Cuad. Invest. Filol., 59 (2026), 493-519. http://doi.org/10.18172/cif.6716

ISSUES OF HOMONYMS/HOMOFORMS AUTOMATIC RECOGNITION IN APPLIED LINGUISTICS

Akerke Meirbekova
A. Baitursynov Institute of Linguistics, Republic of Kazakhstan
meirbe.erke@gmail.com

Anar Fazylzhanova
A. Baitursynov Institute of Linguistics, Republic of Kazakhstan
fazylz.anar@outlook.com

Aiman Zhanabekova
A. Baitursynov Institute of Linguistics, Republic of Kazakhstan
aiman_zhan@hotmail.com

Aigul Amirbekova
A. Baitursynov Institute of Linguistics, Republic of Kazakhstan
amirbekova.ai@outlook.com

Gulnara Talgatkyzy
A. Baitursynov Institute of Linguistics, Republic of Kazakhstan
talgatkyzy-gul@hotmail.com

ABSTRACT: The phenomenon of homonymy creates additional difficulties for automatic text recognition processes, requiring the use of more complex algorithms and processing methods. The research aims to address the issues of automatic homonym recognition in Kazakh, Russian, English, Turkish and Tatar. The study identified the main problems that arise in the automatic detection and identification of homonyms/homoforms, in particular, the lack of a clear sentence structure (Russian, Turkish), significant morphological diversity of the language (the presence of a large number of grammatical categories, high level of affixation), contextual ambiguity, insufficient development of theoretical issues of homonymy study. The most effective methods of distinguishing homonyms in national corpora of Russian (Markov model with maximum entropy), English (Embeddings from Language Model), Turkish (hybrid method) and Tatar (method of removing homonyms from a corpus marked

with homonyms) were considered. The present study is more aimed at solving the problem from the point of view of analysing morphology and syntax and requires a more detailed study in semantic and contextual aspects.

KEYWORDS: language corpus, morphological analyser, algorithm, homographic pairs, formal features.

CUESTIONES DE RECONOCIMIENTO AUTOMÁTICO DE HOMÓNIMOS/HOMOFORMAS EN LINGÜÍSTICA APLICADA

RESUMEN: El fenómeno de la homonimia crea dificultades adicionales para los procesos de reconocimiento automático de texto, lo que requiere el uso de algoritmos y métodos de procesamiento más complejos. La investigación busca abordar los problemas del reconocimiento automático de homónimos en kazajo, ruso, inglés, turco y tártaro. El estudio identificó los principales problemas que surgen en la detección e identificación automática de homónimos/homoformas, en particular, la falta de una estructura oracional clara (ruso, turco), la significativa diversidad morfológica de la lengua (presencia de un gran número de categorías gramaticales, alto nivel de afijación), la ambigüedad contextual y el insuficiente desarrollo de las cuestiones teóricas del estudio de la homonimia. Se consideraron los métodos más eficaces para distinguir homónimos en corpus nacionales de ruso (modelo de Markov con máxima entropía), inglés (incrustaciones del modelo lingüístico), turco (método híbrido) y tártaro (método de eliminación de homónimos de un corpus marcado con homónimos). El presente estudio se centra en la solución del problema desde la perspectiva del análisis morfológico y sintáctico, y requiere un estudio más detallado de los aspectos semánticos y contextuales.

PALABRAS CLAVE: corpus lingüístico, analizador morfológico, algoritmo, pares homográficos, rasgos formales.

PROBLÉMATIQUE DE LA RECONNAISSANCE AUTOMATIQUE DES HOMONYMES/HOMOFORMES EN LINGUISTIQUE APPLIQUÉE

RESUMÉ : Le phénomène d'homonymie crée des difficultés supplémentaires pour les processus de reconnaissance automatique de texte, nécessitant l'utilisation d'algorithmes et de méthodes de traitement plus complexes. La recherche vise à aborder les problèmes de reconnaissance automatique des homonymes en kazakh, russe, anglais, turc et tatar. L'étude a identifié les principaux problèmes qui se posent dans la détection et l'identification automatiques des homonymes/homoformes, en particulier, l'absence d'une structure de phrase claire (russe, turc), la diversité

morphologique importante de la langue (présence d'un grand nombre de catégories grammaticales, niveau élevé d'affixation), l'ambiguïté contextuelle, le développement insuffisant des questions théoriques de l'étude de l'homonymie. Les méthodes les plus efficaces de distinction des homonymes dans les corpus nationaux de russe (modèle de Markov avec entropie maximale), d'anglais (Embeddings from Language Model), de turc (méthode hybride) et de tatar (méthode de suppression des homonymes d'un corpus marqué par des homonymes) ont été considérées. La présente étude vise davantage à résoudre le problème du point de vue de l'analyse de la morphologie et de la syntaxe et nécessite une étude plus détaillée des aspects sémantiques et contextuels.

MOTS CLÉS : corpus linguistique, analyseur morphologique, algorithme, paires homographiques, caractéristiques formelles.

Received:17/04/2025. Accepted:18/07/2025

1. Introduction

In the context of globalisation, the development of political, social, economic and other relations, the number of information flows is growing rapidly. Since their dissemination in society is carried out in natural language, the scope of linguistics is expanding every day. In this connection, since the last decade of the twentieth century, the issue of creating language corpora capable of automatic natural language processing has become relevant in many world languages. One of the central problems of applied linguistics in this regard is the emergence of ambiguity/multiple meanings in the process of automatic text recognition, which is expressed in the need to develop an algorithm capable of automated analysis of homonyms/homoforms depending on the context.

A homonym denotes words that possess the same spelling or sound yet convey distinct meanings. Homonyms are classified into two categories. The first category is homographs, which are words that share the same spelling yet possess distinct meanings (e.g., lead [to guide] and lead [the metal]). The second category is homophones – a term derived from the Greek 'homo-' (same) and 'phone' (sound) – which refers to words that sound the same but differ in meaning and spelling (e.g., 'to', 'too', and 'two'). A homoform is a word form that possesses a morphological structure common with other words, varying in meaning based on context or grammatical attributes.

The degree of research into the creation of language corpora, as well as the algorithms for automatic recognition of homonyms/homoforms that underlie them, varies from country to country. Thus, addressing the socio-cultural and linguistic situation of Kazakhstan (the "trinity" of Kazakh, Russian and English), as well as the fact that Kazakh, Turkish and Tatar belong to the same language group, it seems appropriate to study this issue based on intercultural experience, designating these languages as the objects of this study (Shyngyssova and Skripnikova, 2018: 41). Woldeyohannis and Meshesha (2022) point out that many automatic natural language processing tasks are based on morphological markup of texts, which is also necessary for machine translation systems. Elov et al. (2023) note that the morphological record is used in semantic annotation and is an important part of parsing. Reproducing large morphologically marked texts manually is a complicated, time-consuming and labour-intensive process, so ready-made morphological analysers (such as Myself, Pymorphy) are used for marking texts. However, when performing word analysis, such programs offer all theoretically possible variants, i.e. almost all word forms in the text have several variants of analysis, which makes it necessary to get rid of those variants that do not fit the given context. In this regard, researchers of applied and computer linguistics pay great attention to the search for an algorithm aimed at determining the only semantically and contextually correct variant in such cases.

