Management Application and Research Center (YUVAM)

MACHINE LEARNING

Editors:

Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci

YEDİTEPE UNIVERSITY
PRESS

Editors:
Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci

MACHINE LEARNING

Yeditepe University Press: 63

Machine Learning
Editors: Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci

ISBN: 978-975-307-139-0
Istanbul, 2023

Cover and page design: Savaş Yıldırım

Yeditepe University Press
Yeditepe University
Inonu Mah. Kayisdagi Cad. 326A 26 Agustos Yerlesimi 34755
Atasehir – Istanbul/TURKEY
Ph. 0090 216 578 00 00 / 37 16
Certificate number: 41307

Editors:

Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci

MACHINE LEARNING

YEDİTEPE UNIVERSITY
PRESS

CONTENTS

Computational Model of Morphology and Stemming of Karakalpak Words on a Complete Set of Inflectional Endings

Ualsher Tukeyev, Shyngys Ryskulbek, Kazbek Amankeldi
Al-Farabi Kazakh National University, Almaty, Kazakhstan
ualsher.tukeyev@mail.ru, rysklbeksyngys@gmail.com,
98.kazbek@gmail.com

Abstract.

For low-resource languages, the development of resources and tools that improve the use of neural machine translation is relevant. However, large dictionaries and corporations require considerable memory, which limits the use of neural machine translation and can cause memory errors. These restrictions can be removed by dividing each word into morphemes in parallel primary corpuscles. Therefore, it is advantageous to use the segmentation method (CSE), which is based on a complete set of suffixes, which allows you to reduce the vocabulary of the initial corpuscles.

The purpose of this work is to get acquainted with the features of the morphological grammar of the Karakalpak language, to consider and summarize all sets of suffixes, stems, stop words. And the use of accumulated linguistic resources in the segmentation of texts.

Keywords:

Karakalpak language, text segmentation, complete set of language endings, agglutinative languages, Turkic languages.

1. Introduction

The huge flow of information on the Internet has led to the rapid development of the natural language processing industry (NLP). Currently, various research mechanisms are developing their own projects, such as the exchange of information between users, machine translation of information, the presence of spam filters, checking e-mail and the development of question-answer systems. However, due to the lack of knowledge of the structure of some languages, there are issues that do not fully meet the needs of the user.

Today, Karakalpak is one of the least resourceful Turkic languages. The increase in information resources available on the Internet in the Karakalpak language leads to the need to develop search engines in the Karakalpak language to access them. This project provides an algorithm for segmenting words from texts in the Karakalpak language, in particular, a segmentation algorithm based on a complete set of language endings, its features.

The algorithm takes into account the specifics of the input data and is designed so that the necessary calculations are performed in a single pass. For the algorithm to work, you need a complete set of connections in that language, as the name implies. Since Karakalpak is a low-resource language, no one has created a complete set of prefixes before me, so compiling a complete set of prefixes was a full-fledged study.

Scientific contribution: Experiments were performed using the accumulated linguistic resources and machine translation algorithm based on the morphological model of the CSE.

In addition, various research mechanisms are developing their own projects, such as the exchange of information between users, machine translation of information, the presence of spam filters, checking e-mail and the development of question-answer systems. However, due to the lack of knowledge of the structure of some languages, there are issues that do not fully meet the needs of the user. One of the problems of research engines today is the morphological and morpheme analysis of words encountered in the processing of user requests. An example of such a language is Uzbek, which belongs to the Turkic language family.

Karakalpak is one of the agglutinative languages. That is, in this language, each grammatical meaning is expressed by individual affixes. The term affix in the grammar of the Karakalpak language is taken in the same general sense as in the grammar of other Turkic languages. It means prefixes, infixes, suffixes, suffixes, prefixes. Today, the structure of the Karakalpak language is complicated by the influence of Arabic, Persian and Russian. To solve such problems, we pay special attention to the morphological grammar of the Karakalpak language, which allows us to get accurate results from search queries.

2. Motivation and related works

Currently, there is no segmentation algorithm based on a complete set of connections to Karakalpak texts. And the complete set of suffixes in the original language is very important for the morphological analysis of sentences in that language, as it guarantees that any word is analyzed in terms of grammatical (lexical) characteristics. This is an important issue for a machine translation system, as the complete set of connections of one language in machine translation determines the completeness of the system of conversion from one language to another at the morphological level.

Getting acquainted with the features of the morphological grammar of the Karakalpak language, reviewing and summarizing all sets of suffixes, stop words, creating programs for stemming and segmentation of texts in the Karakalpak language.

The task of morphological segmentation was proposed by Harris [2]. One of the most important tasks in neural machine translation is to segment the source text. This task is associated with the reduction of NMT vocabulary, as well as the solution of the problem of rare and unknown words. Research work related to segmentation for NMT can be divided into work based on the BPE method, Morfessor and final state converters. [3,4,5].

Sennrich et al. developed methods for dividing corpses into frequent strings of symbols. Specifically, the known byte-pair encoding (BPE) compression method was used in these methods in the Anglo-German

and Anglo-Russian NMT systems. The authors adapted the BPE algorithm to the segmentation task and created an open dictionary. The advantage of the BPE-based method of segmentation is that it allows you to often translate rare words into phrases and translate unknown words, which is one of the main goals of the NMT. The BPE segmentation method is the preferred approach to subword segmentation.

The BPE-based method involves dividing words into different variants; however, this approach does not apply to languages with a rich morphology, such as Kazakh. For example, during the training period, the words "projects", "project of" and "of the project" are presented in the form "project" (correct segmentation of "projects" by the BPE method, respectively, "project". (correct segmentation - "project critique") and "project </w>" (without segmentation). When the BPE method is applied to files during testing, these words are not separated, but are left as whole words. It is explained that the experiments in Section 4 confirm this hypothesis, so the vocabulary of BPE-based segmentation is greater than that of morphological segmentation.

Sanchez-Cartagena and others demonstrated morphological segmentation using Apertium and combined the results of a rule-based machine translation (Sánchez-Cartagena et al., 2019). They segmented the original text using a rule-based morphological analyzer. If the word does not have the correct segmentation, many segmentation variants have been created, as there have been known suffixes that correspond to the word.

3. Morphological calculation model of the Karakalpak language

The Karakalpak language is closest to the Kazakh and Nogai languages. He also mastered a large vocabulary and, to some extent, the grammar of the Uzbek language. Like the vast majority of Turkic languages, Karakalpak has a vowel consonant, no agglutinative and grammatical gender. The word order is usually a subject-object-verb.

