Article
IRSTI 16.31.21 https://doi.org/10.55491/2411-6076-2024-2-172-181

N. Amirzhanova,[1] D. Sadyk[2]
[1]Candidate of Philological Sciences, Nazarbayev University,
Kazakhstan, Astana, e-mail: nazira.amirzhanova@mail.ru
[2]Corresponding author, Doctoral student, Al-Farabi Kazakh National University,
Kazakhstan, Almaty, e-mail: sadyk_didar@mail.ru

SCIENTIFIC AND PRACTICAL PROBLEMS OF AUTOMATION OF THE KAZAKH ORTHOGRAPHY THROUGH THE METHOD OF MODELING

Abstract. The rapid development of digital technologies, including the emergence of the Internet, led to a radical change in the communicative landscape of the 20th century. The possibility of constant access to the Internet, as well as the use of applications for the exchange of instant messages online gave an impetus to the formation of a new digital language, the development of a new language environment. Currently, the Kazakh language entered the digital space and the life force of the Kazakh language was formed in the virtual space. In this regard, an inventory of knowledge related to the Kazakh language was carried out and the development of Kazakh language resources in the virtual space was put on the agenda.

Automation of orthography is the process of using computer programs and tools to check the spelling of words in a text. This allows one to automatically correct spelling errors and reduce the time spent on text editing. Innovative functions require the development of a formal language model. And for formal modeling of orthography, first of all, it is necessary to conduct an inventory of knowledge related to orthography, to identify difficulties in spelling. Currently, scientists of the Institute of Linguistics named after A. Baitursynuly are conducting research within the framework of the program-targeted financing project "Automatic recognition of Kazakh text: development of linguistic modules and IT solutions". As a part of the research, linguistic resources are differentiated and analyzed, allowing to automatically correct errors in word spelling.

The article analyzes in detail the automation of knowledge related to orthography, scientific and practical problems of its automatic representation, modeling of knowledge related to Kazakh orthography, definition of model concepts, common types of modeling, types of modeling used in automation. Also, for the automation of Kazakh orthography, types of modeling using phonetic-phonological and orthographic distinctive signs are proposed, the stages of automation of Kazakh orthography are described.

Keywords: orthography, automation, model, modeling, orthogram.

Н.С. Әміржанова,[1] Д.А. Садық[2]
[1]филология ғылымдарының кандидаты, Назарбаев университеті,
Қазақстан, Астана қ., e-mail: nazira.amirzhanova@mail.ru
[2]автор-корреспондент, докторант, Әл-Фараби атындағы Қазақ ұлттық университеті,
Қазақстан, Алматы қ., e-mail: sadyk_didar@mail.ru

МОДЕЛЬДЕУ ӘДІСІ АРҚЫЛЫ ҚАЗАҚ ЕМЛЕСІН АВТОМАТТАНДЫРУДЫҢ ҒЫЛЫМИ-ПРАКТИКАЛЫҚ МӘСЕЛЕЛЕРІ

Аңдатпа. Цифрлық технологиялардың қарқынды дамуы, соның ішінде интернеттің пайда болуы XX ғасырдағы коммуникациялық ландшафттың түбегейлі өзгеруіне әкелді. Интернетке тұрақты қол жеткізу мүмкіндігі, сондай-ақ онлайн түрде жедел хабар алмасу қосымшаларын пайдалану жаңа цифрлық тілдің қалыптасуына, жаңа тілдік ортаның дамуына серпін берді. Қазіргі кезде қазақ тілі цифрлық кеңістікке еніп, қазақ тілінің виртуалды кеңістіктегі өміршеңдігі нысанға алынды. Осыған байланысты қазақ тіліне қатысты білімдер түгенделіп, қазақ тілінің виртуалды кеңістіктегі ресурстарын әзірлеу күн тәртібіне қойылды.