Liu et al. (2018) prove that a significant number of errors can occur in the process of neural machine translation, mostly related to the recognition of homonyms and polysemy. Special automated processing of homonymous words can significantly improve the quality of translation and reduce the time required to select the correct variant manually (Kondratenko, 2014). When building a high-quality algorithm for recognising ambiguity in a text, an important aspect is its training based on texts (examples) that will form a certain "knowledge base" used in the future for automatic detection of homonyms/homoforms. van den Beukel and Aroyo (2018) used free and publicly available electronic sources of information WordNet and Wikipedia to develop the algorithm. The researchers' findings have several positive results, but since Wikipedia is not a reliable and trustworthy source of information, WordNet as the main dictionary database does not contain information on homonymy, and the processes of manual correction of results are not sufficiently minimised, further study is required.

The problem of the lack of a full-fledged, high-quality source that would provide a basis for identifying homonyms is also noted by Rice et al. (2019). Having identified this problem, the authors manually annotated more than 500 homonyms taken from the comprehensive Wordsmyth dictionary. The study of the researchers is of high practical importance, but it should be noted that the development of a special

algorithm capable of performing homonym recognition automatically would significantly increase the number of homonyms removed in a shorter time. Among the researchers who have studied the issues of homonym identification in Turkic languages are Khakimov et al. (2016), developed a tool based on context analysis rules to recognise and remove grammatical ambiguity in the Tatar language, and Sak et al. (2007), created a perceptron-based algorithm for similar purposes in Turkish. Such tools are effective in almost half of the cases and can serve as a basis for developing similar or improved models in related languages. The proposed models also require further refinement, considering the creation of rules for all variants of homonyms.

Since 2009, the Baitursynuly Institute of Linguistics in Kazakhstan has been compiling a language corpus of Kazakh texts. As part of this project, various styles of the Kazakh language were collected, and a programme was created that automatically divides words in texts into root words and word modifiers and assigns morphological signs to them (Zhubanov and Zhanabekova, 2016). When analysing a text, this programme encounters certain problems with distinguishing homonyms and parts of speech. At the same time, the Kazakh language lacks a sufficient theoretical basis that would allow for the development of rules for automatic homonym recognition. In addition, the experience of using data and processing methods from other languages to improve work with the Kazakh language remains unexplored.

The study aims to create rules that will improve the accuracy of homonym detection in the general corpus of the Kazakh language to include them in the automatic recognition algorithm. Following the set goal, the following research objectives have been identified: to analyse homonyms in Russian, Kazakh, English, Turkish and Tatar by their features in a sentence; to identify the peculiarities of the composition of each of the languages under study and their influence on the complexity of homonym recognition; to conduct an experiment to identify homonyms and parts of speech in the Kazakh corpus; to make a classification of homophones in the Kazakh language; to develop rules for recognising homophones in the Kazakh language to include them in an automatic identification algorithm.

2. Materials and Methods

The study of problems and algorithms for recognising homonyms/homoforms in the Kazakh language was carried out based on a language corpus compiled by specialists of the A. Baitursynuly Institute of Linguistics. The corpus is based on data (texts) from reliable resources on the Internet, electronic libraries and other sources. Eight texts, two for each style (scientific, official business, journalistic, and fiction)

were selected as research materials from similar sources that meet the requirements of reliability.

To achieve the research goal, an experiment was conducted based on the institute, during which a database of materials (a specified number of texts) was processed using a morphological parser developed by the institute's computer linguistics specialists. While processing, the words and homonyms identified by the morphological analyser, as well as the word classes in each text, were assigned conventional signs. As a result, the study found that the part of speech of many words and homonyms/homoforms remained unidentified. Since the text of the analyser marks such words as "unknown", the morphological parser generates a list of such words in an automated manner. The words that were not recognised by the morphological parser were then sorted and assigned to the correct part of speech manually. Since the relation of homonyms/homoforms to the word class was manually extracted from the context, this automatically extracted their quantitative indicators.

The obtained quantitative indicators became the basis for the formation of a classification of homoforms of the Kazakh language. The frequency of homonyms in the Frequency Dictionary of the Kazakh Language was analysed to assess the accuracy of recognizing homonyms and their parts of speech (Temirgalieva, 2016). Based on this data, a classification of 402 of the most common homophones was compiled. In the current study, these were categorised into 20 types according to their grammatical characteristics, namely: verb-noun; verb-noun-adjective; noun-adjective; verb-adjective; verb-pronoun; noun-pronoun; numeral-pronoun; conjunction-pronoun; adverb-noun; verb-auxiliary noun; noun-numeral; verb-adverb; adverb-numeral; verb-conjunction; noun-conjunction; adjective-conjunction; noun-auxiliary noun; adjective-pronoun; verb-modal word; noun-noun.

The classification is based on the principles of certainty and mutual exclusivity. Numerical indicators reflecting the degree of distribution of this or that type of homoform in the Kazakh language are also indicated. The diagram tool shows how each type relates to each other in terms of frequency of use, and which types are more commonly used in the Kazakh language. To determine the model for recognising homonyms in Kazakh, Russian, English, Turkish and Tatar, the article also analyses the homonyms/homoforms of each language by features, including the analysis of syntagmatic, syntactic and lexical relations of the units.

Based on the results of the study of the applicability of features for inclusion in automatic recognition algorithms in each language, comparative and descriptive methods were also used to compare the composition of the analysed languages and identify key features that reflect the main similarities and differences between them. Various methods were utilised to differentiate homonyms and homoforms within the

examined corpora. The principal methodologies encompassed syntagmatic analysis, which investigates the interrelations of words according to their placement and role within sentences, taking into account word order, semantic valence, and syntactic structure. Contextual analysis was employed, concentrating on adjacent words and phrases to ascertain the accurate meaning of homonyms, examining the relationships among adjectives, verbs, and nouns inside a sentence to establish if a word should be regarded as a noun or verb. Morphological analysis was used to investigate the structure of words, including their affixes, aiding in the differentiation of homonyms that share the same basic form but belong to distinct grammatical categories, such as nouns and verbs. Furthermore, statistical techniques were employed, encompassing the Markov model with maximum entropy for Russian, the Embeddings from Language Model (ELMo) for English, and a hybrid approach integrating contextual rules and statistical analysis for Turkish and Tatar corpora. These models were tested and refined based on corpus-specific features and contextual ambiguities.

The selection of statistical methods was determined by their relevance to the languages under examination and the characteristics of homonymy in each language. The maximum entropy Markov model was selected for Russian due to its effective management of the language's morphological complexity. The ELMo model was utilised for English because of its effectiveness in contextual word embedding for disambiguating homonyms in syntactically simple languages with minimal morphological variability. The hybrid method was favoured for Turkish and Tatar, as it amalgamates statistical and contextual criteria and proved particularly efficacious in agglutinative languages characterised by complex morphological structures. Despite their prevalent application in linguistic classification tasks, decision trees were not evaluated in this study. Decision trees, although efficient in certain settings, exhibit limited adaptability to the extensive morphological variety present in the corpus of languages such as Kazakh, Turkish, and Tatar, where various affixations and word ordering might affect homonym distinction. Additionally, decision trees might not capture the complex relationships between words in context as well as the chosen models that use probabilistic and contextual analysis. Thus, we selected models that could more effectively address the linguistic complexities inherent to the languages being examined.

3. Results

The creation of an algorithm for recognising homonyms in the Kazakh language corpus may allow the development of an accurate and reliable model for automatic homonym recognition not only in Kazakh but also in other languages, primarily those belonging to the Turkic group. Even though Kazakh, English, Russian, Turkish and

Tatar belong to three different language groups, they are united by the presence of many homonyms/homoforms in their lexical composition. The effectiveness of the algorithm can be assessed by improving the quality of natural language processing (increasing the number of automatically reliably recognised homophones) and the possibility of using it to analyse texts in other languages. In addition, efficiency is determined by the accuracy of the underlying rules. The phenomenon of homonymy creates significant difficulties in annotating a corpus of texts since a computer language understands only formal structures, and semantic differences between language units (word forms) are only possible with human intervention. This requires manually developed and detailed rules.

