байланыс технологиялары бойынша қазақ тіліндегі терминдер мен олардың ағылшын, қытай тілдеріндегі аударамаларының 4000 бірлік көлеміндегі базасын жасалды;

3) көп тілді дыбыстық сөздіктегі сөздерді синтездеу алгоритмін және программасын жасау, қазақша сөйлеуді синтездеуге арналған дифон базасын құрлды;

4) көп тілді дыбыстық сөздікке енгізу үшін сөздерді тану алгоритмін жасап және програмалық жүзеге асырдым;

5) көп тілді дыбыстық сөздіктің программасын жасау құралдарын және интерфейсін сипаттадым.

Көп тілді дыбыстық сөздік түзудің ғылыми-теориялық мәселелерін зерттеуде таза лингвистикалық факторлар да, алуан түрлі ғылым салаларының ерекшеліктерінен туындайтын экстралингвистикалық факторлар да ескерілуі тиіс екенін тәжірибе көрсетіп отыр.

Сондықтан көп тілді дыбыстық сөздіктерді ғылыми-теориялық негіздерін зерттеу осы салада түзілетін туындылардың авторлардың тілдік түйсігіне, субъективті жағдайларға тәуелді болмауы үшін, объективті ғылыми негізде жасалуы тиіс екендігі аталмыш зерттеудің тақырыбын белгіледі.

Шын мәнінде, ғылыми ұғымдардың жүйесі қай тілдегі сөздікте болса да, жүйелі түрде берілуі керек. сөздіктерде де бір-біріне туыстас, төркіндес құбылыстарды атауда да ұқсас тілдік құрылымдардың қолданылуы шарт. Тілдік жүйе құрамында ұғымдардың бір-біріне байланыстылығын, олардың арасындағы сатылы бағыныңқылықты дәл бейнелей алатын терминологиялық бірліктер болу керек. Бұл жерде белгілі бір ұлт тіліндегі білім мен ғылым, техника, мәдениет салаларының сол ұлтты құрайтын халықтың дәстүрлі білім жүйесіне, ұлттық мәдениет ерекшелігіне бұрыннан тән болып, онымен сабақтаса өріліп, өрбіп жатуы аса маңызды. Сонда ғана ұлттық тілдің салалық терминологиядағы икемділігі, қолданым дәрежесі жоғары болмақ. Мысалы, қазақ халқына тән дәстүрлі шаруашылық түрлері бойынша жаңа терминдерді түзу ісі қиыншылық тудырмайды. Себебі бұл салада терминжасамның бұрыннан қалыптасқан жүйесі бар.

Осымен байланысты, ұлттық терминология атаулы ғылым-білім, техниканың дамуына байланысты жаңа салалармен, күрделі ұғымдармен толығып отыратын болғандықтан және оларды терминологиялық сөздіктерде беру мәселелері арнайы зерттеудің нысанына айналмағандықтан, терминдердің тілдік табиғаты түсіндірмелі терминологиялық сөздіктерде толық ашылып, анықталып болды, түпкілікті зерттелді деп кесіп айту қиын.

Қазіргі кездегі қолданыста жүрген көп тілді дыбыстық сөздіктер тұрғысынан қарастырсақ, көп тілді дыбыстық сөздіктердің құрылымы сөздіктерді жасау және өңдеу үшін редактор және сөздіктерді қарастыру программасы. тұратынына тоқталдық. сөздіктерді жасау және өңдеу үшін редакторі, Мұнда сөздік жеке файл ретінде болады. Оны сөздіктер базасына қосып алуға және таратуға болады, ал сөздіктерді қарастыру программасы – әрбір сөздік, әр термин туралы ақпарат, мәліметтердің жиынтығы болып табылады. Сөздіктің мақсаты мен міндетіне қарай олардың әртүрлі терминологиялық сөздік түрлеріне бөлінетіндігі, түрлі атауға ие болатындығы көрсетілді.

Алда осы зерттеу жұмысын әртүрлі салалар бойынша деректер қорын одан әрі толтыра отырып, программаның көлемін үлкейтуге болады.және мобилді платформаларда орналастыруға тілді үйрену құралы ретінде дамытуға болады

Әдебиеттер

1.Монахов В.В. Материалы курса «Язык программирования Java» http://ru.sun.com/research/ materials/Monakhov_Java/

ТҮРІК ТІЛДЕРІНЕ ОҚЫТУДЫҢ ТЕХНОЛОГИЯЛАРЫ МЕН ИНТЕЛЛЕКТУАЛДЫ ЖҮЙЕЛЕРІ
ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ ДЛЯ ОБУЧЕНИЯ ТЮРКСКИМ ЯЗЫКАМ
INTELLIGENT SYSTEMS AND TECHNOLOGIES FOR LEARNING TURKIC LANGUAGES

ОМАРБЕКОВА А.С., ШАРИПБАЙ А.А.

Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан

ТЕХНОЛОГИЯ СОЗДАНИЯ ЭЛЕКТРОННЫХ УЧЕБНЫХ ИЗДАНИЙ НА ЛАТИНИЦЕ

В научно-исследовательском институте «Искусственный интеллект» ведутся исследования по переходу казахского языка с кириллицы на латиницу. В рамках финансируемого проекта «Методология, алгоритмы и программы генерации электронных учебных изданий» разработана технология создания электронных учебных изданий (ЭУИ). С помощью которой каждый преподаватель-непрограммист может создать ЭУИ со своим контентом, с унифицированным интерфейсом, составом элементов обучения и структурой управления, который полностью соответствует государственному стандарту СТ РК 34.017-2005 «Информационные технологии. Электронное издание. Электронное учебное издание» [1].

Подробнее о проекте можно ознакомиться на сайте **www.e-zerde.kz/metod.**

Система состоит из трех подсистем: авторской (для работы тьютора), пользовательской (для работы обучающегося), контролирующей (для работы администратора).

Авторская система

После установки программы в главном меню появится группа «Генератор ЭУИ», в ней четыре ярлыка: «Регистрация», «Справка», «Учебник», «Формирование ЭУИ».

При выборе пункта «Формирование» запускается программная оболочка разработки ЭУИ, первое окно которой предлагает выбрать язык ЭУИ (казахский кириллица, казахский латиница, русский, английский).

С помощью следующего окна вводится информация об авторах, аннотация, заголовок, содержание, состав урока, также возможен просмотр и сохранение ЭУИ (рисунок 1).

Рисунок 1

Кнопка «Abhtorlar» («Авторы») запускает программу формирования списка авторов ЭУИ.

Кнопка “Ahdatpa» («Аннотация») позволяет ввести текст аннотации, кнопка «Taqyryby» («Заголовок») - заголовок ЭУИ.

Кнопка «Mazmvny» («Содержание») позволяет сформировать трехуровневую структуру учебного материала, состоящую из модуля, блока и урока.

После окончания формирования структуры необходимо тщательно проверить наличие в каждом модуле всех блоков, в каждом блоке всех уроков.

Кнопка «Sabaq quramy» («Состав урока») открывает окно формирования содержимого урока, где будут указаны такие элементы урока, как: «Teoryja» («Теория»), «Mysaly» («Примеры»), «Tapsyrma» («Задания»), «Suraq» («Вопрос»), «Testiler» («Тесты»), «Mvltijmedyja» («Мультимедиа»), «Anyqtama» («Справочник»), «Tezawrvs» («Тезаурус») (рисунок 2).

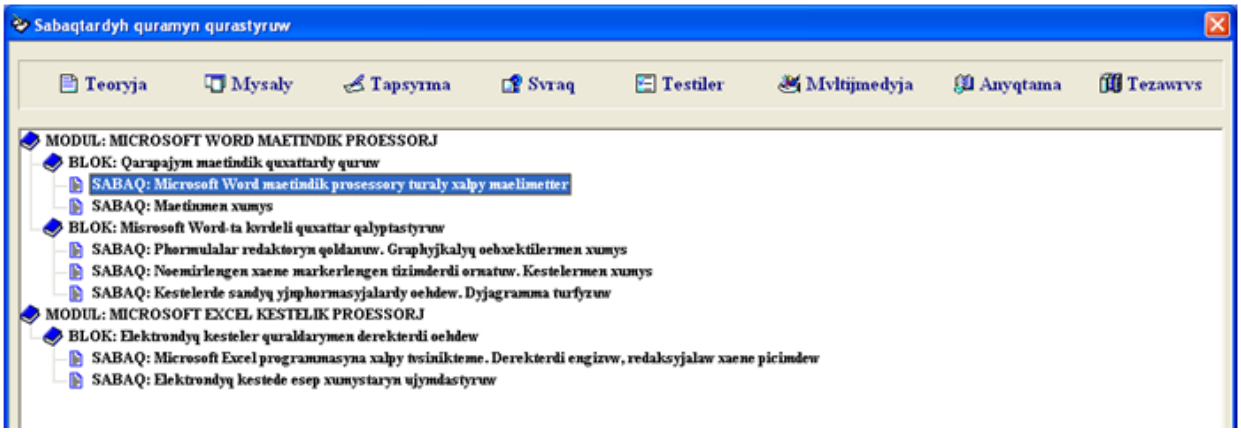


Рисунок 2

После ввода содержимого каждого урока, необходимо проверить корректность введенной информации. Для этого необходимо нажать кнопку «Qaraw» («Просмотр»). Отображается вся информация в том виде, как она будет показана обучаемому (рисунок 3).