Орфографияны автоматтандыру – бұл мәтіндегі сөздердің емлесін тексеру үшін компьютерлік бағдарламалар мен құралдарды пайдалану процесі. Бұл сөздердің жазылуындағы қателерді автоматты түрде түзетуге мүмкіндік береді және мәтінді өңдеуге кететін уақытты азайтады. Инновациялық мүмкіндіктер тілдің формалды моделін әзірлеуді талап етеді. Ал орфографияны формалды модельдеу үшін ең алдымен орфографияға қатысты білімдерді түгендеу, емледегі қиындықтарды саралау қажет. Қазіргі кезде А.Байтұрсынұлы атындағы Тіл білімі институты ғалымдары «Қазақ мәтінін автоматты тану: лингвистикалық модульдер мен IT-шешімдер әзірлемесі» атты бағдарламалық-нысаналы қаржыландыру жоба аясында зерттеу жүргізіп келеді. Зерттеу аясында сөздердің жазылуындағы қателерді автоматты түрде түзетуге мүмкіндік беретін тілдік ресурстар саралану, талдану үстінде. Осыған байланысты мақалада орфографияға қатысты білімдерді автоматтандыру, оны автоматты танытудың ғылыми-практикалық мәселелері, қазақ орфографиясына қатысты білімдерді алгоритмдеу түсініктерінің анықтамасы, кең таралған модельдеу түрлері, автоматтандыруда ұстанылатын модельдеудің типтері жан-жақты талданады. Сондай-ақ қазақ орфографиясын автоматтандыру үшін фонетика-фонологиялық және

орфограммалардың айырымдық белгілері арқылы модельдеудің түрлері ұсынылып, қазақ орфографиясын автоматтандырудың кезеңдері сипатталады.

Тірек сөздер: орфография, автоматтандыру, модель, модельдеу, орфограмма.

Н.С. Амиржанова,[1] Д.А. Садык[2]
[1]кандидат филологических наук, Назарбаев университет,
Казахстан, г. Астана, e-mail: nazira.amirzhanova@mail.ru
[2]автор-корреспондент, докторант, Казахский национальный университет имени аль-Фараби,
Казахстан, г. Алматы, e-mail: sadyk_didar@mail.ru

НАУЧНО-ПРАКТИЧЕСКИЕ ПРОБЛЕМЫ АВТОМАТИЗАЦИИ КАЗАХСКОЙ ОРФОГРАФИИ ПОСРЕДСТВОМ МЕТОДА МОДЕЛИРОВАНИЯ

Аннотация. Стремительное развитие цифровых технологий, в том числе появление интернета, привело к радикальному изменению коммуникативного ландшафта XX века. Возможность постоянного доступа в интернет, а также использование приложений для обмена мгновенными сообщениями в режиме онлайн дали толчок к формированию нового цифрового языка, развитию новой языковой среды. В настоящее время казахский язык вошел в цифровое пространство, сформировалась жизнеспособность казахского языка в виртуальном пространстве. В связи с этим проведена инвентаризация знаний, касающихся казахского языка, и поставлена на повестку дня разработка ресурсов казахского языка в виртуальном пространстве.

Автоматизация орфографии – это процесс использования компьютерных программ и инструментов для проверки правописания слов в тексте. Это позволяет автоматически исправлять орфографические ошибки и сокращать время, затрачиваемое на редактирование текста. Инновационные функции требуют разработки формальной языковой модели. А для формального моделирования орфографии прежде всего необходимо провести инвентаризацию знаний связанных с орфографией, выделить трудности в правописании. В настоящее время ученые Института языкознания имени А.Байтурсынулы проводят исследования в рамках проекта программно-целевого финансирования «Автоматическое распознавание казахского текста: разработка лингвистических модулей и IT-решений». В рамках исследования дифференцируются и анализируются лингвистические ресурсы позволяющие автоматически исправлять ошибки в написании слов.

В статье подробно анализируются автоматизация знаний, связанных с орфографией, научные и практические проблемы ее автоматического представления, моделирование знаний, связанных с казахской орфографией, определение модельных понятий, распространенные виды моделирования, виды моделирования, используемые в автоматизации. Также для автоматизации казахской орфографии предложены виды моделирования с использованием фонетико-фонологических и орфограммных отличительных признаков, описаны этапы автоматизации казахской орфографии.

Ключевые слова: орфография, автоматизация, модель, моделирование, орфограмма.

Introduction

Nowadays, the issue of automatizing the culture of the written Kazakh speech is being considered, and the data related to the graphics and orthography of the Kazakh language is accumulated. This necessity is borne, firstly, out of increasing the livability of the Kazakh language in the virtual space and, secondly, out of improving the process of obtaining linguistic data through utilizing contemporary technological advancements.

There is no compensation for the request to transfer Kazakh texts on paper into electronic format, processing, obtaining various information, increasing the number and quality of electronic textbooks and content. Compared to a scanned graphic file, the electronic format is ideal: it reduces losses on information storage, distribution, use of the document, and allows to realize all possible scenarios of analysis. There are such programmes in active use, for example ABBYY FineReader. But, despite the fact that their algorithms are effective, printed, handwritten English characters obtained in the database are inadmissible for recognition of Kazakh alphabet characters. Therefore, to begin with, it is important to provide spelling automation for the recognition of Kazakh texts with built-in linguistic modules: they "penetrate" into the electronic text, including information of interest to the consumer on the elements of the text.