Homonym recognition algorithms developed for automated morphological analysis programmes are based on the syntagmatic relations of words in a language (Fashwan and Alansary, 2021: 207). This means that, before determining which class of words the homonyms belong to, rules are made based on the location (distribution) of the word. Such rules are included in the algorithm. This is called the deterministic form of homonym distinction. Since the basis of the morphological record is the relation of words to the word class, the purpose of distinguishing homonyms is to determine the relation of the homonym word to the word class. The following signs of homonymy in a sentence are taken as a basis: word order in the sentence; semantic relations in the phrase; word-formation affixes attached to the word; punctuation marks at the end or before the word (Zhubanov and Zhanabekova, 2016). To test the applicability of the homonym recognition features, a case study for each of the languages under investigation was conducted.

The Kazakh homonym word “ат” [ʌt] can be a noun in the sense of “name”, a noun in the sense of “horse” and a verb in the sense of “shoot”. In the Kazakh language, a noun as a part of speech usually comes at the beginning or middle of a sentence, and a verb at the end of a sentence. Thus, according to the first feature, it is possible to state that when the word “ат” is placed at the beginning or in the middle of a sentence, it is more likely to be a noun, for example, “ат тыныш тұрды” (the horse stood calmly). If the word “ат” comes at the end of a sentence, it is a verb, for example, “ортадагы нысаманы ат” (to shoot the target in the middle). According to the second feature, words are arranged by semantic valence. For instance, “торы ат” means a breeding horse, “әдемі ат” means a beautiful horse, and the word “ат” can be combined with adjectives (торы, әдемі) only when it is a noun. The verb is more often attached to an adverb, for example, тез ат, дәлдеп ат – to shoot quickly, to shoot accurately. The third feature assesses the presence of affixes in a word in each context. If the word “ат” is attached to suffixes that can only be attached to a noun, it is a noun, and if it is attached to suffixes that can only be attached to a verb, it is a verb.

According to the fourth feature, if the word “aт” is a verb, in most cases it is followed by punctuation marks (commas, periods, exclamation marks, other signs), which complete the thought or sentence.

The English homonym “box” [bɒks] can be a noun meaning "container" or "box," a verb meaning "to box," and a verb meaning "to pack”. According to the first feature, if the word “box” is in the beginning or middle of a sentence, it is a noun. If the word “box” appears at the end of a sentence, it can be either a noun or a verb. The order of words in a sentence in English is distinguished by a clear structure, but despite this, the variety of meanings of homonymous words makes it difficult to recognise them (Hromko et al., 2023: 29). This problem can be solved with the help of part-of-speech tagging and simple methods. In this case, algorithms based on statistical methods, such as the Markov model and the conditional Markov model, are used, which show an accuracy of at least 96% for English (Cing and Soe, 2020: 2025). In addition, support vector machines and decision trees from relevant statistical methods can be used. For instance, when SVMs were tested on the text of news articles from The Wall Street Journal corpus, the accuracy of homonym ambiguity removal was 97.2%, which is quite high (Tang, 2006: 38). According to the second feature, the word “box” can be combined with various adjectives, for example, “big box”, or “small box”, which indicates its noun nature. For the word “box” as a verb, the most frequent combinations have not been identified, but the method of exclusion can be used, which will determine that if the combination contains an adjective, the verb variant is not considered significant.

Researchers of the contextual method for distinguishing homonymy in English note that the most important aspect of the homonym’s location is the qualitative aspect of the word (Kanerva et al., 2021: 565). The most important context in terms of semantic load is the one in which the analysed polysemous/dominant word is surrounded by one word each on the left and right. If at least one of the words next to the analysed word is an auxiliary word – a “particle” – the number of words on both sides of the context should be increased to two (Wilson and Marantz, 2022: 146). In combination with the features being analysed, this variant of the contextual method can be additionally specified in the recognition algorithm. According to the third feature, if the word “box” is prefixed with suffixes that are typical for a noun (for example, “-es” in the plural form – “boxes”), it is a noun. If affixes characteristic of a verb is added (for example, “-ing” in the participle form of “boxing”), it is a verb. The fourth feature is that if the homonym “box” comes at the end of a sentence and is followed by a period or other sentence ending mark, it may indicate that the word is a verb.

The homonym “печь” [п’эч’] in Russian can be a noun, denoting a structure, made of stone, brick or metal, adapted for heating a room for cooking, and an imperfect verb, used to denote the process of cooking food under the influence of very high temperature (heat). Since the word order in Russian is not stable according to the first feature, it is quite difficult to unambiguously designate which class of words the homonym “печь” belongs to in a given context. As a noun, the word can be used at the beginning of a sentence (e.g. “печь занимала очень много места” – “the oven took up a lot of space”), in the middle (“мы собирались чистить печь только этим средством” – “we were only going to clean the oven with this stuff.”), and at the end (“она всегда делает выпечку в печи” – “she always makes baked goods in the oven”). Similarly, as a verb, the word can be placed at the beginning of a sentence (“печь пироги – ее любимое занятие” – “baking pies is her favourite pastime”), in the middle (“я хотела печь пироги сегодня днем” – “I wanted to bake pies today afternoon”) and at the end (“она любит печь” – “she loves to bake”).

Given this feature of the Russian language, as well as the fact that morphological ambiguity in Russian is not limited to interclass homonyms, but also includes many grammatical features, it is advisable to use algorithms based on the hidden Markov model with context detection in the whole sentence to solve functional homonymy. According to the second feature, if the word “печь” is combined with a subsequent noun, it is a verb (“печь пиццу” – “bake pizza”, “печь хлеб” – “bake bread”). According to the third feature, if affixes are added to the word “печь”, it is a verb (for example, the prefix “за-” – “запечь”). The absence of theoretically affixes may indicate that it is a noun. Due to the variability of word order in Russian sentence construction, information about punctuation marks located near the analysed word does not carry the intended semantic load. Thus, it is impossible to draw unambiguous conclusions about the fourth feature, as well as the first, without additional text processing.

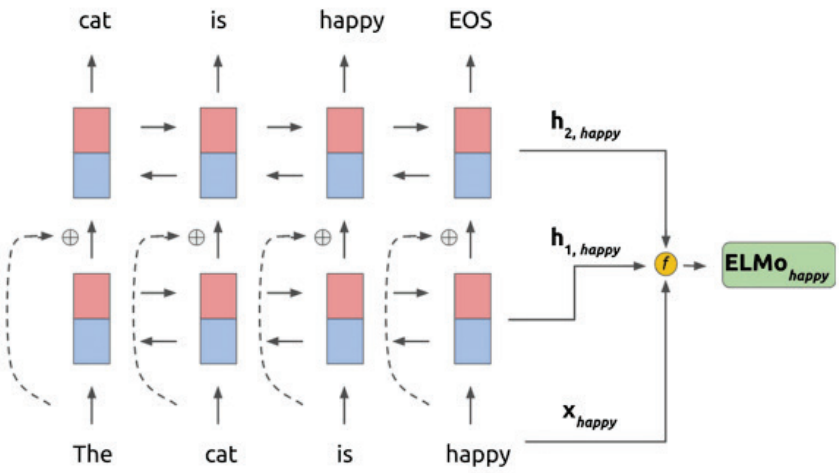
The Turkish word “yaz” can be a noun in the sense of “summer” and an imperative verb in the sense of “write”. The first sign is that if the word “yaz” appears at the beginning or in the middle of a sentence, it is most likely a noun, for example, “yaz çok sıcaktı” – “the summer was very hot”. If it comes at the end of a sentence, it is more likely to be a verb, for example, “bu metni yaz” – “write this text”. According to the second feature, the word “yaz” is a verb if it follows a noun (“bütün bu kelimeleri kağıda yaz” – “write all these words on paper”). According to the third feature, if the word “yaz” is prefixed with suffixes that are typical for verbs (for example, “-mak” – “yasmak”, “to write”), it is a verb. According to the fourth feature, if the word “yaz” is not followed by any punctuation marks or is followed by a dash,

it is a noun; if it is followed by a colon or punctuation marks indicating the end of a sentence, it is a verb.