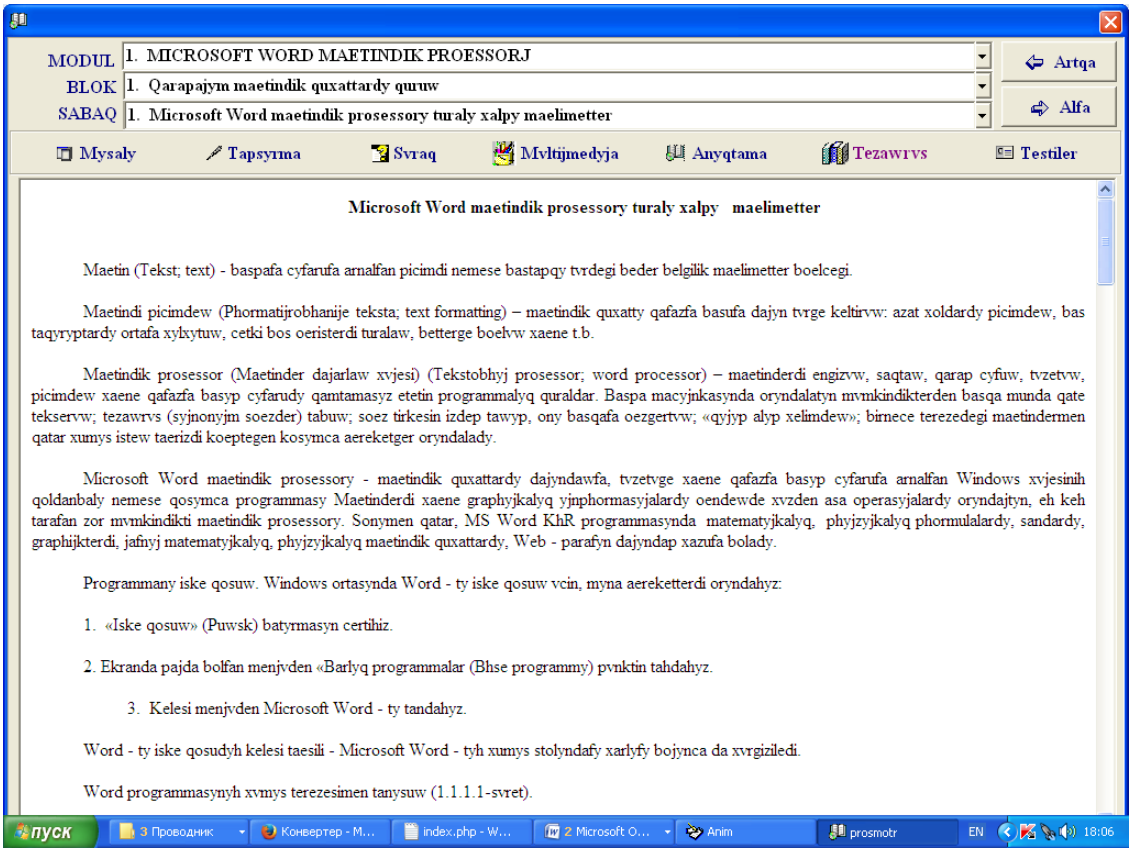


Рисунок 3

После того как редактирование содержимого баз данных, уроков завершено, внесены все данные, необходимо нажать кнопку «Saqtqw» (“Сохранить”).

Пользовательская система

Теперь можно просмотреть готовое ЭУИ, выбрав пункт главного меню «Учебник». Титул представлен на рисунке 4.

Рисунок 4

Кнопки «Abhtorlar», «Ahdatpa» отражают ту информацию, которая была введена при формировании содержимого ЭУИ. При нажатии кнопки «Taqyryptama» открывается окно наглядно отражающее структуру учебника.

Кнопка (“Mazmvny”)«Содержание» позволяет обучаемому выбрать режим работы (рисунок 5).