Agglutinative structure is an important feature that should be taken into account when translating from Karakalpak and bring it closer to other languages of the Turkic group of languages. Agglutination is the addition of many suffixes, each of which has its own grammatical meaning. For example, the suffix -ажақ is used to denote the future tense

(Алажақ – Ол алады); The suffixes -ман / -мен and -сан / -сен are used to denote this moment (аламан – мен алмаймын, бересен – сіз бересіз); and the suffix -n (алыппан - алдым) serves as a suffix of the past tense (with the meaning of execution).

Another important morphological feature of the translation into Karakalpak is the productive use of gerunds ending with the verb phrases -a / -e / -у, -п, which are formed in accordance with the analytical principle and represent the method of action (Жеп тойып алдым – Мен тойдым). The main objectives of the study are literature search, segmentation analysis and review of dissertation materials, compilation of complete sets of endings in Karakalpak, segmentation and study of Karakalpak text based on a computational model of morphology based on a complete set of endings.

Karakalpak language connection system. Let's divide the system of word endings in the Karakalpak language into 2 different classes. The first is for nouns (nouns, adjectives, numerals) and the second is for verbs. Below I show the system of suffixes for nouns in the Karakalpak language.

In the Karakalpak language, there are four types of suffixes that connect to the base of nouns:

- plurals (denoted by K),
- dependency connections (denoted by T),
- auxiliary connections (denoted by C),
- split connections (denoted by J),
- Stem (denoted by s).

Let's look at different options for connecting types of connections. It can be a single connection, it can be two different connections, it can be three and four different connections. We determine the number of all possible variants in the sequence of this placement by the following formula:

$A_n^k = n!/(n-k)!$

Then the number of placements is determined as follows:

$A_4^1 = 4!/(4-1)! = 4$,
$A_4^2 = 4!/(4-2)! = 12$,
$A_4^3 = 4!/(4-3)! = 24$,
$A_4^4 = 4!/(4-4)! = 24$.

Then the number of placements is determined as follows: aIt is calculated by a pure formula. Now let's count the possible sequences according to the language rules.

1) **K, T, C, J**

For example: балалар, балам, баламның, баламын.

2) **KT, TC, CJ**, JK
KC, TJ, CT, JT
KJ, TK, CK, JC.

For example: китапларым, китапларымды, бизлер балалармыз, китабымнын, баланбыз, баладанбыз.

3) **KTC**, **KTJ**, **TCJ**, TCK, CJK, CJT, JKT, JKC
KCJ, KCT, TJK, TJC, CTK, CTJ, JTK, JTC
KJT, KJC, TKC, TKJ, CKT, CKJ, JCK, JCT.

For example: китапларымды, балаларынбыз, баланнанбыз, балалардынбыз.

4) KTJC, TKJC, CKTJ, JKTC
KTCJ, TKCJ, CKJT, JKCT
KJTC, TJKC, CTKJ, JTKC
KJCT, TJCK, CTJK, JTCK

KCTJ, TCJK, CJKT, JCKT
KCJT, TCKJ, CJTK, JCTK

In the order of placement, there are four in 1 sequence, six in 2 sequences, and four in three different sequences. The total number of placement sequences is 15. In accordance with the sequence of placement of the connections we have identified, I took the children's word and gave an example of all these sequences.

In general, Karakalpak is the most similar language to Kazakh among the Turkic-speaking languages. Sound harmony, as in the Kazakh language, depends on the law of synharmonism. That is, the connections connected depend on the thickness and thinness of the final sounds, or in other words, the law of synharmonism.

4. Experiments and results

Stem dictionary is a necessary resource for the application of stemming, morphological analysis algorithms in any natural language. This is due to the fact that when applying the stemming algorithm to the text of the Karakalpak language, there were many mistakes due to the lack of the necessary stem words. That is, they recognized the last letters of the stems, which should not be divided, as a connection and separated them.

During the study, different methods were used to collect stem words:

1. The first Lexicon Free stemming algorithm was compiled on the basis of distributed stem words, using Uzbek language texts. However, it was very difficult to identify many words due to the fact that this algorithm incorrectly divided many words. But as a result of this algorithm, more than 1,000 stem words were collected.
2. With the Lexicon Free stemming algorithm, it became difficult to collect stem words, so we moved to collect words based on books. We took the Karakalpak-English translation dictionary, cleaned it and added about 19,000 new stems.
3. A common Karakalpak-Russian dictionary written in the Cyrillic alphabet was found and translated into the Latin alphabet. There are about 1,500 of them.

4. Many stem words are based on the use of the Stemming algorithm with stem stemming algorithm. Using this stemming algorithm, many stem words that were not on the list were identified.
5. About 2,000 names of people in the Karakalpak language were added to the list.

As a result, a stem dictionary of 25,161 words was created.

Application of stemming algorithm to the text of the Karakalpak language. To use the morphological segmentation algorithm for the Karakalpak language, it is necessary to perform two main stages:

1. Divide all words in the text into bases and suffixes;
2. Divide the connections into individual segments.

Accordingly, the stemming algorithm is first performed on the texts. Stemming is the process of finding the basis (stem) for a given original studied word. Stemming algorithms are divided into several types depending on their performance, accuracy, and how to overcome any stemming problems. One of them is a stemming algorithm based on a set of connections used in the research. Basically, it is said that these algorithms are performed for all languages belonging to the group of Turkic-speaking languages, because the languages belonging to this group are agglutinative. That is, the structure of words is realized as a result of the addition of suffixes and suffixes to the root. And on the basis of this model, the Kazakh language texts were stemming, segmented and showed excellent results. The next step was to test these programs in other Turkic-speaking languages. In particular, according to the topic of my research, the research language is Karakalpak. The CSE-based stemming algorithm includes the following steps:

1) The length of the word is determined: L (w).
2) Determine the length of the conjunction of the word under consideration: L [e (w)] max = L (w) - 2, where 2 is the minimum length of the base.