To automatize Kazakh orthography, it is necessary to develop the base of linguistic algorithms and models. It is only through models that one can create textual editors, autocorrecting tools, and programs of analysis and synthesis. Hence the orthographic models may serve as a material for creating a technology of the new methods of error prevention and identification.

Materials and methods

To automatize the Kazakh orthography, conceptual assumptions and works of Kazakh grammotologists are taken into consideration. In particular, the works of such prominent professors as R. Syzdyq, N. Uali, and Q. Kuderinova are taken as a basis.

Apart from such commonly scientific methods as modeling, algorithm creation, analysis, differentiation and generalization, and compiling, the methods of distributive analysis, formal linguistic analysis, and orthographic modeling are utilized. Orthographic modeling has become quite a significant method of scientific cognition and research. The method of modeling is used in every science, on every stage of scientific cognition; its heuristic power is immense. The mentioned power is identified through the following: via this method, it is possible to ease the difficult, to turn the invisible into visible and intangible into tangible, the unknown into known – hence the method allows to study a complex phenomenon in a thorough and multifaceted way.

Literature review

The issue of automatizing the orthography is viewed in relation to the fields of computational, applied, and mathematical linguistics and to the method of modeling which is widespread in these fields. Modeling a language is the process of creating a compressed and abstract variant of a language which serves for researching various linguistic phenomena. A model may be presented as a scheme, a mathematical formula, or as a computer program. Linguistic modeling is usually utilized in researching phonetics, morphology, syntax, and semantics. It provides linguists with an insight into how language works and how its elements interact.

Modeling is the process of creating models which are necessary for researching and analyzing a system. Linguistic modeling is helpful for automatic translation, cognizing a word, analyzing a text, creating virtual helpers and other natural language processing applications. Also, linguistic modeling is useful in studying language and its structure, and developing new theories related to linguistic phenomena.

A model is a simplified form which aids in analyzing and comprehending complex processes. Hence modeling refers to creating linguistic models which aid in analyzing and comprehending linguistic phenomena. These models consist of algorithms which help to understand the language through statistical, semantic, or syntactical methods.

Creating linguistic models is extremely necessary for analyzing texts, studying grammatical rules, encompassing the research of lexical interrelations and other linguistic phenomena, and preserving the literacy in the virtual space. Apart from preserving literacy, modeling is significant for developing computer programs which are capable of synthesizing speech and recognizing speech, translating texts, analyzing text content and even creating new texts. Thus, linguistic modeling may be used for studying the language and its structure. This helps linguists and other researchers create new theories related to linguistic phenomena and develop exact versions of language. In general, linguistic modeling is a significant tool for researching and analyzing natural language. It may be used to create computer systems and study the language and its structure.

A model refers to demonstrating a concrete object or process in a compressed and exact way. A model is a prototype developed for a computer. Various models of the same object may be developed. For instance, in automatizing orthographic knowledge, different modes of modeling may be utilized. In other words, a model can be described as creating an abstract or written plan of performing a certain activity. Hence modeling refers to creating a scheme of objects and processes.

In linguistics, there are various scholars who study modeling. For instance, one may note Noam Chomsky, an American linguist who created the theory of generative grammar and who utilized modeling for language research. Noam Chomsky is a prominent linguist who contributed to the field of linguistic modeling. Studying generative grammar theory, he claimed that a language relies on certain rules, and the number of correct sentences can be increased through these rules. N. Chomsky believed his theory to allow for better understanding of linguistic phenomena and easing the process of learning foreign languages (Chomsky N., 2000).

George Lakoff is an American linguist who utilized modeling to study metaphors. He suggested a conceptual theory which describes how we use well-known notions and images to comprehend new or

intangible ideas. G. Lakoff claims that metaphors are not only the means of furnishing the speech, but also the main tool through which we think and understand the world. Also, the scholar modeled metaphors, showing their usage in different spheres starting with politics and ending with science. For instance, he showed how the metaphor "life is game" is utilized in different contexts to describe the situations requiring strategic thinking and risk-taking. Additionally, he studied the metaphor of "emotions" and identified its influence on our understanding of emotions and their representation. G. Lakoff examines the traditional method of linguistic modeling and proves its excessive formality and its omission of context and cultural differences. The scholar suggests the use of flexible and contextual modeling methods which consider the interrelation of language and thinking, and claims that metaphors are utilized not only for colorizing the speech, but also for comprehending the world (Lakoff, 1996).