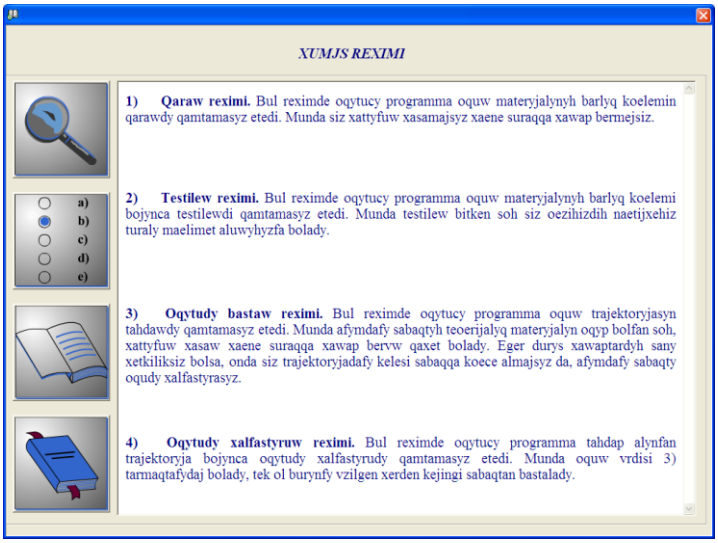


Рисунок 5

Первый режим просмотра. В этом режиме обучающая программа обеспечивает просмотр только учебного материала. При этом доступа к заданиям, вопросам, тестам не будет. (рисунок 6).

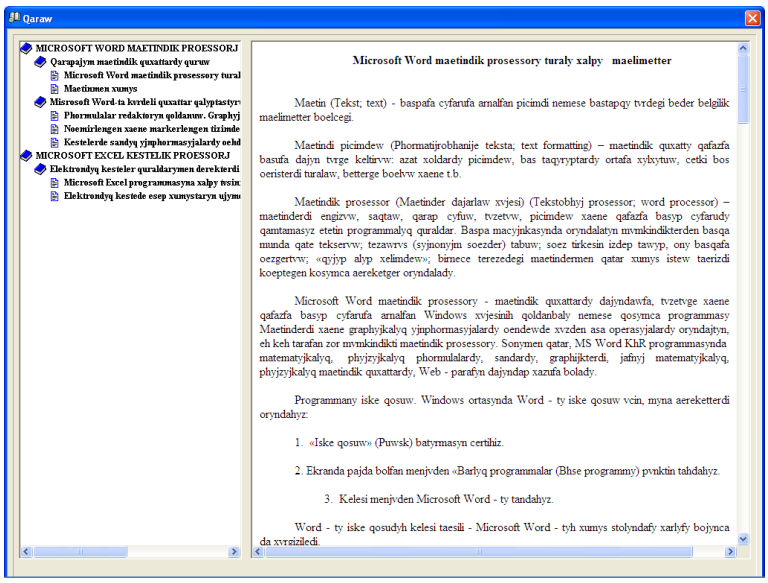


Рисунок 6

Второй режим тестирования. В этом режиме обучающая программа обеспечивает тестирование по всему объему учебного материала. При этом после тестирования можно получить информацию о результате тестирования (рисунок 7).

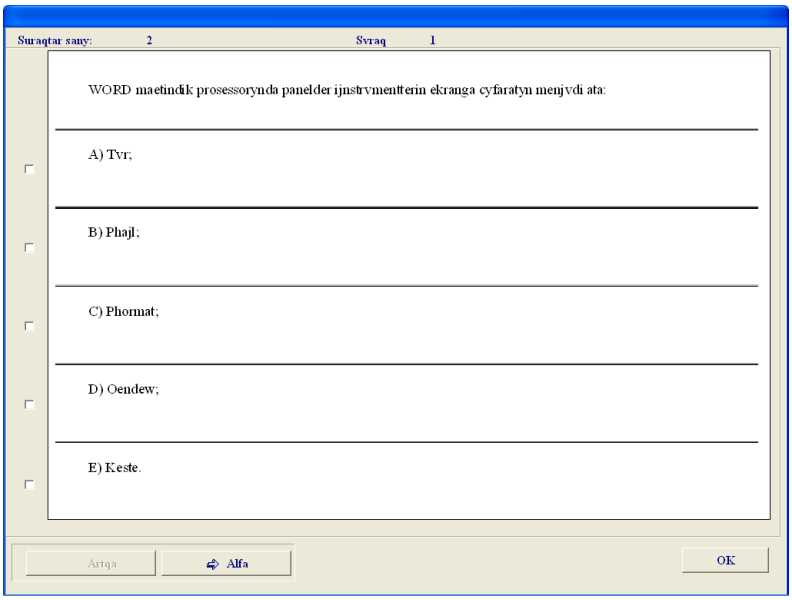


Рисунок 7

Третий режим начала обучения. Для начала обучения необходимо обучаемому зарегистрироваться.