3) If L (w) ≤ L (e) max; then L [e (w)] is replaced by L [e (w)] max: L [e (w)] = L [e (w)] max.
4) The value of L [e (w)] is replaced by the value of L (e) max: L [e (w)] = L (e) max.
5) Select the suffix e (w) corresponding to the length L [e (w)] for the word w under consideration.
6) Check that e (w) matches the length L [e (w)] with the suffixes in the complete lists. If it corresponds to any connection, then the root word is determined by the formula st (w) = w - e (w). If this step is performed, the algorithm completes.
7) In case of discrepancy, the length of the measured connection of the word under consideration is reduced by another number: L [e (w)] = L [e (w)] - 1.
8) If L [e (w)] <1, then the word w has no suffix. The algorithm ends here. In other cases, it goes back to step 5 until the corresponding connection is determined.

In accordance with the morphological model of the CSE, we experimented using stem vocabulary, using the above-mentioned linguistic resources and the stemming algorithm. Uzbek texts about computers and books were obtained from the Internet as a text resource. In total, experiments with 82 sentences of 1,046 words were performed using a basic algorithm based on the CSE morphological model. We manually checked each word of the result obtained by the experiment separately. As a result, the accuracy of the model with Karakalpak texts showed very good results. As these algorithms experimented with the text, suffixes that were not in the list were identified, and new stem words were added to the list. During the experiment with the use of accumulated linguistic resources, the accuracy increased from 90.6% to 94.5%.

Professor U. Tukeyev developed and proposed an algorithm for machine translation from Kazakh into Karakalpak languages (Tatar, Turkish, Uzbek, Karakalpak, etc.) based on the model of a complete set of connections.

Here is a step-by-step description of this advanced machine translation algorithm:

We consider the given text or sentence in words.

1. A word is given: w_i.
2. Distinguish the si system and the e_i suffix of the given word w_i using the Stemming algorithm with stems-lexicon according to the CSE (Complete Set of Endings) morphology model [1, 2]. $w_i = s_i + e_i$.
3. The corresponding translation of the system s_i in the original language (SL) in the target language (TL) is searched in the dictionary of identified systems in the Kazakh and Karakalpak languages. If the dictionary finds the system s_i^{TL} corresponding to the system s_i^{SL}, then s_i^{TL} is taken as a translation of the given s_i^{SL}. If not found, the given s_i^{SL} translation will be represented by itself, ie s_i^{SL}.
4. The corresponding translation of the suffix e_i in the original language (SL) in the target language (TL) is searched according to the table of morphological connections of the Kazakh and Karakalpak languages. The EndingsSL column in the table contains rows with the e_i^{SL} connection. The e_i^{TL} connection in the EndingsTL column corresponding to the e_i^{SL} connection from the table is taken as a translation. If e_i^{TL} is more than one, then e_i^{TL} is selected based on the rules of the language for s_i^{TL}.
5. s_i^{TL} and e_i^{TL} will be combined and presented as a translation of the original word w_i.
6. The end.

Table 1. Examples of algorithm steps.

Algorithm steps	For example
w_i	тәсілдері
$w_i = s_i + e_i$	тәсіл – дері
s_i^{SL} according to s_i^{TL}	тәсіл – eritpe
e_i^{SL} according to e_i^{TL}	дері – leri
$w_i^{TL} = s_i^{TL} + e_i^{TL}$	eritpeleri

In general, the preparation of a database of basics takes a lot of time and manpower. Due to the time constraints of scientific and technical work, a small translation model with 461 bases and 22 closing words was developed. The results of the translation are given in the table below. The higher the stem vocabulary, the higher the translation result.

When testing the operation of a machine translation algorithm based on the model of a complete set of Kazakh and Karakalpak suffixes, a text consisting of 45 parallel sentences in the Kazakh and Karakalpak languages was used.

Table 2. Results of BLEU metrics.

	TER	*WER*	*BLEU*
Kazakh-Karakalpak	0.36	0.32	49

The lower the TER value, the higher the translation quality. Although the results of the official evaluation of machine translation using the proposed technology showed a high percentage of errors, the translation itself was very close to the translation model. This indicates that the proposed technology can be used. In addition, the proposed machine translation technology does not require a large number of parallel housings for training.

5. Conclusion and future work

This article develops a method of morphological segmentation of the Karakalpak language based on CSE. The paper presents a dedicated language resource of Karakalpak language suffixes needed to solve the problem of segmentation of Karakalpak texts for neural machine translation. To perform morphological segmentation, a complete set of suffixes was collected and possible types of noun and verb endings of words were obtained using a combinatorial approach based on the semantic correspondence and compatibility of words. The proposed CSE-based segmentation may also apply to other languages in the Turkic language group.

The main part of the article provides a general and morphological description of the Karakalpak language. Based on the description, a list of

endings in the Karakalpak language was obtained, and a combinatorial definition was given for each type of connection. A review of the works on the morphological segmentation and its status in the Karakalpak language was identified. Based on the developed model, the results were tested in morphological segmentation and stemming problems and described.

As a result of this project, a morphological model of the Karakalpak language was developed and experiments were performed on such models as stemming algorithms and morphological segmentation. The obtained results will help to make a comparative analysis with the structure of other modern Turkic languages and to solve the problem of language distance. In addition, based on the obtained model, it can help to perform neural machine translation into Turkic languages other than Karakalpak.

One of the difficulties encountered during the implementation of this project is the lack of electronic records in the Karakalpak language. In the future, it is planned to increase the size of the case in Karakalpak and translate machine translation into Kazakh. Also, CSE-based Kazakh, Turkish, Uzbek, etc. On the basis of modern languages used in the Turkic group, such as It is also possible to study the problems of obtaining information, morphological analysis on the basis of the obtained CSE.

REFERENCES

Tukeyev, U.: Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. In: Proceedings of the International Conference "Turkic Languages Processing" TURKLANG 2015, Kazan, Tatarstan, Russia, 17-19 September, pp. 91-100 (2015)

Harris, Z Methods in structural linguistics. Chicago University Press (1951)

Sennrich, R., Haddow, B., & Birch, A. 2016. Neural machine translation of rare words with subword units. Proceedings of the 54th annual meeting of the association or computational linguistics, 1, 1715-1725.

Ataman, D., Negri, M., Turchi, M., Federico, M.: Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. Prague Bull. Math. Linguist.108 (1), 331-342 (2017)

Sanchez-Cartagena, V. M., & Toral, A. Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. Proceedings of the first conference on machine translation. ACL, 2016.

Tacorda, A. J., Ignacio, M. J., Oco, N., & Roxas, R. E. 2017. Controlling byte pair encoding for neural machine translation. 2017 international conference on Asian language processing, 168-171.