The Tatar homonym “бит” can be a noun meaning “face, cheek”, a noun meaning “page, leaf”, and a particle “because”. According to the first feature, if the word “бит” is at the beginning or middle of a sentence, it is a noun (for example, “бер бит кәгазь бир әле” – “give me a piece of paper”), if it is at the end, it is most likely a particle (“мин сезгә әйттем бит!” – “I told you so!”). In some cases, before independent words, the particle “бит” can also appear in the middle of a sentence (“мин бит сезгә әйттем!”). To improve the accuracy and quality of recognition, separate rules can be set for such variants. According to the second feature, if the homonym “бит” is used with adjectives (“бит күнелгән” – “tired face”) or pronouns (“мин бит” – “my page”) following it, it is most likely a noun; if the word “бит” is preceded by a verb or adverb, it is most likely a particle (“ул турыда син миңа язмадың бит” – “you did not write to me about it”; the verb is язмадың). According to the third feature, if the word “бит” is attached to suffixes that are typical for nouns, it is a noun (for example, the suffix (case affix) -кә – “биткә чибәр”, “handsome in person”). According to the fourth feature, if the word “бит” is followed by commas or other punctuation marks that complete the thought, it is a particle.

Based on the analysis, the following conclusions were drawn. In Kazakh, English and Tatar, the word order in a sentence is strict, while in Turkish and Russian there are certain rules for sentence construction, but in general, the order is freer (especially in Russian) (Papa et al., 2025; Bilous and Barladiuha, 2020). Since word position changes can also affect changes in semantics and context, and since there are many more theoretically variants to analyse, it is easier to develop rules based on the above features for languages with a clear sentence structure in this aspect. The Kazakh and Turkish languages are characterised by a significant number of word forms, while the Tatar language is also relatively flexible and can change the shape of words by adding various morphemes. As words with different semantics and grammatical characteristics may have similar forms or affixes (in particular, suffixes), the process of homonym recognition in such languages becomes more complex and requires a more detailed study of the properties of words and their frequency of occurrence through the analysis of a large corpus of texts. English is characterised by a low level of morphological diversity, while Kazakh, Turkish, Tatar and Russian are highly diverse (Chaika, 2023; Hoff and Barboza, 2025). Russian is the most difficult language to distinguish between homonyms/homoforms because the same word form can have several grammatical interpretations (the influence of many grammatical categories, such as gender, case, and other morphological features).

The model of homonym identification in Kazakh, Turkish and Tatar is similar: if X is the end of S, then XV (X is a homonym, S is a sentence, V is a verb). The phenomenon of lexical homonymy significantly complicates the process of forming such a model. In this case, additional algorithms are required to consider semantic and contextual features in each of the analysed languages, as well as deep learning methods (Serdiuk, 2023: 172). One of the most effective approaches to solving the problem of contextual homonym recognition in English is ELMo, which is based on a trained two-level language model and a vector of word embeddings in the text (Peters et al., 2018: 2231). The algorithm operates by calculating a vector for each of the words in the analysed text, i.e., any of the words necessarily has its vector defined by the context, and then using the capabilities of the Long Short Term Memory architecture and k-values are grouped into pairs of homonyms (Figure 1).



Note: x – verbal representation; h_1 and h_2 – bidirectional representations of the inner layer.

Figure 1. How ELMo operates. Source: compiled by the authors based on Y. Lee (2021).

Thus, since the algorithm can find the appropriate context for each word, the homonyms will also be recognised in the vector space and thus will be correctly identified in the given context. To apply this approach to the Kazakh linguistic corpus, it is necessary, firstly, to create a trained model that will form the basis of the algorithm corresponding to the Kazakh language, and secondly, to create additional conditions (rules, models) that would consider the grammatical features (affixation) of the Kazakh language.

The Markov model of maximum entropy (MEMM) is also used as an algorithm for removing morphological ambiguity. The algorithm is a two-layer model (latent indicator layer and observation layer) aimed at recognising homonyms by building and evaluating the dependence of each subsequent state on the preceding one (Duro and Kondratenko, 2015; Bashmanivskiy, 2016). The indicators of the hidden layer depend on the observed ones. Model training consists of recovering the conditional distribution by applying the MEMM principle to some features calculated from the corpus (Alwateer et al., 2023). The accuracy of ambiguity removal is at least 90%. This value of the model's applicability to the Russian language is quite high. Due to the similarity of the morphological structure of the Russian and Kazakh languages, as well as the high morphological complexity, MEMM algorithms can be used to compile a corpus of the Kazakh language. In any case, when training the model, it is necessary to consider as many features of the Kazakh language as possible, as well as the fact that the accuracy of homonym detection using MEMM deteriorates when examined in increased size. In addition, the algorithms often make mistakes in distinguishing proper names, pronouns, Roman numerals, initials and abbreviations, and in distinguishing certain participle forms (in Russian, the distinction between the nominative and accusative cases). In the Kazakh language, there should be no difficulty in distinguishing the cases of words. Nevertheless, the quality of homonym recognition using this model, including the above factors, can only be assessed after practical testing.

Proper nouns, pronouns, and other commonly confusing phrases are addressed in the homonym recognition process using a blend of context-based strategies. Proper nouns are typically processed by named entity recognition, which recognises and categorises them as unambiguous according to context, such as differentiating Paris as a geographical location or an individual. Pronouns are clarified by anaphora resolution, wherein the system associates the pronoun with its appropriate antecedent by examining sentence structure and syntactic connections (Borysova, 2023; Strilets, 2021). For other ambiguous terms that can serve various syntactic functions (such as "can" as a verb or noun, or "bank" as a financial institution or a riverbank), part-of-speech tagging and contextual analysis are utilised to ascertain the most probable meaning, frequently augmented by statistical models like hidden Markov models or neural networks trained on extensive corpora. These strategies collectively aid in clarifying ambiguity and ensuring the accurate identification of phrases according to their context.

To resolve morphological ambiguity in the Tatar language, a hybrid method seems to be the most effective (Khakimov et al., 2016: 238). This method is based on contextual rules and statistical and probabilistic methods. This is because, although

the method based on contextual rules shows good accuracy in resolving ambiguity, creating such rules for all types of ambiguity is a very large task that requires careful linguistic analysis. The author proposes to use a generalised contextual method of removing ambiguity, which includes three stages. The first is to compile a complete classification of homoforms. The second is the extraction of the minimum context (the most easily identifiable nondecomposable simple condition in each context) required to eliminate the ambiguity of functional homonyms for each type. The third is the creation of a generalised rule management structure that will ensure maximum accuracy of homoform recognition. By the beginning of the third decade of the twenty-first century, the classification of homophones in the Tatar language had not been made. The method is based on some regularities of context determination, which indicate that the stability of a certain amount of agglutinative syntactic structure allows us to define clear contextual restrictions.