В этом режиме обучающая программа обеспечивает выбор траектории обучения. При этом после изучения теоретического материала по текущему уроку необходимо будет отвечать на тестовые вопросы. В случае недостаточного количества правильных ответов на тесты, обучаемый не сможет перейти к следующему уроку в траектории и будет продолжать изучение текущего урока. Кроме текущего тестирования предусмотрены промежуточное тестирование (при переходе к следующему блоку), рубежное (при переходе к следующему модулю) и итоговое (при завершении обучения).

Четвертый режим продолжения обучения. В этом режиме обучающая программа обеспечивает продолжение обучения по выбранной траектории. При этом процесс обучения начинается со следующего урока после прерывания.

Режим начала обучения позволяет выбрать одну из трех траекторий обучения: ручной выбор, тестовый выбор и полный выбор (рисунок 8).

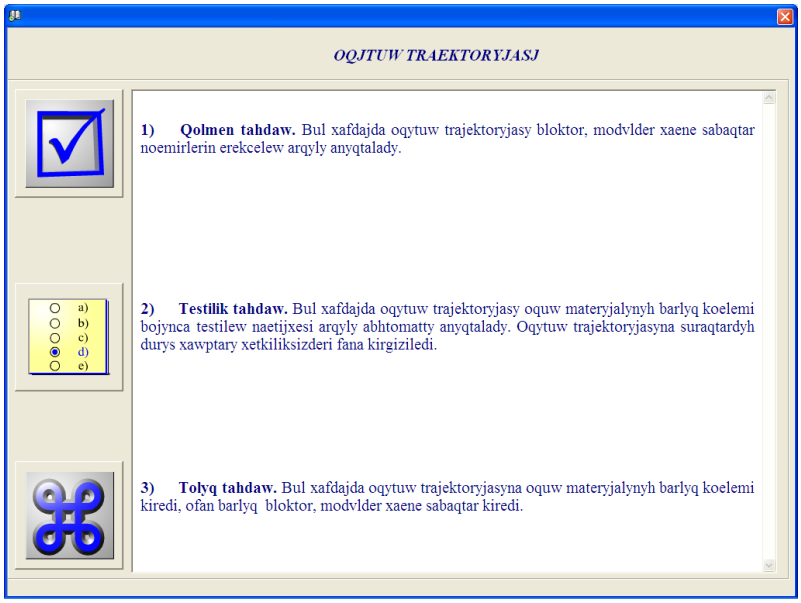


Рисунок 8

При ручном выборе траектория определяется обучаемым самостоятельно путем отметки номеров модулей, блоков, уроков.

При тестовом выборе траектория определяется автоматически по результатам тестирования по всему объему учебного материала. В этом случае в траекторию обучения включаются только те уроки, по вопросам которых были получены недостаточное количество правильных ответов. При полном выборе в траекторию включается весь объем учебного материала данной дисциплины, включая все уроки, модули и блоки.

После определения траектории пользователь переходит непосредственно к окну обучения (рисунок 9).

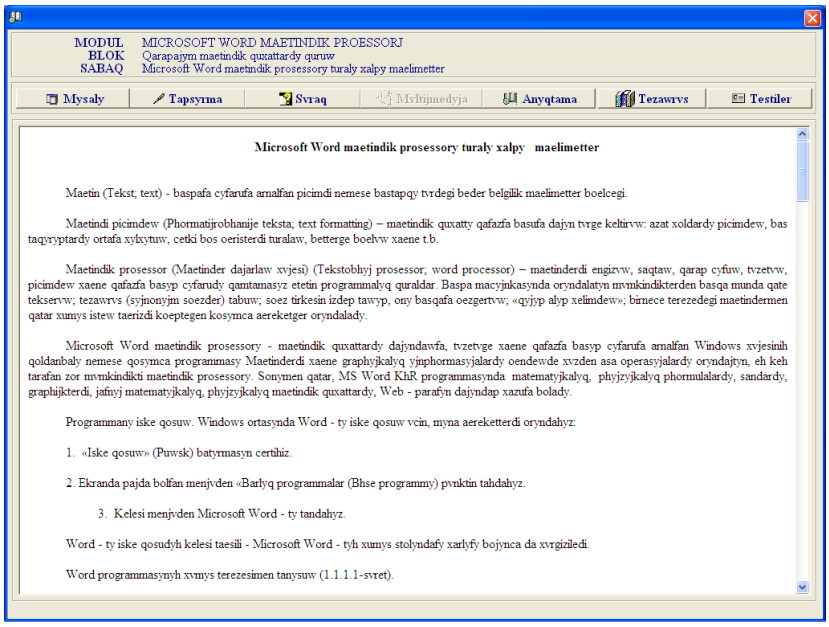


Рисунок 9

Таким образом, каждый преподаватель может создавать свое электронное учебное издание, структура и содержание которого зависят от целей его использования.