Creutz, M., & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, 1(106-113), 51-59.
https://tuhat.helsinki.fi/ws/portalfiles/portal/77193625/Creutz05akrr.pdf

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing (TSLP), 4(1), 1-34. DOI: 10.1145/1187415.1187418

Creutz, M., & Linden, K. 2004. Morpheme segmentation gold standards for Finnish and English (Tech. rep.A77). Publications in Comput-

er and Information Science, Helsinki University of Technology.

Banerjee, T., & Bhattacharyya, P. (2018). Meaningless yet meaningful: Morphology grounded subword-level NMT. Proceedings of the second workshop on subword/character level models (pp. 55-60)

Papli, K. (2017). Morpheme-aware subword segmentation for neural machine translation [bachelor thesis].University of Tartu Institute of Computer Science. p.27.

Briakou, E., & Carpuat, M. (2019). The University of Maryland's Kazakh- English neural machine translation system at WMT19. Proceedings of the fourth conference on machine translation (pp. 134-140), August 1-2.

Casas, N., Fonollosa, J. A. R., Escolano, C., Basta, C., & Costa-Jussa, M. R. (2019). The TALP-UPC machine translation systems for WMT19 news translation task: Pivoting techniques for low resource MT. Proceedings of the fourth conference on machine translation (WMT19), August 1-2 (pp. 155-162).

Rakesh Kumar, Minu Bala, and Kumar Sourabh.: 2018. A study of spell checking techniques for indian languages. JK Research Journal in Mathematics and Computer Sciences, 1(1).

Rakhimova D., Turganbayeva A. Approach to Extract Keywords and Keyphrases of Text Resources and Documents in the Kazakh Language. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, 12496 LNAI, pp. 719–729.

Tukeyev U.A., Turganbaeva A.O.: Lexicon-free stemming for the Kazakh language. In: Materials of the International Scientific Conference "Computer science and Applied Mathematics" dedicated to the 25th anniversary of the Independence of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computational Technologies, Part II, Almaty, September 21-24, 2016.

Ualsher Tukeyev, Aliya Turganbayeva, Aidana Karibayeva, Dina Amirova, and Balzhan Abduali.: Language_Resources_for_Kazakh_language. https://github.com/NLP-KazNU/Language_Resources_for_Kazakh_language

Study and development of an approach for identifying incorrect words for the Kazakh language in semi-structured data

Diana Rakhimova, Yntymak Abdrazakh
Al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru, yntymak05099@gmail.com

Abstract.

Nowadays, the amount of information that humans and machines produce in natural language is increasing significantly. Information retrieval systems, machine translation, spell checkers, and conversational systems perform analysis and processing of natural language texts. Thus, the range of automatic text processing systems is wide and includes a variety of tasks. One of the most important tasks (directions) of natural language processing is the search for errors in the text and words in it, the identification and correction of incorrect words. The article discusses methods for identifying incorrect words of the Kazakh language in semi-structured data, a model and methods for identifying incorrect words of a general language character. The study identified the most common types of errors in the words of the Kazakh language, obtained from semi-structured data. The necessary materials for identifying incorrect words in the language are collected and the results of the experiment are presented.

Keywords:

Semi-structured data, Kazakh language, Approach, Internet, Social network, incorrect words, Stemming.

6. Introduction

The huge flow of information on the Internet and social networks has led to the rapid development of the natural language processing industry, computer linguistics. Currently, various research mechanisms are developing their own projects, such as the exchange of information between users, machine translation of information, verification of e-mail, and the development of question-answer systems [1]. In general, the task of finding and correcting errors in texts and words in them is one of the main tasks of word processing in natural language. For more than half a century, this topic has not lost its relevance, new methods have emerged, the scope of its application is expanding. On the Internet and Instagram, Vkontakte, Twitter, and other social networks, applications are very attractive in terms of receiving and analyzing information in messages, because the information in these systems is real, it appears at this time [2]. However, the text on the Internet often differs from the generally accepted norms of the language. There are errors, deliberate distortion of words, mistakes [3]. Words with such errors, that is, incorrect words, can be processed and analyzed by identifying and correcting the necessary information. And these errors makes the text harder to read and, worse, harder to process. Natural language processing requires normalized forms of a word because incorrect spelling or digitization of text decreases informational value. A spelling error, for example, in a database of medical records, diminishes efficiency of the diagnosis process, and incorrectly written comments and publications of users on the Internet can influence research or organizational processes [4].

Due to the fact that the Kazakh language belongs to the group of low-resource languages, it is known that there are few translation systems, dictionaries, corpus (multilingual and bilingual), systems, and programs for detecting and correcting errors in words. Accordingly, today it is important to develop resources and tools, systems, as spelling errors detection programs that improve the use of languages with limited resources, such as the Kazakh language.

6.1 Semi-structured data

Semi-structured data is data that does not conform to the rigid structure of tables and relationships in structural relational database models. Online information is not always relevant to a particular field of knowl-

edge. In this regard, many organizations and scientists are developing specific algorithms for creating the structure of the text, which is not related to the field of education [5]. Examples of systems with semi-structured data include comments, publications and texts written by users on the Internet, websites, and social networks [6]. Data from such systems is of great interest for research and applications, as it is possible to publish people's opinions and moods on any issue in real-time and contribute to the increase of information. It also helps to change people's attitudes towards business, politics and the social system today. Each type of data has its own characteristics, which must be taken into account when collecting, preparing, pre-processing, and describing objects. We used information from the Internet and social media and this data is semi-structured as mentioned above. And this data has been used in research and experiments.

7. Related works

The problems and works on spelling error detection and correction in text began in the 1960s and continues to the present day. There are good reasons to continue research in this area in order to improve quality and performance, as well as to expand the range of possible applications. For example, even though system programs (language processors, etc.) are becoming more powerful and complex, they do not help the user (with very few exceptions) to correct many obvious spelling errors in the input source data [7]. For 50 years of solving the problem of finding and correcting errors, researchers have tried a huge number of different methods. From character codes and n-gram acceptance tables and the direct application of the Damerau-Levenshtein distance to the active use of various machine learning methods, phonetic information about a word and machine translation methods. And the construction of systems for detecting and correcting errors in the text faces a number of fundamental and unresolved issues: compact storage of dictionaries, effective methods of morphological and syntactic analysis, and a system of scientific editors, i.e. a person who conducts literary and scientific processing of scientific and technical texts [8]. English text correction systems include «Grammarly», «ReversoSpeller» and more applies. Russian text editing systems include: «Orfogramka», «ORFO» and others.