Another notable name is Richard Montague, an American scholar who studied the issue of linguistic modeling in its relation to semantics. He is known for developing semantics and utilizing modeling for studying the meaning of words and sentences. The scholar also created the linguistic model based on the ideas of categorization and types theory. Richard Montague, studying lexemes and identifying the main principles of modeling them, makes the following assumptions:

1. A language can be described as a collection of semantic units (lexemes) which refer to specific notions.

2. Every lexeme has a variant which defines its syntactical and semantic function in a sentence.

3. A sentence consists of lexemes interconnected through logical operators and quantitative links.

4. The meaning of a sentence is defined via the meaning of every lexeme that constitutes its structure and context.

Richard Montague applies these findings to the automatic analysis and generation of natural language texts. The scientist creates a system of formal semantics that allows to accurately determine the meaning of a sentence. He argued that only such an approach can ensure accuracy and reliability in automatic language processing. One of Richard Montague's main works on this topic is called *Language, Thought and Reality*.

One of the scholars who studied the theory of modeling and drew unique conclusions is Martin Kay. The scholar made several assumptions related to modeling. Namely:

"1. Modeling is an important scientific tool which allows to exactify and verify the hypothesis.

2. Modeling might come in multiple forms: from mathematical models to computer modeling.

3. It is important to understand that models do not depict reality fully, but only reflect it partially.

4. Modeling might lead to comprehending complicated processes, but it may also lead to mistakes and erroneous conclusions" (Martin, 2000).

Another scholar who examined the theory of modeling is David Black. In his work, Black reviews the issues of modeling in science and answers the questions related to how mathematical models are utilized in linguistics and how they are connected to reality.

Along with the mentioned scholars, the names of S.I. Arkhangelskiy, V.V. Yermilova, and M.Vortofskiy can be noted in relation to automatizing and modeling language. S.I. Arkhangelskiy studied the theory of modeling and divided it into simple and complicated, schematic and discrete (Arkhangelskiy, 1980; Yermilova, 2001). S.I. Arkhangelskiy notes that every modeling process requires preparation and demonstrates that the schematic model is the best for modeling language-related knowledge. At the same time, M. Vortofskiy claims that modeling is not only collecting knowledge, but also "creating future" (Vortofskiy, 1988).

In the language automatization field of Kazakh linguistics, the names of Q. Bektayev, A.Zhubanov, A. Zhanabekova can be noted. Q. Bektayev developed the frequency dictionary of Kazakh syllables and introduced the pioneering field of Kazakh linguistics through utilizing the methods of analysis and synthesis of language materials. As a result, computational linguistics, applied linguistics, mathematical linguistics and other such disciplines became a part of Kazakh language studies. A. Zhubanov is among the scientists who studied the applied and mathematical sides of language and made unique assumptions in the field of automatization. A. Zhubanov claims: "The process of finding necessary information is being carried out via traditional (non-automatic) methods. However, it is proven that using such methods of retrieving information produces negative influence on the scholar's

time efficiency. For instance, as statistics shows, the user of traditional methods spends 80 percent of their time on finding, analyzing, and selecting the necessary data. For these reasons, the innovations in the field of data service require new methods of processing linguistic data" (Zhubanov, 2012). Indeed, the issue is relevant. Contemporary linguistics requires new inquiries and new methods of automatizing language.

According to the findings of the above-mentioned scientists, the concept of linguistic model first appeared in structural linguistics. We would like to say that concepts of model and modeling are often used in language teaching as well as in applied linguistics because modeling creates analytical capabilities.

Results and discussions

The cognitive-linguistic base of Kazakh orthography has been forming for centuries. The cognitive-linguistic base includes orthological tools (dictionaries, spelling rules, manuals). These tools serve as the resource of automatizing the orthography, thus helping write certain words and expressions correctly. The resources are helpful for verifying the spelling and grammar (proper utilization of grammatical rules and entering corrections in case the word combination is written erroneously), for auto-completion (in order to be able to finish the words and word combinations correctly), and for translation (to translate words and word combinations into other languages).

Modeling can be implemented in the form of a formula, table, writing in words, block diagram. Scientists divide it into two types depending on its usage: verbal and schematic. If the sequence of actions to be performed is given in words or sentences, this is counted as verbal modeling. Verbal modeling can be used for Kazakh punctuation. And schematic modeling is effective for spelling.