Контролирующая система

Для установки готового ЭУИ на другом компьютере необходимо запустить файл AutoRun.exe в текущей дирректории. Откроется следующее окно:

Рисунок 10

Выберите пукт «Электронное учебное издание». Запуститься программа установки ЭУИ. После установки в меню «Пуск» появится группа «Электронное учебное издание», в нем ссылки «Регистрация», «Электронное учебное издание», «Администрирование».

При выборе пункта «Администрирование» необходимо ввести пароль. Далее откроется форма позволяющая просмотреть список обучающихся, их успеваемость, также удалить записи обучающихся и сменить пароля администратора (рисунок 11).

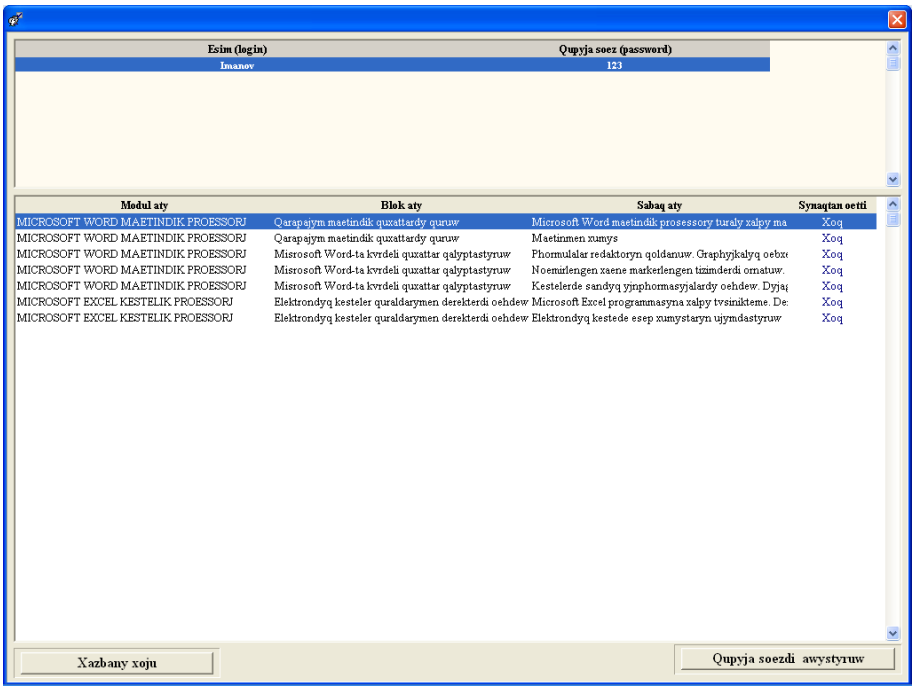


Рисунок 11

Таким образом, разработанная технология позволяет тьютору, не владеющему навыками программирования, в кратчайшие сроки самостоятельно создать качественное ЭУИ на казахском (латинице, кириллице), русском, английском языках [2].

Практическая ценность результатов работы состоит в том, что результаты проекта могут быть использованы преподавателями различных учебных учреждений для создания базы ЭУИ по всем предметам, что повысит качество и эффективность обучения.

Литература

1. Абдыманапов С.А., Шарипбаев А.А, Омаров А.Н., Баймуратова Г.Г., Нургужин М.Р., Байгелов К.Ж., Альжанов А.К., Омарбекова А.С. СТ РК 34.017-2005 «Информационная технология. Электронное издание. Электронное учебное издание».

2. Omarbekova A.S., Seifullina A. Automatization create electronic learning. // Journal of International Scientific PublicationsEducation Alternatives - Volume 10, Part 1, ISSN 1313-2571. Bulgaria, 2012– С.242-250

А.Т.АЛСЕИТОВА, Р.С.НИЯЗОВА

Л.Н.Гумилев атындағы Еуразия Ұлттық университеті, Астана қ., Қазақстан.

АВТОМАТТАР ТЕОРИЯСЫ БОЙЫНША МУЛЬТИМЕДИАЛЫҚ ОҚЫТУ ҚҰРАЛЫН ЖАСАУ

Қоғамның қазіргі даму кезеңі жедел ақпараттандырылу процесімен сипатталады. Ақпараттық компьютерлік технологиялардың білім беру жүйесіне енуі жаңа тарихи – әлеуметтік жағдайлардың талабы болып отыр. Білім беру жүйесіндегі жаңа ақпараттық технология дегеніміз – оқу және оқу-әдістемелік материалдар жинағы, оқу қызметіндегі есептеуіш техниканың техникалық құралдары, олардың рөлі мен орны туралы ғылыми білімнің жүйесін және оқытушылар еңбектерін жүзеге асыру үшін оларды қолдану формалары мен әдістері деген анықтама беруге болады. Яғни, ақпараттық технология – білім беру мекемесі мамандарының жұмысын жүзеге асырушы әдістер мен формалар және оқушыларға білім беруші құрал.