For agglutinative languages (such as Turkish, Kyrgyz, etc.) there are error checking systems, for example, a built-in spell checker in MS Word. But these systems are not suitable specifically for the Kazakh language. And unfortunately, there are no (in the public available) analogues of the previously listed systems for the Kazakh language.

An analysis of text verification and correction systems for English and Russian languages. During the research and analysis, texts of different styles, texts on the Internet and social networks, messages were considered. The advantages and disadvantages of the systems are shown in Table 1.

Table 1. Comparative characteristics of text verification and correction systems (system for English and Russian).

Text checking and editing systems	Disadvantages	Advantages
Microsoft.com	Recognizes some words that are not in the inserted dictionary and classifies them as incorrect words.	In most cases, it identifies incorrect words in a large text.
ORFO - https://online.orfo.ru/	Shows good results in the use of individual words, can not find some errors in large texts.	Finds complex errors and mistakes, different options. Shows all possible word correction options.
Orfogramka - https://orfogrammka.ru/	In some cases, it can't detect word errors.	Identifies difficult errors and mistakes, different options.
Online-spellcheck.com	In some cases, it can't correct the wrong words.	Identifies different versions of incorrect words.

To date, the analysis of text correction systems has been carried out. The disadvantage of these text-correction systems is that they cannot be applied to the Kazakh language, as it is an agglutinative language with a complex morphological and lexical form [1].

To identify errors in the text and develop systems for correcting them in the Kazakh language, it is necessary to focus on the specifics of the language. Kazakh language is an agglutinative language group with the participation of morphological and syntactic rules and is a language with semantics depending on the structure of sentences.

At the same time in the process of automatic editing of Kazakh texts to ensure the correctness of the words in the text, control the correct transfer of word forms, etc. such control levels are included. As a logical result of this work, there is an electronic dictionary of the Kazakh language and systems for checking the accuracy of the Kazakh language texts, which have reached the industrial level of use. However, systems for checking the accuracy specifically for weakly structured texts in the Kazakh language are not publicly available, and even commercial software is difficult to find on the Internet [9].

7.1 Types of errors

Errors in the text are divided into two classes: typographic errors and cognitive errors. Spelling errors - a person does not know how to spell a word. . Typographic errors (real word errors) are related to the keyboard and hand/finger movement where spelling errors happen because of two letters keys' closeness on the keyboard. But this is not limited to the type of error. To date, several types of errors have been identified, especially in semi-structured data [10, 11].

In the course of the study, the following cases were identified when detecting incorrect words from words of the Kazakh language obtained from semi-structured data, i.e. the following types of errors in words are indicated:

- spelling errors (мұхит (muhit)– мухит (mýhit));
- typographic errors (кітап (kitap)– кіап (kiap));

- deliberate distortion of words (алғаaa (al*ǵaaa*), тағда (ta*ǵda*));
- punctuation errors;
- spelling words with Russian alphabet;
- spelling words with Latin alphabet;
- abbreviations, etc.

8 Approach for identifying spelling errors

Spelling error detection is associated with identifying a word as an incorrect word. Spelling checking methods include dictionary search methods that compare words and place them in a language dictionary. Failure to identify a word in the dictionary indicates a spelling error. Also, the most commonly used method is n-grams algorithms. Other methods used to determine to spell include morphological analysis, finite-state transducers, and machine learning algorithms. Hybrid methods are also used to check to spell [12].

There are many algorithms for checking the spelling of text documents. Figure 1 shows the use case diagram showing what types of algorithms are available to detect word errors.

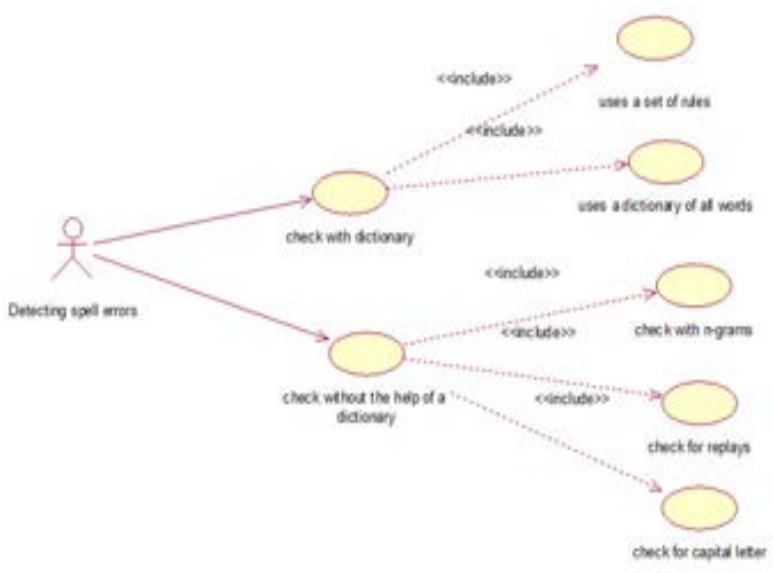

Fig. 1. Use case diagram showing the algorithm for detecting errors in word.

As shown in the diagram, the first method is to check the spelling with a dictionary. Dictionary verification is divided into verification through a dictionary of all words and verification through a dictionary using a set of rules.

Checking with a set of rules. A dictionary that uses a set of rules checks that all words are spelled correctly using language rules.

Checking with the dictionary of all words. A dictionary is a file in .txt format, which contains all the words in the Kazakh language, including all forms of words. The words are arranged in alphabetical order, each word is in a new line. Checking is performed by a regular search for a word in the dictionary. If all the letters of the word match the word in the dictionary, then it is the correct word. If there is no such word, then it is wrong or incorrect [13].

The second method is to check the spelling without the help of a dictionary, which includes checking the capital letter at the beginning of the sentence, checking the repetition of words, and checking with n-grams. Capitalization, that is, each letter after the dot should automatically become an uppercase letter. A repeat test shows that the user wrote two identical words in a row. N-gram analysis is formulated as a method of finding misspelled words in a text array. Instead of comparing each word in the text with a dictionary, only n-grams are checked. If no or rare n-grams are detected, the word is marked as misspelled, otherwise, it is correct. This method does not depend on the language, because it does not require knowledge of the language used [14, 15].