In order to model Kazakh orthography, it is necessary, first of all, to determine the types of invariants, variants, and variations in the theory of phonology, secondly, to find the position of the orthogram, differential signs of orthograms, and its variants, and thirdly, to propose types of modeling depending on the type and kind of orthogram. This system is the main linguistic goal for the automation of Kazakh orthography.

The linguistic basis of developing orthographic modeling is studied in relation to identifying the invariant phoneme in the theory of phonology. In Kazakh phonology, it is possible to create phonetic-phonological models which serve as the main and primary stage of orthographic modeling, based on the theoretical conceptions of N. Uali (Uali, 1993) and Z. Bazarbayeva (Bazarbayeva, 2008).

The invariant phoneme is represented by a letter. When it comes to variants (*басшы – башшы*) and variations (*хатшы – хатчы*) of the invariant, they are only audial, not visually represented. Automatizing orthography requires clarifying variant and variation phenomena. The reason is, writing does not take into account the additional phonic meanings, i.e., variants of the phonemes which appear due to their positional and combinatory changes. Such positional and combinatory changes of sounds and their auxiliary meanings are the characteristic of speech only. Due to the omission of such linguistic tendencies in writing, orthographic difficulties arise. That is why we believe it is necessary to create phonetic-phonological and orthographic models while identifying the shades of phonemes in weak positions. This is the base of automatizing orthography.

Phonetic-phonological model is identified via the triad *letter-phoneme-sound*. As an example, the phonemes *а* and *л* can be shown.

Table 1 – Phonetic-phonematic modeling

№	letter	phoneme	sound	phonation	inscription
1	*а*	*а*	*[а]* *[ä]*	*[қара]* *[кітän]* *[жäйлäу]*	*қара (look/black)* *кіtап (book)* *жайлау (summer pastures)*

Continuation of the table 1

2	*ң*	*ң*	*[ң]* *[ң']* *[ң˚]* *[ң˚']*	*[маң]* *[м'ең']* *[м˚оң˚]* *[м˚үң˚үл˚']*	*маң (morning)* *теңң (equal)* *моң (frost)* *мүңіл (get disappointed)*

The model above helps identify variants and variations of every Kazakh phoneme in their weak positions. This is, beyond doubt, the linguistic-cognitive knowledge necessary for automatization. Hence this process is the primary work which needs to be done while automatizing Kazakh orthography. The second stage encompasses such linguistic work as systematizing and complementing knowledge related to orthography and orthograms. Orthographic knowledge is primarily connected to orthographic rules. The orthographic rules are developed by the linguistic-cognitive model of any orthographic difficulties. Orthographic rules are laws which manifest different features of orthograms, i.e., the definition of graphic signs. The phonemic motivation of the graphic sign of writing is reflected in the content of any rule of orthography. The unit of spelling rules is included in the phonemic motivation, as is the unit of orthography. Graphic signs of writing are motivated by the basic shades of phonemes.

The notions of orthogram and orthographic rule are symmetric. Orthographic rules guide towards writing correctly and serve for avoiding or preventing difficulties in writing practice. While automatizing orthographic knowledge, orthographic rules and distinctive features of orthograms are taken into consideration. That is why, while automatizing the orthographic knowledge, tendencies of Kazakh orthography, types and kinds of orthograms, their distinctive features, and their variants must be considered.

The notion of distinctive features of orthograms was used by N.N. Algazina, the scholar in the field of Russian language teaching methodics, in the middle of the 60s of XX century (Algazina, 87). She studies distinctive features of orthograms in relation to the issues of attentiveness and agility. In 1970s, M.T. Baranov studied distinctive features of orthograms and defined them as specific signs which can be encountered within or between words, describing them as the conditions of choice which lead to proper inscription of words. Thus, distinctive features of orthograms are the signs which depict the presence of an orthogram within a word or between two words, the signs that show the usage of rules. Certain (phonetic, phonological, semantic) peculiarities of words are also the features of orthograms.

Distinctive features of orthograms are evident when a letter and a sound do not correlate in writing and speaking. Such features are only activated when a word is heard and its letter image comes to mind simultaneously. Distinctive features of orthograms are demonstrated in the position of “dangerous” sounds and sound combinations (letters and letter combinations). This poses difficulties mainly in orthography. Also, to identify the distinctive features of orthograms, it is necessary to be able to find morphemes and morphemic combinations within words. An artificial intelligence trained to identify morphemes looks for an orthogram in a programmed way, as the models of how orthograms are presented in roots, affixes, or between morphemes. Those distinctive features might be described as common for various orthograms. Apart from common distinctive features, there are certain features which characterize types or kinds of orthograms. To clarify, we propose the following model of automatizing orthograms (Table 2).