ҚР-ның «Білім туралы» Заңында «Білім беру жүйесінің басты міндеті – ұлттық құндылықтар, жеке адамды қалыптастыруға және кәсіби шыңдауға бағытталған білім алу үшін қажетті жағдайлар жасау, оқытудың жаңа технологияларын енгізу, білім беруді ақпараттандыру, халықаралық ғаламдық желілерге шығу» делінген. Осыған орай, қазіргі уақытта білім беру жүйесінде компьютерлік технология кең қолданысға ие болды. Ақпараттық технологиялар оқу материалдарын иллюстрация жасау кезінде (мысалы анимациялы слайд-фильмдер) қолданылады. Компьютерлік графика мен визуалды білім мүмкіндіктерімен байланысты еңбектер дамытылды. Оқыту бағдарламасында компьютерлік графиканы қолдану шығармашылық жеке адамды тәрбиелеуді жүзеге асырады. Бұл оқу үрдісін қозғалыста бейнелеуге мүмкіндік береді. Компьютердің көмегімен дыбыстық және бейнефрагменттерді де демонстрация жасауға болады. Менің магистрлік жұмысым, терең зерттеуді қажет ететін автоматтар теориясы бойынша мультимедиалық оқыту құралын жасау болып табылады.

Мультимедиалық технологиялар мультимедиа өнімдерінің жасалу үрдісімен байланысты яғни, электрондық кітаптар, мүмкіндігінше энциклопедиялар, компьютерлік фильмдер. Бұл өнімдердің өзіндік ерекшеліктері мәтіндік, графикалық, аудио-бейне, ақпараттық анимацияның бірігуі болып табылады. Мультимедиа технологиясы компьютерді тең дәрежедегі әңгіме-дүкен құрушыға айналдырып қана қойған жоқ, білім алшыға бір орында отырып, үлкен ғалымдар мен педагогтердің дәрістеріне қатысуына, өткен және қазіргі тарихи оқиғаларға куә болуына, әлемнің ең белгілі мұражайлары мен мәдени орталықтарына,

жер шарының ең алыс және қызық түкпірлеріне сапар шегуіне, белгілі бір салалар бойынша түсінік қалыптастыруға мүмкіндік жасайды.

Информатика адамның интеллектуалды өміріндегі әртүрлі есептерді шығару және жеке үдерістерді автоматтандыру үшін жасанды тілдерде жазылған программалар арқылы компьютерде жүзеге асырылатын ақпараттық технологияларды жасаумен зерттеу проблемаларымен айналысады.

Заманауи ақпараттық технологияларда лингвистикалық (морфологиялық, синтаксистік, семантикалық және сөйлеу) анализ бен синтез әдістері маңызды рөл атқарады. Әртүрлі тілдік процессорлар (трансляторлар, компиляторлар, интерпретаторлар, конверторлар, редакторлар және т.б.) лингвистикалық әдістерді қолдануға негізделген. Тілдер мен автоматтар теориясы информатика ғылымының іргелі саласы бола тұра, осы әдістердің ғылыми негізін құрады және тілдерді тудыру мен тану механизмдерін зерттеумен шұғылданады.

Тудыру механизмдер (формалды грамматикалар) теориясының негізін XX-ғасырдың 40–50 жылдары табиғи тілдерге арналған лингвистикалық жұмыстарға байланысты Н.Хомский (N.Chomsky) қалады. Ал танушы механизмдер (ақырлы автоматтар) теориясының негізін сол жылдары М.О.Рабин (*M.*O.Rabi*n*), Д.Скотт (D.Scott) және В.М.Глушков жасады. Сол кезде тілдер мен автоматтар теориясы программалау тілдерін жасау және жүзеге асыру саласында кең практикалық қолданыс тапты.

Қазір лингвистикалық әдістерге негізделген технологиялық жабдықтардың қолдануы кең етек жайды. Олар табиғи тілдерді өңдеу жүйелерін, соның ішінде орфографиялық түзеткіштерді, морфологиялық және синтаксистік талдау, мағыналы іздеу мен шешім қабылдау жүйелерін, сонымен қатар, сөйлеу технологияларын жасау кезінде қолданылады. Дегенмен оларды сауатты және тиімді қолдану пайдаланушыдан, ең кемінде, олар негізделетін математикалық теорияны білуді талап етеді.