In addition, research has been conducted on models that identify incorrect words in the language. As a result, an approach was developed to identify incorrect words in the Kazakh language. Figure 2 shows an approach of incorrect words.

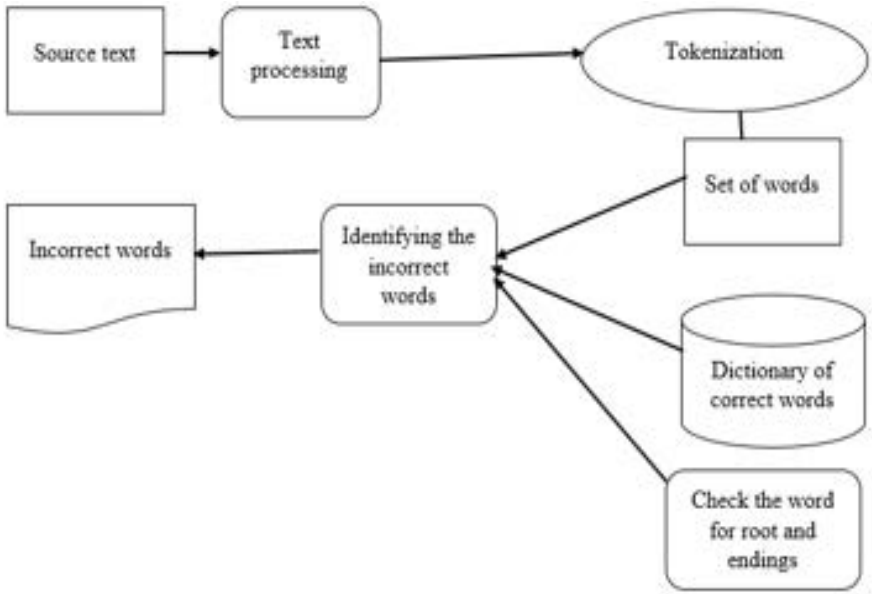

Fig. 2. Scheme of the approach for identifying incorrect words.

As shown in the approach, first the texts are collected from the semi-structured data, then the text is pre-processed, then we divide the text into sentences, sentences into words, and a set of words is formed from those words, respectively. Using the created dictionary of correct words, it is checked that the words in the set of words belong to the words in the dictionary of correct words, ie words are considered correct if they are defined in that dictionary, otherwise, they are considered incorrect. To find incorrect words for the Kazakh language, a stemming algorithm based on the CSE-model developed by Professor Ualsher Tukeyev and his team [10, 16] is used. With the help of the stemming algorithm based on the CSE-model with stems-lexicon [10] the word is separated into stem and endings, if these stem and ending of given word are found in the stem dictionaries and in the complete set of endings of the Kazakh language [17], then the word is correct, otherwise case the word is incorrect.

The proposed approach provides high performance due to the simplicity of the structure. Covers the grammar of the Kazakh language

based on the CSE-model and stem dictionary, and therefore more accurately finds errors in the text. To further improve the quality, an addition to the stem dictionary is proposed.

9. Experiments and results

During the research, a program was developed to assemble the corpus in the Kazakh language. In the Google Colab application, a corpus of 117356 sentences in the Kazakh language was compiled corpus using code written in the Python programming language. We assembled corpus from websites using Python scripts that use the BeautifulSoup and Request libraries in Python. This assembled dataset was analyzed by scripts based on the HTML structure. The corpus includes information on 1200 web pages published in the Kazakh language on the sites akorda.kz and nur.kz. Various statistical calculations were performed on the assembled corpus, the words in the corpus were analyzed and a dictionary of correct words was created, supplementing the data. Table 2 shows the information required to implement the program.

Table 2. Input data for the implementation of the program.

Text in corpus	Number of sentences
akorda.kz	91678
nur.kz	25678

The problem that arose when assembling the corpus was the need to pre-process the corpus data. This is due to the fact that information from sites comes in a variety of formats. It is necessary to consider cases of getting rid of unnecessary symbols, correcting sentences, splitting sentences into words. It will take some time.

For the experiment was used the stemming program described and available in [10, 16], also, were used the language resources of the Kazakh language, namely the dictionary of stems and stop words, complete set of endings available in [17]. The program was tested several times. Table 3 shows the results of the experiment.

Table 3. Experimental results of the developed stemming algorithm for the Kazakh language.

Number of checked words	Number of words for which the stemming algorithm was performed correctly	Number of words for which the stemming algorithm was performed incorrectly	Accuracy, %
1339	1210	229	90
3233	3007	266	93
939	882	57	94
1239	1189	50	96
869	843	26	97
741	726	15	98

After working with the program, the results of the experiment were analyzed. Using the stemming algorithm, the roots of words (stem words) and endings were obtained. After the stemming process for words in the input text, the following errors were encountered in the analysis of the results:

- the stemming algorithm was performed on some words, assuming that there are endings on the basis of the word;

- the stemming algorithm for words was not performed correctly because some endings were not found in the affix file.

The solution of these problems was considered by adding words to the list of stem words, supplementing the set of endings. And another problem to consider is the difficulty of collecting stem words. With each experiment, we added stem words. In the next experiment, there were stems from the previous experiment. Taking them into account, it became necessary to collect universal and unique stem words with each test. It took a long time. To solve this problem, the stem words collected in each experiment were checked against the list of stem words by the program.

In the process of working with the program were collected stem words, stop words, as well as the ending of the Kazakh language. The following table (Table 4) lists the linguistic resources collected in the experiment.

Table 4. Linguistic resources were collected as a result of the experiment.

Linguistic resources	Number of elements of resources
dictionary of stop words	495
dictionary of Kazakh language endings	3140
dictionary of stem words	56398
dictionary of correct words	87376

Also, a dictionary of correct words was used to identify incorrect words in the Kazakh language from semi-structured data. This dictionary was created using the corpus and taking into account the morphological features of the Kazakh language. Experiments were carried out on the method of checking the dictionary and the method of checking the roots and endings of words. A comparative analysis of about 3,000 messages and texts was conducted. The following table shows (Table 5) the results of the experiment.

Table 5. Results of comparative analysis messages.