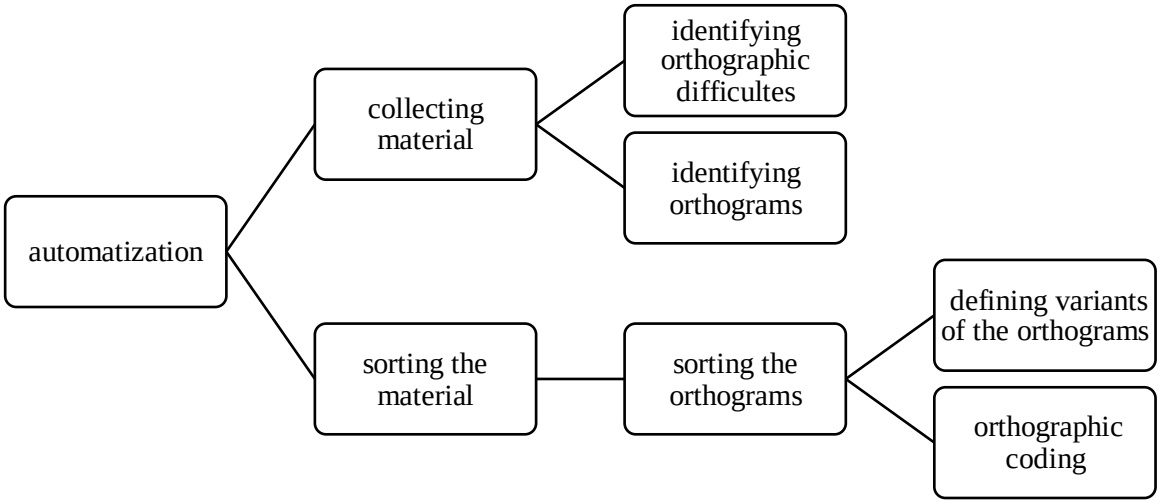
Table 2 – Modeling orthograms based on their distinctive features

Distinctive orthogrammatic feature of the letter A	Distinctive orthogrammatic feature of the letter Ң
- the phoneme *а* when it comes between *ж, ш, й* (*жайлау (summer pasture), шай (tea)*); - writing or omitting the letter *а* in a compound	- when it neighbors the consonants *қ, ғ, г, к: қара[ң] қалу (helplessness), і[ң]ген (female camel), ше[ң]гел (claws), әме[ң]гер (the tradition of marrying the*

Continuation of the table 2

word (*мамаағаш (a stake to which a horse is bound)*); - the letter *а* sounds like *ә* in the second syllable *(кіman (book)*	*spouse of a deceased brother), жы[ң]ғыл (a thorny bush), баста[ң]ғы (the so-called party of Kazakh youngsters), же[ң]ге (sister-in-law), жәрме[ң]ке (fair)*

Distinctive features of orthograms are an intrinsic component of the content of orthographic rules. It is crucial to pay attention to this problem when spelling a word. To spell a word correctly, it is necessary to know, first of all, the orthographic principle, second, the spelling rules, third, orthograms of a language, their types and kinds, and positions of orthograms within words. Such positions are identified through the distinctive features. That is, a certain graphic sign in a certain position is formed through the orthographic rule, on the basis of recognizing an orthogram. That is why this system is the basis of automatizing. In a nutshell, automatizing Kazakh orthography consist of the following steps (Scheme 1):



Scheme 1 – Stages of automatization

The process of automatization is directly linked to solving complicated technical problems. For this reason, this process requires complementing orthographic data and creating a base consisting of different orthological tools. “Orthographic dictionary of the Kazakh language” will be taken as the linguistic base of automatizing Kazakh orthography. Several editions of the given dictionary have been published. Each edition complemented the dictionary and improved its base, and, as is known, variant words with different inscriptions were eventually codified as one. Among these editions, we chose “Orthographic dictionary of the Kazakh language” published in 2013.

The orthographic dictionary is compiled to help spell words correctly. Its purpose is to provide users with the information of how to spell words, including the words with complicated spelling. Also, the information about using words in different contexts and meanings is provided. Orthographic dictionaries are utilized to automatize Kazakh-language texts. They encompass the database of words, the spelling of which is checked within the text. In case the word is spelt incorrectly, the program suggests correction options. This speeds up the process of writing and processing texts, increases the grammatical quality of materials, and helps users avoid mistakes when creating texts which increases the quality of those texts.