Автоматтар теориясы бойынша жасалатын мультимедиалық оқыту құралы әр жылдары Информатика, Ақпараттық жүйелер, Есептеу техникасы және программалық қамтама мамандықтары бойынша оқытылатын болады.

Аталмыш құралды дайындауға мазмұны информатика мамандығының жалпы білім беру мемлекеттік стандартына сәйкес болатындай оқу басылымдарының жоқтығы себеп болды. Сондықтан оқу процесін толыққанды қамту үшін, ішінде бір ғана ұғым әртүрлі терминдер, анықтамалар және белгілеулерге ие болатын, түрлі әдебиеттер мен интернет ресурстары пайдаланылады. Бұл жағдай бір ұғымға тек бір ғана термин, анықтама және белгі пайдаланатын осы оқулықты жазу идеясына әкеліп соқтырды. Жасалынатын оқулықтың оқылатын пән бойынша оқу басылымдарының жетіспеушілігін жөндеуге, әсіресе осы пәнге бағытталған мультимедиалық оқулықтың жоқтығын ескере отырып, бірдей ұғымдардың терминдерін, анықталуларын және белгілеулерін бірыңғай етуге мүмкіндік береріне үміт артады.

Оқу материалын беру әдістемесе мультимедиалық құралды пайдаланушыны тілдер мен автоматтардың классикалық теориясының іргелі фактілері туралы айтуға және оқылатын саладағы тұжырымдарды дәлелдеу әдістерімен таныстыруға, сонымен қатар, оны қандайда бір мысалдармен жабдықтауға негізделген. Теоремалар дәлелдеулерінің барлығы жуық конструктивті сипат алады, сондықтан олар практикаға пайдалы материал береді.

Мультимедиалық оқыту құралы орыс және қазақ тілдерінде бірдей мазмұнды жасалынады және қолданылатын белгілеулер, ұғымдар, түсініктер, қысқартулар, математикалық негіздер және тілдерді анықтау механизмдері талқыланып, регулярлық тілдерді тудырушы механизмдер (регулярлық жиындар, регулярлық өрнектер, регулярлық алгебра, регулярлық теңдеулер мен олардың жүйесі, оң сызықты грамматика), регулярлық тілдерді танушы механизмдер (бейдетерминді және детерминді ақырлы автоматтар және олардың эквиваленттігі), регулярлық тілдердің қасиеттері (регулярлық өрнектер, оң сызықты грамматикалар және ақырлы автоматтардың эквиваленттілігі) және олардың алгоритмдік проблемалары қарастырылады. Сонымен қатар контекстісіз тілдерді тудырушы механизмдер (контексті-бос грамматикалар), контекстісіз тілдерді танушы механизмдер (бейдетерминді

мен детерминді стекті автоматтар және олардың эквиваленттілігі), контекстісіз тілдердің қасиеттері (контексті-бос грамматикалармен стекті автоматтардың эквиваленттілігі) және олардың алгоритмдік проблемалары беріледі. Контексті тілдерді тудырушы механизмдер (контексті-тәуелді грамматикалар), контексті тілдерді танушы механизмдер (бейдетерминді мен детерминді сызықты шенеуленген автоматтар және олардың эквиваленттілігі), контексті тілдердің қасиеттері (контексті-тәуелді грамматикалар және сызықты шенеуленген автоматтардың эквиваленттілігі) және олардың алгоритмдік проблемаларына да тоқталамыз.

Бесінші бөлімде шенеуленбеген тілдерді тудырушы механизмдер (шенеуленбеген грамматикалар),шенеуленбеген тілдерді танушы механизмдер (бейдетерминді және детерминді Тьюринг машиналары және олардың эквиваленттілігі),шенеуленбеген тілдердің қасиеттері (шенеуленбеген грамматикалар және сызықты шенеуленген автоматтардың эквиваленттілігі) және олардың алгоритмдік проблемалары ұсынылады.

Мультимедиалық оқыту құралы студенттерге, магистранттарға, докоранттарға, оқытушыларға, ғалымдарға және тілдер мен автоматтар теориясын өздігінен оқып білем деген барлық азаматтарға, оның ішінде қашықтықтан білім алушыларға арналған.

«Түркі тілдерін компьютерлік өңдеу»
атты I халықаралық конференция
ЕҢБЕКТЕРІ

ТРУДЫ
I Международной конференции
"Компьютерная обработка тюркских языков"

PROCEEDINGS
Of the I International Conference
on Computer processing of Turkic Languages (TurkLang-2013)

Типография ЕНУ им. Л.Н. Гумилева
Г. Астана, ул. Мунайтпасова, 13