The process of separating homonyms is as follows (Figure 2). To begin with, the functional form of the homonym of the analysed word is determined and a generalised ambiguity elimination rule based on contextual rules is found. Next, the algorithm structure establishes the order in which the rule is applied. Each time a rule is used, it is checked for compliance with the minimum context that eliminates ambiguity in the rule. If the rule is found to be correct, the functional homonym is recognised as corresponding to this homoform structure. If there is another rule, new actions will be performed by moving to the next rule. Unless otherwise specified, the construction type is selected by default. If the default type is not found, the ambiguity is marked as unresolved.

The introductory ones shown in Figure 2 are as follows: S (sentence) – the context in which the homonym occurs; MK11°MK12°…°MK1n – the minimum distinguishing context for a functional homonym of type T1; MK11°MK12°…°MK2n – the minimum context for type T2. Otherwise, a functional homonym is recognised as a TK (Khakimov et al., 2016: 243; Quecedo et al., 2020: 3575). To mark up the Tatar language corpus, a web-based programme based on crowdsourcing (voluntary participation of people) was developed. The processes included in the hybrid method can be used to create a Kazakh language corpus. To ensure the effectiveness of their application, it is advisable to include the classification of homonyms of the Kazakh language in the algorithms of the hybrid method. Since the Kazakh language corpus is still in the process of formation, the use of crowdsourcing (necessarily based on certain rules), including the development of a crowdsourcing programme, could also be a positive experience that would speed up the process of collecting information for training the model with the statistical probabilistic method.

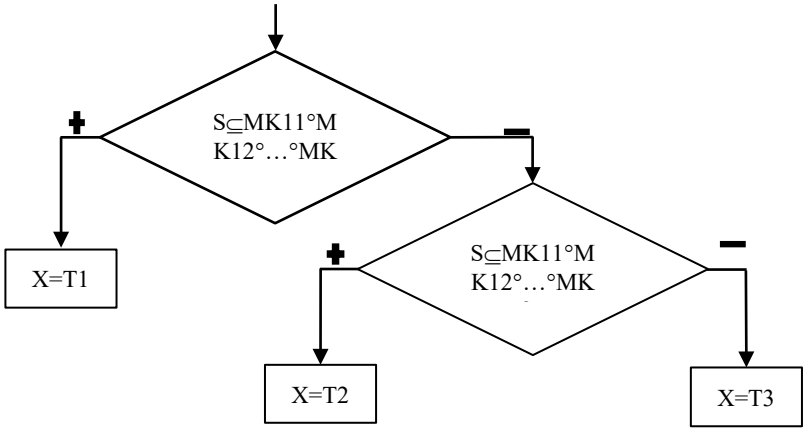


Figure 2. The sequence of the process of eliminating homonym ambiguity. A set of minimal distinguishing contexts for distinguishing homoforms of the type T1, T2, TK (X).

To eliminate morphological ambiguity in Turkish, a method of removing homonyms from a corpus marked with homonyms was developed (Yuret and Türe, 2006: 330). The peculiarity of the method is that each morpheme has its contextual restriction, but not all types of affixed chains. The method is based on a new learning algorithm for the Turkish language that combines decision lists, statistical and rule-based methods, considers each character individually to avoid data fragmentation, and does not require the previous word to be identified and relies on external features. The method is used with the grade point average algorithm to learn the rules generated by a morphological analyser that generates tags consisting of 126 unique characters. For each such character, a subset is found that contains its analysis for each example, after which the examples included in the subset are divided into correctly and incorrectly recognised depending on the accuracy of the analysis of this character. The method is based on the peculiarities of agglutinative languages and has at least 96% morphological ambiguity elimination, which means it is potentially applicable to the Kazakh linguistic corpus.

The application of the experience of homonym/homoform recognition in corpora of other languages can only be appropriate if the specifics of the language are considered, and the problems and gaps in the theoretical aspects of homonymy in the Kazakh language are addressed. In addition to those identified in this article, the Kazakh language also has difficulties in recognising lexical items that form a homographic pair. For such homographical pairs or several homographical words, special rules must also be established. To do this, it is initially necessary to classify

possible homographic pairs based on their specific characteristics, and only the next step is to create rules. To distinguish between homonyms and homographs, a computer program must first of all be guided by the formal features outlined earlier. A morphological analyser is performing better when it is based on as many instructions as possible that do not contradict each other and consider the specifics of the recognition language (Almalki et al., 2025; Kulyk, 2023).

In the current study's homophones' classification, the most common type is the verb-noun type (264 homophones, which is 65.7% of the total number of homophones in the sample). As such, there is a fairly high probability that the analysed homoform will be a coincidence of a verb and a noun in a certain form. Verb-noun-adjective, noun-adjective, verb-adjective, verb-adjective and verb-adverb types are also relatively frequent (Figure 3).

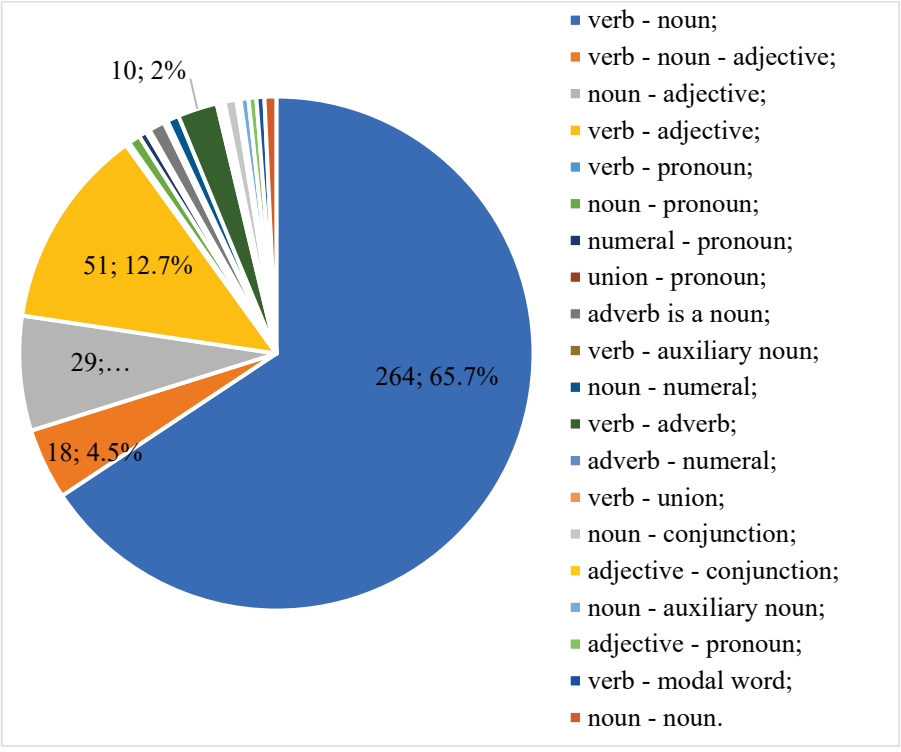


Figure 3. The ratio of each type of homoform (only the most frequent types are presented).