The orthographic dictionary plays a crucial role in automatizing the processes which relate to processing Kazakh texts. They help create text editing programs, messengers and other applications which involve text, i.e. programming products. The spelling dictionary provides an opportunity of checking the word inscriptions automatically and suggesting the options of correction. This allows to accelerate the processes of creating and editing texts and increase the quality of the grammatical design

of texts.

Apart from the Orthographic dictionary, such issues as creating the database of Kazakh orthograms and identifying the number of Kazakh orthograms must be taken into consideration when dealing with the linguistic base of automatizing. The reason is that the artificial intelligence memory requires not only the database of orthograms, but also their modeled system. The examples are provided above and below (Table 3).

Table 3 – Base of the orthograms

Orthographic norm	*Variants of the orthogram*	*Explanation of the orthogram*
абайтану	абай тану	The words that contain “tanu” (cognition) are written together. These words, uniting with their initial component, signify a branch of science or a discipline, serving as semi-words.
абжылан	әбжылан, әб жылан, аб жылан	The names of snake types are written together. In orthoepy, the word “абжылан” is pronounced as “әбжылан”
абжыланға	абжылаңға	When the sound [н] combines with the sounds [г], [ғ] in the middle of a word, it is pronounced as [ң] but written as н.
абырой	аброй	In the words that start with [л], [р], [п], [м] and contain an open vowel, the closed vowels [ы] and [і] are not omitted
абысын	абсын	In the flow of oral speech, the letter ы is usually pronounced quickly. In the written form, it is not omitted as it is significant.
абысынға	абысыңға	When the sound [н] combines with the sounds [г], [ғ] in the middle of a word or at the end of words, it is pronounced as [ң] but written as н.
авансахна	аван сахна, аван сахына, авансахына, ауансахна, авансақына, авансақна	In practice, there are certain peculiarities of writing the letters қ and х. With relation to the traditional norms of orthography, some of such words are spelled with қ (рақмет), and the rest with х (сахна).
автобекет	авто бекет, аутобекет, ауто бекет	The traditional orthographic norm prescribes that the letter в in such prefixes as авто- and евро- is preserved and written together with the main word.

The orthographic database is a specific dictionary that encompasses the information about the correct spelling of Kazakh words. It is utilized for automatized checking of the spelling in texts and program products. The orthographic base encompasses the rules of spelling, considering all the words in a text, all the peculiarities and non-standard usages. The orthographic base is peculiar for its identification of incorrect orthograms. That is, the orthographic base includes the list of words which considers their correct spelling and all the possible variants.

The database allows users to identify spelling mistakes quickly and without distortion, providing opportunities for correction options. The orthographic base is a significant instrument of increasing the quality of working with texts and speeding the process of operating texts.

The database is constructed via analyzing numerous Kazakh texts. The texts are sourced from books, journals, newspapers, internet websites, etc. During the construction of the base, it is necessary to consider the issues of automatic analysis of the text, identification of correct spelling and all the possible variants of spelling. The base must contain information about grammatical rules.

Creating an orthographic database can be done manually or with the help of special software. In the first case, linguists analyze texts and create a list of words, taking into account all possible variants. In the second case, programs that automatically analyze texts and create an orthographic database based on the received data are used.

The orthographic database is constantly updated and supplemented with new words and spelling rules. This is necessary for it to work correctly with new texts and to take into account changes in the rules of the Kazakh language.

Conclusion

In the automation of Kazakh orthography, several linguistic aspects are considered together. The first is an inventory of phonetic (composition of vowels and consonants in the Kazakh language, system of pronunciation, hearing, formation and distinguishing features of each sound from another sound) and phonological knowledge (identifying the most basic phoneme that distinguishes the meaning, distinguishing the main and secondary tones (variant/variation) of sounds) necessary to define phonetic-phonological code types. Second, by sorting words that are difficult to write, it is necessary to determine the differential signs of orthograms in order to determine the types of orthographic code (finding types of orthograms from written words). If these two main tasks are fulfilled, the automation of Kazakh orthography will be on its way.

The fields of phonetics, morphology, vocabulary and syntax are mastered in the practice of speaking, and the rule of correct writing from a normative point of view requires the conscious use of special rules. If it is necessary to pay individual attention to language units for the formation of speaking skills, then writing is based on how to use each language unit and how they are reflected in writing. Therefore, for artificial intelligence, orthological tools that show the correct spelling of each linguistic unit (Orthographic dictionary of the Kazakh language, 2013) are also provided. This can be the material needed to determine the correct spelling of a word in an automatic system, to recognize the meaning of a word.