Objects that were analyzed	Percentage of errors in 400 entries, %	Common types of errors	Percentage of types of errors, %
Instagram	54,6	spelling, typography, spelling in Russian and Latin alphabets, abbreviations	spelling – 16,43, typography – 25,56, spelling in Russian and Latin alphabets – 34,88, abbreviations – 17,12, other – 6
Facebook	39,64	spelling, typography, spelling in the Russian alphabet, abbreviations	spelling – 11,13, typography – 21,87, spelling in Russian alphabet – 43,89, abbreviations – 19,11, other – 4
VKontakte	47,78	spelling in Russian and Latin alphabets, abbreviations, spelling, typography, deliberate distortion	spelling in Russian and Latin alphabets –48,93, abbreviations – 9,57, spelling – 10, typography – 17,86, deliberate distortion – 9,41, other – 4.63
https:// massaget.kz/	32,57	spelling, typography, abbreviations	spelling – 22,5, typography - 53, abbreviations – 20,06, other – 4,44
https://kaz. tengrinews. kz/	31,78	spelling, typography, spelling in the Russian alphabet	spelling – 18,32, typography – 28,65, spelling in the Russian alphabet – 46,91, other – 5,12

Here, at first, data from sites and social networks were collected, both automatically using the program and manually. Then the process of identifying incorrect words was carried out using the developed software product. The percentage of errors was calculated as the number of erroneous words divided by the total number of words in messages of one object. The percentage among the types of errors was calculated in this way. To classify and analyze errors, experts-linguists were involved and these experiments were carried out. As shown in the table 6, posts on sites and social networks were considered for comparative analysis. 600 messages were received from each object, respectively, 3000 messages were received from 5 objects in total. Based on this developed approach from semi-structured data (for example, comments from social networks and news web pages) in the Kazakh language were found incorrect words and defined basic types.

10. Conclusion and future work

The article reviews the literature in accordance with research and analyzes methods and technologies for identifying incorrect words in natural languages. A comparative analysis of text correction systems has been carried out. The disadvantage of these systems is that they cannot be applied to the Kazakh language, since the Kazakh language is an agglutinative language with a complex morphological and lexical form. In this regard, analyzing the methods and systems for detecting errors in the text, models for identifying incorrect words of the Kazakh language were developed, and a dictionary of correct words of the Kazakh language was compiled. The corpus is assembled in the Kazakh language using a program code created taking into account the features and peculiarities of the Kazakh language. An approach was developed for identifying incorrect words from semi-structured data, which worked on the basis of the stemming algorithm with basic vocabulary according to the CSE (Complete Set of Endings) morphological model, and the main types of errors were identified. The experiment was conducted on data from social networks and news web portals. In general, according to the results of the experiment, the accuracy of determining incorrect words was more than 90%.

The main contribution to this work is made by the automated selection of the corpus for the Kazakh language. Development of an approach to identifying incorrect words of the Kazakh language, taking into account the linguistic properties of the language and the creation of linguistic resources and programs for collecting data in semi-structured data.

In the future, methods for correcting errors in the Kazakh language, expanding the accumulated and processed corpus and supplementing information will be developed and implemented.

REFERENCES

Computational processing of the Kazakh language: collection of scientific works (materials) / edited by Rakhimova D.R. –Almaty: Qazaq Universiteti, 2020. -146 p.

Han B., Baldwin T.: Lexical normalisation of short text messages: Makn sens a# twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Volume 1. – Association for Computational Linguistics, 2011. – P. 368-378.

Farra N. et al.: Generalized Character-Level Spelling Error Correction. ACL (2). – 2014. – P. 161-167.

Hladek, Daniel, et al.: Survey of Automatic Spelling Correction. Electronics [Basel], vol. 9, no. 10, 2020, p. 1dj+. Accessed 30 Apr. 2021.

Peter Buneman.: Semistructured data. In: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. May 11-15, 1997, Tucson, Arizona, United States. – P. 117-121.

Brill E., Moore R.C.: An improved error model for noisy channel spelling correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 2000. – P. 286-293.

Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger.: Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. Polibits (40) 2009.

Kaufmann M., Kalita J.: Syntactic normalization of twitter messages. International conference on natural language processing, Kharagpur, India. – 2010.

Rakhimova D.R.: Research of models and methods of semantics of machine translation from Russian into Kazakh language. Dissertation. – Almaty, 2014.

Ualsher Tukeyev, Aliya Turganbayeva.: Universal program of stemming algorithm with stems-lexicon according to the CSE (Complete Set of Endings) morphology model. http://github.com/NLP-KAZNU

Rakhimova D., Assem S. Problems of Semantics of Words of the Ka-

zakh Language in the Information Retrieval. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11684 LNAI, pp. 70–81.

Shaalan K., Aref R., Fahmy A.: An approach for analyzing and correcting spelling errors for non-native Arabic learners. Published 2010, Computer Science, 2010 The 7th International Conference on Informatics and Systems (INFOS).

Taktashkin D.V., Mokrousova Ye.A.: Methods and algorithms for checking the spelling of test documents (Paper in Russian). Electronic scientific & practical journal «Modern scientific researches and innovations». 2017. №5. URL: https://web.snauka.ru/issues/2017/05/72892 - (date of access: 12.04.2021)

Rakesh Kumar, Minu Bala, and Kumar Sourabh.: 2018. A study of spell checking techniques for indian languages. JK Research Journal in Mathematics and Computer Sciences, 1(1).

Rakhimova D., Turganbayeva A. Approach to Extract Keywords and Keyphrases of Text Resources and Documents in the Kazakh Language. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, 12496 LNAI, pp. 719–729.

Tukeyev U.A., Turganbaeva A.O.: Lexicon-free stemming for the Kazakh language. In: Materials of the International Scientific Conference "Computer science and Applied Mathematics" dedicated to the 25th anniversary of the Independence of the Republic of Kazakhstan and the 25th anniversary of the Institute of Information and Computational Technologies, Part II, Almaty, September 21-24, 2016.