Thus, based on the study of automatic homonym/homoform recognition in the Kazakh language corpus, it is shown that the work on frequency extraction and obtaining statistical data helps to determine which type of homoforms requires more detailed study to further improve the accuracy of extraction in the general corpus. The classification revealed that homophones formed from nouns and verbs account for more than half of the homophones in the Kazakh language. When developing algorithms for recognising homoforms of these types, the following general rules can be specified:

1. If there is a comma or semicolon before and after the homonymous word X, X must be identified as one of the individual members and the homonyms must be distinguished based on the remaining non-unit single members. If singular members are replaced by adjectives, singular numbers and conjunctions, the rule is invalidated.
2. If the homonym X comes before or after the words “мен”, “бен”, “пен”, if the conjunctions “да”, “де”, “та”, “те” come before X, and if X comes after the conjunctions “не...не”, it is necessary to define the homonym as one of the homogeneous parts of a word or speech and to distinguish the homonym based on the other non-homonymous homogeneous parts of speech.
3. It is impossible to add two or more case suffixes one after the other, i.e. if the analyser outputs two case suffixes in a row in one parsing variant when distinguishing a homonym, this parsing variant is automatically considered invalid.
4. Unions cannot be modified; they cannot be formed from other parts of speech employing affixes. These are auxiliary words formed because of the loss of their lexical meaning.
5. Two infinitive verbs in a row are not used, and thus such homonyms, where both words can be a verbal infinitive or a noun, such as “қара шаш”, “бас қала”, “қара ат”, “ақ ат”, are resolved.

The suggested regulations for homonym recognition in Kazakh can be modified and expanded to cover other Turkic languages by considering their linguistic characteristics while preserving the fundamental principles of the current algorithm. Turkic languages, including Turkish, Tatar, and Uzbek, exhibit shared morphological and syntactic features, such as agglutination and vowel harmony, facilitating the adaptation of rules across several languages. The management of affixes, a fundamental characteristic of the Kazakh language, can be applied to other Turkic languages, which also display analogous affixal structures. Each language possesses

distinct characteristics (Brait et al., 2025; Shlapak, 2016). For example, Turkish exhibits a somewhat rigid word order in contrast to languages such as Kazakh and Tatar, necessitating adjustments to the syntactic rules for homonym disambiguation. Furthermore, lexical elements and semantic differences in each language must be considered. In certain Turkic languages, like Tatar, morphological alterations within word forms may occur more frequently, requiring more detailed rules to address these variations. Using language-specific data, like frequency dictionaries or semantic classifications, in the analysis methods would make it easier to identify words that sound the same but have different meanings. By combining the analysis methods that worked well in Kazakh and adjusting them to fit the unique word forms and sentence structures of other Turkic languages, the suggested rules could be changed to better identify homonyms in these languages, making the study more useful and relevant.

4. Discussion

When studying the problems associated with the detection of homonyms in texts, different linguists choose different aspects on which to base their research. Among the most relevant are the contextual and semantic approaches to automatic homonym recognition in a language corpus.

Habibi et al. (2021) based on the work of van den Beukel and Aroyo (2018), considering the work of Rice et al. (2019) use an approach to eliminate morphological ambiguity, which is based on building a map of relations between the semantics of words and analysing the meanings in one coherent component as multiple meanings. Unlike previous studies, A.A. Habibi et al. build the corpus of texts on verified sources, noting that Wikipedia is not such a resource. The study conducted in this article also identifies this problem, so when forming the material base for the experiment aimed at improving recognition algorithms in the Kazakh language corpus, only those texts that meet the requirements of reliability were selected. This is because the presence of semantically, grammatically and syntactically unrelated words in texts can worsen the recognition quality, causing errors in the analyser.

The use of contextual relations to detect word ambiguity was studied, in particular, by Hassan et al. (2009), Shashank et al. (2019). Hassan et al. used algorithms trained on recognising associative rules for all meanings of a word in a context, which allows them to identify the most appropriate meaning in terms of automatic determination of its reliability. U. Shashank et al. developed rules for a multi-meaning word recognition algorithm that includes text processing with punctuation, identifying words to exclude in homonymy research, and identifying keyword indicators in context based on trained data and dictionaries (explanatory,

historical). Several studies in the field of word ambiguity identification, mainly in the field of computational linguistics (Toews and van Holland, 2019; Kim and Kim, 2020; Patil et al., 2020: 1188), focus on improving the accuracy of automatic homonym/multiple word detection by embedding words. This method can also be referred to as a contextual method, since in the hidden layers (in the vector space) the vector of the word under study is analysed concerning other words in the sentence. It is possible to note that many of these methods or models are not adaptive for the automatic recognition of homonymy (Li et al., 2021).

Analysing the above models, it is possible to conclude that they do not pay enough attention to the study of grammar and syntactic relations. Many languages, including Kazakh, are marked by a complex grammatical structure, which requires a deeper study from the point of view of morphology. This article analyses the features of homonyms concerning the class of words in a sentence and establishes that these features are fully applicable to the Kazakh language and can be used as the basis for rules for recognition algorithms. Shashank et al. (2019) point out the expediency of using tokenizers to identify words and punctuation marks in the text. The study presented in this article also shows that considering punctuation marks, including those indicating sentence completion, as well as assessing the relationship of homonymous words to them, allows us to determine their part of speech and, using additionally specified rules, to increase the accuracy of eliminating morphological ambiguity. Casas et al. (2019) conducted a study of word frequency in English, Spanish, and Dutch corpora, and examined its relationship with other features of a homonym word. This approach seems to apply to many languages and also implies obtaining significant information that will become the basis for the formation of new rules and patterns for detecting morphological ambiguity in these languages. The study by Casas et al. was conducted using the value-frequency law and Zipf's law. In the experiment to which this article is devoted, the frequency of homophones in the Kazakh language was studied only using the law of meaning-frequency, which implies the study of the relationship between the frequency of distribution and the tendency of more frequently occurring words to be polysemous. Since the study of the frequency of homonyms based on the Frequency Dictionary of the Kazakh language for general educational purposes was conducted for the first time, the use of only one law seems quite logical. The results obtained first need to be tested for their practical application, and only then further investigated using a larger corpus of analysed texts and additional data (laws, algorithms, approaches).

Kessikbayeva and I. Cicekli (2016) developed a rule-based morphological analyser for the Kazakh language, using tagging previously developed for use in Turkish. The advantages of such an analyser include the inclusion of all

morphological features of the language (including derivational and inflectional morphemes; rules of vowel and consonant harmony). The accuracy of the coverage is estimated by the researchers themselves at 99%. This analyser does not cover technical, borrowed words and proper names. It has a recognition accuracy of 87%. The research base of the researchers compiled four texts. In the experiment described in this article, the volume of the base material was twice as large, which increased the accuracy of the results collected. Comparison of recognition accuracy using the rules developed in this study is only possible after their practical application. The authors also note problems in the study of grammatical categories in the Kazakh language due to variations in different data sources or the complete absence of development of certain aspects (for example, different quantitative values of verb tenses – from three to twelve). Such gaps in the theory of linguistics significantly complicate the processes of automatic homonym recognition. In the course of the study described in this article and the formation of rules for inclusion in the homonym/homoform recognition algorithms, the grammatical category of verb tense was not taken into account.

Many theoretical questions about the phenomenon of homonymy in Turkic languages remain unresolved by the beginning of 2024. Among these issues is the compilation of a classification of homonyms that would comprehensively consider the peculiarities of a particular Turkic language. For example, the Karakalpak language uses a classification of four types of homonyms (lexical, lexico-grammatical, grammatical and mixed) (Kaljanov, 2021: 197). This division is quite general, and each of the defined types requires a separate detailed classification/typology. The solution of theoretical issues is a necessary basis for training algorithms for automatic recognition of ambiguity in text. This article offers a classification of the type of homonyms – homoforms in the Kazakh language. The results of the study do not claim that the developed classification covers all aspects of grammatical relations in the Kazakh language, but it addresses the predominant (most frequent) part of the homophones of this language. Concerning Turkic languages, it is also possible to note that recently there has been a tendency to borrow words and terms from English and Persian (Doszhan, 2016: 24; Kessikbayeva and Cicekli, 2016: 101). The use of such words of foreign origin with violation of their lexical and semantic properties may also cause additional ambiguity in the analysed corpora. The results of the study conducted in this article do not take this factor into account, as well as the recognition of proper names and abbreviations, which may be the basis for further research in this area.