The article was written within the framework of the program of the Science Committee MSHE RK "Automatic recognition of Kazakh text: development of linguistic modules and IT solutions" (BR18574183)

References

Algazina N. (1987) Formirovanie orfograficheskih navykov. – M., 1987. – 158 s. [Algazina N. (1987) Formation of orthographic skills. – M., 1987. – 158 p.] (in Russian)

Arhangel'skij S.I. (1980) Uchebnyj process v vysshej shkole i ego zakonomernye osnovy i metody. – M.: Vysshaja shkola, 1980. – 368 s. [Arkhangelskiy S.I. (1980) Study process in higher education and its logical bases and methods. – M.: Vysshaya shkola, 1980. – 368 p.] (in Russian)

Bazarbaeva Z. (2008) Qazaq tili: intonologija, fonologija. – Almaty, 2008. – 284 b. [Bazarbayeva Z. (2008) The Kazakh language: morphology, intonology. – Almaty, 2008. – 284 p.] (in Kazakh)

Chomskij N. (2000) Logicheskie osnovy lingvisticheskoj teorii. – Birobidzhan: IP «Trivium», 2000. – 146 s. [Chomsky N. (2000) Logical bases of linguistic theory. – Birobidzhan: IP «Trivium», 2000. – 146 p.] (in Russian)

Ermilova V., Ermilova D. Ju. (2001) Uchebnoe posobie dlja uchrezhdenij srednego profobrazovanija. – M., 2001. – 184 s. [Yermilova V., Yermilova D. Yu. (2001) A manual for secondary education institutions. – M., 2001. – 184 p.] (in Russian)

Lakoff G. (1996) Kognitivnoe modelirovanie. – Jazyk i intellekt. – M., 1996. – S. 143-184. [Lakoff G. (1996) Cognitive modeling. – Language and Intelligence. – M., 1996. – pp. 143-184]. (in Russian)

Martin K. (2000) O modelirovanii lingvistiki – M., 2000. – 258 s. [Martin K. (2000) On modeling linguistics. – M., 2000. – 258 p.] (in Russian)

Uali N. (1993) Qazaq grafikasy men orfografijasynyng fonologijalyq negizderi: filol. gyl. kand. ... avtoref. – Almaty, 1993. – 162 b. [Uali N. (1993) Phonological basics of Kazakh graphics and orthography: philol. science. kand... abstract. – Almaty, 1993. – 162 p.] (in Kazakh)

Vartofskij M. (1988) Modeli. Reprezentacija i nauchnoe ponimanie. – M., 1988. – 510 s.

[Vartofsky M. (1988) Models. Representation and scientific understanding. – M., 1988. – 510 p.] (in Russian)

Zhubanov A. (2012) Qazaq til bilimi: qoldanbaly lingvistika. – Almaty, 2012. – 696 b. [Zhubanov A. (2012) Kazakh linguistics: applied linguistics. – Almaty, 2012. – 696 p.] (in Kazakh)

Әдебиеттер

Алгазина Н.Н. (1987) Формирование орфографических навыков. – М., 1987. – 158 с.

Архангельский С.И. (1980) Учебный процесс в высшей школе и его закономерные основы и методы. – М.: Высшая школа, 1980. – 368 с.

Базарбаева З. (2008) Қазақ тілі: интонология, фонология. – Алматы, 2008. – 284 с.

Вартофский М. (1988) Модели. Репрезентация и научное понимание. – М., 1988. – 510 с.

Ермилова В.В., Ермилова Д.Ю. (2001) Учебное пособие для учреждений среднего профобразования. – М., 2001. – 184 с.

Жұбанов А. (2012) Қазақ тіл білімі: қолданбалы лингвистика. – Алматы, 2012. – 696 б.

Лакофф Дж. (1996) Когнитивное моделирование // Язык и интеллект. – М., 1996. – С. 143-184.

Мартин К. (2000) О моделировании лингвистики. – М., 2000. – 258 с.

Уәли Н. (1993) Қазақ графикасы мен орфографиясының фонологиялық негіздері: филол. ғыл. канд. ... автореф. – Алматы, 1993. –162 б.

Хомский Н. (2000) Логические основы лингвистической теории. – Биробиджан: ИП «Тривиум», 2000. – 146 с.

The article has been received: 31.10.2023