5. Conclusions

The phenomenon of homonymy creates additional difficulties for automatic text recognition processes, requiring the use of more complex algorithms and processing methods. This article studies the degree of development of this topic in Kazakh, Turkish, Tatar, Russian and English languages, identifies the positive experience of each of the analysed language corpora, considers the possibilities of their application in the compilation of the Kazakh language corpus being formed based on the A. Baitursynuly Institute of Linguistics, and develops proposals for its improvement. Following the analysis of the relationship of a homonym to a class of words according to its features in a sentence, the peculiarities of the composition of each language and their influence on the complexity of homonym detection are studied. The study determined that Kazakh, Tatar and English are characterised by a clear sentence structure, the latter also having a low level of morphological diversity, which simplifies the process of identifying homonyms/homoforms in the corpus of texts. At the same time, Kazakh, Tatar and Turkish languages are characterised by the possibility of adding many morphemes to words, while Russian has a significant variety of grammatical categories, which complicates the process of eliminating morphological ambiguity in the text.

The article examines the ELMo model on the example of the English language, the MEMM model on the Russian language corpus, the hybrid method (contextual and statistical-probabilistic methods) on the Turkish language corpus, and the method of removing homonyms from a corpus marked with homonyms on the Tatar language. Initially, the methods used in the practice of Turkish and Tatar corpora seem to be the most suitable for use in the formation of the Kazakh language corpus. This is due to the simpler structure of the underlying algorithms, as well as their focus on considering the specifics of agglutinative languages. ELMo and MEMM algorithms, due to their more complex structure and the need to prepare more materials for their training, can be used in the future, considering the peculiarities of the Kazakh language. The study proposes a classification of homophones of the Kazakh language (20 groups) and determines that homophones formed from verbs and nouns make up 65.7% of the homophones of the Kazakh language. Rules have been developed for such homophones that can be included in the language corpus of Kazakh texts. Thus, the use of Turkish and Tatar experience in the identification of homonyms/homoforms with the inclusion of general rules developed based on the experiment in the Kazakh language corpus created at the Institute in the automatic recognition algorithms will significantly improve the accuracy of their detection at this stage.

The present study is more aimed at solving the problem from the point of view of analysing morphology and syntax and requires a more detailed study in semantic

and contextual aspects. Recommendations on the use of algorithms applied to corpora of texts in other languages in Kazakh are theoretical and based on modelling processes due to linguistic similarities and therefore need to be further studied with due regard to the peculiarities of the Kazakh language from a practical perspective. In addition, the rules for inclusion in the recognition algorithm proposed in this paper have been developed only for more than half of the homophones in the Kazakh language, which means that the remaining groups of homophones still need to be studied in more detail to create rules for their identification in the text.

References

ALMALKI, I., METWALLY, A.A., and ASIRI, E. (2025). "The Role of Online Dictionaries in Translating Literary Collocations: An Experimental-based Case Study of BA Students at KKU". *Dragoman,* 17, 136-156. https://doi.org/10.63132/ati.2025.therol.99950666

ALWATEER, M.M., ELMEZAIN, M., FARSI, M., and ATLAM, E. (2023). "Hidden Markov Models for pattern recognition". In Khalil, H., and Ghaffari, A. (eds.), *Markov Model – Theory and Applications*. Rijeka, IntechOpen. https://doi.org/10.5772/intechopen.1001364

BASHMANIVSKIY, O. (2016). "The problems of automated translation of business correspondence using free software". *Society. Document. Communication,* 1(2), 79-90.

BILOUS, N., and BARLADIUHA, A. (2020). "Structurally semantic and stylistic features of phraseological units". *International Journal of Philology*, 24(4), 70-77. https://doi.org/10.31548/philolog2020.04.070

BORYSOVA, N. (2023). "Mind maps as an effective tool in "Practical English language course"". *Scientia et Societus,* 2(2), 96-109. https://doi.org/10.69587/ss/2.2.2023.96

BRAIT, B., SOUZA, G.T., AMORIM, M., S.A.P.F. PENTEADO, P.M.H. CRUZ, R.C.G., STELLA, P.R., and STORTO, L.J. (2025). "Bakhtin and Linguistics: A Dialogue Settled in the Beginning of the 20's/". *Bakhtiniana,* 20(1), e66039e. https://doi.org/10.1590/2176-4573e66039

CASAS, B., HERNÁNDEZ-FERNÁNDEZ, A., CATALÀ, N., FERRER-I-CANCHO, R., and BAIXERIES, J. (2019). "Polysemy and brevity versus frequency in language". *Computer Speech & Language*, 58, 19-50. https://doi.org/10.1016/j.csl.2019.03.007

CHAIKA, O. (2023). "Key advantages of multiculturalism for foreign language teaching and learning". *Humanities Studios: Pedagogy, Psychology, Philosophy*, 11(1), 119-128. https://doi.org/10.31548/hspedagog14(1).2023.119-128

CING, D.L., and SOE, K.M. (2020). "Improving accuracy of part-of-speech (POS) tagging using hidden Markov model and morphological analysis for Myanmar Language". *International Journal of Electrical and Computer Engineering*, 10.2, 2023-2030. http://doi.org/10.11591/ijece.v10i2.pp2023-2030

DOSZHAN, G. (2016). "Semantic and pragmatical aspects of English business lexemes in Turkic languages". *Procedia Economics and Finance*, 39, 24-31. https://doi.org/10.1016/S2212-5671(16)30236-2

DURO, R., and KONDRATENKO, Y. (2015). *Advances in Intelligent Robotics and Collaborative Automation.* River Publishers. https://doi.org/10.13052/rp-9788793237049

ELOV, B.B., HAMROYEVA, S.M., and AXMEDOVA, X.I. (2023). "Methods for creating a morphological analyzer". In Zaynidinov, H., Singh, M., Shanker Tiwary, U., and Singh, D. (eds.), *Intelligent Human Computer Interaction 14th International Conference*. Cham, Springer, pp. 27-38. https://doi.org/10.1007/978-3-031-27199-1_4

FASHWAN, A., and ALANSARY, S. (2021). "A morphologically annotated corpus and a morphological analyzer for Egyptian Arabic". *Procedia Computer Science*, 189, 203-210. https://doi.org/10.1016/j.procs.2021.05.084

HABIBI, A.A., HAUER, B., and KONDRAK, G. (2021). "Homonymy and polysemy detection with multilingual information". In Vossen, P., and Fellbaum, C. (eds.), *Proceedings of the 11th Global Wordnet Conference*. Potchefstroom, Global Wordnet Association, pp. 26-35. https://aclanthology.org/2021.gwc-1.4

HASSAN, M.A., KHAN, R.R., KABIR, F., and RAHMAN, C.M. (2009). "Finding the appropriate meaning of polysemous words using context dependency". In *Proceedings of 2nd International Conference on Data Management (ICDM'2009)*. https://www.researchgate.net/publication/292132545_Finding_the_Appropriate_Meaning_of_Polysemous_Words_Using_Context_Dependency

HOFF, S.L., and BARBOZA, G. (2025). "Languages, Language, and Linguists: The Study of the Diversity of Languages According to Saussure and Benveniste". *Bakhtiniana*, 20(1), e65692e. https://doi.org/10.1590/2176-4573e65692

HROMKO, T., PANCHUK, L., and MATVEIKO, O. (2023). "Versification of language units from the lens of morphological statistics (Modern British and German poetry)". *International Journal of Philology*, 27(2), 23-32. https://doi.org/10.31548/philolog14(2).2023.03

KALJANOV, A. (2021). "Some questions of classification of homonyms in the Karakalpak language". *Bulletin of the Karakalpak Branch of the Academy of Sciences of the Republic of Uzbekistan*, 265.4, 196-199. https://www.researchgate.net/publication/357164236_Qaraqalpaq_tilindegi_omonimler_klassifikaciyasinin_ayirim_maseleleri

KANERVA, J., GINTER, F., and SALAKOSKI, T. (2021). "Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks". *Natural Language Engineering*, 27.5, 545-574. https://doi.org/10.1017/S1351324920000224

KESSIKBAYEVA, G., and CICEKLI, I. (2016). "A rule based morphological analyzer and a morphological disambiguator for Kazakh language". *Linguistics and Literature Studies*, 4.1, 96-104. https://doi.org/10.13189/lls.2016.040111

KHAKIMOV, B., GATAULLIN, R., and GILMULLIN, R. (2016). "Grammatical disambiguation in the Tatar national corpus". In Moreno Ortiz, A. and Pérez Hernández, C. (eds.), *8th International Conference on Corpus Linguistics*. Málaga, University of Málaga, Spanish Association for Corpus Linguistics, pp. 236-244. https://doi.org/10.29007/jkgl

KIM, H., and KIM, H. (2020). "Integrated model for morphological analysis and named entity recognition based on label attention networks in Korean". *Applied Sciences*, 10.11, 3740. https://doi.org/10.3390/app10113740

KONDRATENKO, Y.P. (2014). "Robotics, Automation and information systems: Future perspectives and correlation with culture, Sport and life science". *Lecture Notes in Economics and Mathematical Systems,* 675, 43-55. https://doi.org/10.1007/978-3-319-03907-7_6

KULYK, O.D. (2023). "Training future translators in the age of artificial intelligence". *Scientia et Societus,* 2(1), 48-56. https://doi.org/10.31470/2786-6327/2023/3/48-56

LEE, Y. (2021). "Systematic homonym detection and replacement based on contextual word embedding". *Neural Processing Letters*, 53, 17-36. https://doi.org/10.1007/s11063-020-10376-8

LI, S., PAN, R., LUO, H., LIU, X., and ZHAO, G. (2021). "Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modelling". *Knowledge-Based Systems*, 218, 106827. https://doi.org/10.1016/j.knosys.2021.106827

LIU, F., LU, H., and NEUBIG, G. (2018). "Handling homographs in neural machine translation". In Walker, M., Heng, J. and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Association for Computational Linguistics, pp. 1336-1345. https://doi.org/10.18653/v1/N18-1121

PAPA, E., MANDRI, H., ZOKIROVA, N., RAXMANOVA, A., and PETROVA, E. (2025). "The Impact of Quality Translation on the Correct Interpretation of Literary Works". *Dragoman,* 2025(18), 118-143. https://doi.org/10.63132/ati.2025.theimp.36423928

PATIL, N., PATIL, A., and PAWAR, B.V. (2020). "Named entity recognition using conditional random fields". *Procedia Computer Science*, 167, 1181-1188. https://doi.org/10.1016/j.procs.2020.03.431

PETERS, M.E., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K., and ZETTLEMOYER, L. (2018). "Deep contextualized word representations". In Walker, M., Heng, J. and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Association for Computational Linguistics, pp. 2227-2237. https://doi.org/10.18653/v1/N18-1202

QUECEDO, J.M.H., KOPPATZ, M.W., and YANGARBER, R. (2020). "Neural disambiguation of lemma and part of speech in morphologically rich languages". In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, European Language Resources Association, pp. 3573-3582. https://aclanthology.org/2020.lrec-1.439

RICE, C.A., BEEKHUIZEN, B., DUBROVSKY, V., STEVENSON, S., and ARMSTRONG, B.C. (2019). "A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings". *Behavior Research Methods*, 51, 1399-1425. https://doi.org/10.3758/s13428-018-1107-7

SAK, H., GÜNGÖR, T., and SARAÇLAR, M. (2007). "Morphological disambiguation of Turkish text with perceptron algorithm". In Gelbukh, A.

(ed.), *Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg, Springer, pp. 107-118. https://www.researchgate.net/publication/226961087_Morphological_Disambiguation_of_Turkish_Text_with_Perceptron_Algorithm

SERDIUK, N. (2023). "Methodology and organization of professional research and academic integrity in the formation of a modern Foreign language and literature teacher". *Professional Education: Methodology, Theory and Technologies*, 9(1), 159-179. https://doi.org/10.31470/2415-3729-2023-17-159-179

SHASHANK, U., VENKATESH, B.N., RAJESHWARI, S.B., and KALLIMANI, J.S. (2019). "Identification and contextual semantic retrieval of polysemy words". *International Journal of Recent Technology and Engineering*, 8.2S8, 1201-1204. https://doi.org/10.35940/ijrte.B1038.0882S819

SHLAPAK, I. (2016). "Terms polysemy in translation of scientific and technical literature". *Society. Document. Communication*, 1(1), 227-234.

SHYNGYSSOVA, N.T., and SKRIPNIKOVA, A.I. (2018). "Polylingual periodicals of Kazakhstan". *Bulletin of Al-Farabi Kazakh National University, Series of Journalism*, 49.3, 38-45. https://bulletin-journalism.kaznu.kz/index.php/1-journal/article/view/1029

STRILETS, V. (2021). "Application of corpus technologies in teaching specialized translation". *Humanities Studios: Pedagogy, Psychology, Philosophy,* 9(4), 48-52. https://doi.org/10.31548/hspedagog2021.04.048

TANG, X. (2006). "English morphological analysis with machine-learned rules". In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*. Wuhan, Tsinghua University Press, pp. 35-41. https://doi.org/http://hdl.handle.net/2065/29035

TEMIRGALIEVA, A. (2016). *Frequency dictionary of Kazakh language developed*.

TOEWS, D., and VAN HOLLAND, L. (2019). "Determining domain-specific differences of polysemous words using context information". In *Proceedings of International Conference on Requirements Engineering – Foundation for Software Quality*. Wachtberg, Fraunhofer Institute for Communication, Information Processing and Ergonomics. https://publica.fraunhofer.de/handle/publica/410113

VAN DEN BEUKEL, S., and AROYO, L. (2018). "Homonym detection for humor recognition in short text". In Balahur, A., Mohammad, S. M., Hoste, V., and Klinger, R. (eds.), *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Brussels,

Association for Computational Linguistics, pp. 286-291. https://doi.org/10.18653/v1/W18-6242

WILSON, K., and MARANTZ, A. (2022). "Contextual embeddings can distinguish homonymy from polysemy in a human-like way". In *Proceedings of the 5th International Conference on Natural Language and Speech Processing*. Trento, Association for Computational Linguistics, pp. 144-155. https://aclanthology.org/2022.icnlsp-1.17

WOLDEYOHANNIS, M.M., and MESHESHA, M. (2022). "Usable Amharic text corpus for natural language processing applications". *Applied Corpus Linguistics*, 2.3, 100033. https://doi.org/10.1016/j.acorp.2022.100033

YURET, D., and TÜRE, F. (2006). "Learning morphological disambiguation rules for Turkish". In *Proceedings of the Main Conference on "Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics"*. New York, Association for Computational Linguistics, pp. 328-334. https://doi.org/10.3115/1220835.1220877

ZHUBANOV, A.K., and ZHANABEKOVA, A.A. (2016). *Corpus linguistics: Educational tool*. Almaty, "Kazakh Language" Publishing House.