Ualsher Tukeyev, Aliya Turganbayeva, Aidana Karibayeva, Dina Amirova, and Balzhan Abduali.: Language_Resources_for_Kazakh_language. https://github.com/NLP-KazNU/Language_Resources_for_Kazakh_language

The Impact of Machine Learning on Organizational Agility in the Context of Industry 4.0 Applications: A Conceptual Study

Dr. Esin Yücel Karamustafa
Altınbaş University , esin.karamustafa@altinbas.edu.tr,
Orcid Number: 0000-0003-1440-1203
Burcu Arsan, PhD(c) Yeditepe University,
burcu.arsan@yeditepe.edu.tr,
Orcid Number: 0000-0002-9757-9391

Abstract

Nowadays, it is becoming more difficult for organizations to survive in increasingly competitive and unstable environmental conditions. In such an environment, it is very vital for organizations to be aware of changes and developments, to keep up with them, and perform quickly. What makes this possible for organizations is the degree of their agility. Industry 4.0 has formed a possibility of digital shift and transformation for organizations and also Industry 4.0 applications enable companies to adapt to transformations in the external environment as well as to respond proactively to these related changes. As a matter of fact, Industry 4.0 applications have a crucial role in the development and formation of organizational agility. Machine learning is one of these applications. Machine Learning, which is a bunch of algorithms, that focuses on creating methods and systems that will enhance performance and uses both statistical and mathematical techniques for this. Machine learning applications warn organizations against the rapidly changing and competitive external environment and aim to enable them to take precautions. However, the number of studies aiming to explain this relationship is limited. This study aims to illustrate the effect and significance of machine learning, which is one of the Industry 4.0 applications, on organizational agil-

ity. In this direction, the concepts of Industry 4.0, machine learning, and organizational agility have been comprehensively examined and the role of machine learning, one of the Industry 4.0 applications, in enhancing organizational agility has been discussed in detail with the arguments explained by considering the relevant literature.

Keywords:
machine learning, organizational agility , industry 4.0

Endüstri 4.0 Uygulamaları Bağlamında Makine Öğreniminin Örgütsel Çeviklik Üzerindeki Etkisi: Kavramsal Bir Çalışma

Öz

Günümüzde örgütlerin, dinamik çevre koşullarında ve giderek artan rekabet ortamında ayakta kalabilmeleri giderek daha zorlu hale gelmektedir.Böyle bir çevrede,örgütlerin değişimler ve gelişmelerden haberdar olmalarıve bu değişim ve gelişime ayak uydurabilmeleri ve hızlı aksiyon almaları büyük önem taşımaktadır. Bunu mümkün kılan ise,örgütsel çevikliklerdir kavramıdır.Endüstri 4.0, örgütler için dijital bir değişim ve dönüşüm olanağı oluşturmuş olup, Endüstri 4.0 uygulamaları da şirketlerin dış çevrelerindeki dönüşümlere uyum sağlamalarının yanı sıra, bu değişikliklere proaktif olarak yanıt vermelerini sağlamaktadır. Nitekim Endüstri 4.0 uygulamaları, örgütsel çevikliğin gelişmesinde ve oluşmasında oldukça önemli bir role sahiptir.Bu uygulamalardan biri de makine öğrenimidir. Makine Öğrenimi, örgüt performansını artıracak yöntem ve sistemler oluşturmaya odaklanan ve bunun için hem istatistiksel hem de matematiksel teknikleri kullanan bir dizi algoritma bütünüdür. Öte yandan, makine öğrenimi uygulamaları, büyük bir hızla değişen rekabetçi dış çevreye karşı kuruluşları uyararak önlem almalarını sağlamayı amaçlamaktadır. Ancak bu ilişkiyi açıklamaya yönelik çalışmaların sayısı sınırlıdır. Bu çalışma, Endüstri 4.0 uygulamalarından

biri olan makine öğrenmesinin organizasyonel çeviklik üzerindeki etkisini ve önemini ortaya koymayı amaçlamaktadır.Bu doğrultuda Endüstri 4.0, makine öğrenimi ve örgütsel çeviklik kavramları kapsamlı bir şekilde incelenmiş ve Endüstri 4.0 uygulamalarından biri olan makine öğreniminin örgütsel çevikliği geliştirmedeki rolü, bu kavramlar dikkate alınarak kapsamlı bir literatür taraması gerçekleştirilmiştir.

Anahtar Kelimeler;
makine öğrenimi, örgütsel çeviklik, endüstri 4.0

1. Introduction

The information and digitalization age leads companies to integrate and create digital solutions in their internal structures, procedures and relations with their stakeholders in accordance with the necessities of the age (Nagy *et al.* 2018). Being out of date is a situation that threatens the presence of companies.Hence, the 4th industrial revolution has forced today's organizations and the business world to undergo a crucial transformation.

Industry 4.0 depends on the idea of production systems dominated by digitalization and automation, integrating physical and cyber systems, where these systems interact through a network(Xu et al., 2018) While this revolution deeply affects organizations in every aspect, it stands out as the main tool of competitiveness in the digital age with the advantages of its applications.Industry 4.0 applications make it possible to simplify operational processes, decrease costs, customized production, improve interaction inside and outside the organization, and be rapid in production procedures and decision-making (Mohamed, 2018).

Today, adaptation to the external environment, which is changing rapidly with the effect of digital transformation, is one of the most important competitive advantages of organizations.In this context, businesses should make major changes when necessary by reviewing all their processes and organizational structures in order to adapt to new conditions (Atlı, 2013). Agility requires severe structural and systemic changes within the company.According to Sherehiy et al. (2007), although agility

includes speed and flexibility, it is a much broader concept than themIn this context, for an agile structure; radical change, innovation and innovation practices are inevitable elements.

In this study, initially, the concepts of Industry 4.0 and machine learning were examined in detail, and then the concept of organizational agility was explained.Afterward, the role of Industry 4.0 applications in the development of organizational agility is discussed with the necessary theoretical background and literature knowledge. Finally, this conceptual study is briefly evaluated and suggestions for future studies are presented.

2. LITERATURE REVIEW

2.1. Industry 4.0 Applications

Until today, there have been three major industrial revolutions that have seriously changed our world (Can and Kıymaz, 2016).In the industrial sense, the First Industrial Revolution (Industry 1.0), which formed with the invention of the steam engine in the 18th century and, strived to improve and speed up the production process, was followed by the Second Industrial Revolution (Industry 2.0), which appeared as a shift to mass production at the beginning of the 20th century and paved the way for benefiting from electrical energy.

Afterward, the 3rd Industrial Revolution (Industry 3.0) emerged, where production techniques and systems ceased to be analog and, digital procedures took place in the industry (Sayer and Ulker, 2014). Therefore, after first three industrial revolutions they brought mechanization, information technology (IT) and electricity, to human production (Yang, 2017). All three revolutions that took place were aimed at improving productivity in production processes. (Can and Kıymaz, 2016).

However, companies faced very severe complications due to the internal and external environmental, social, economic, and technological improvements experienced at that time, and improving productivity alone did not bring businesses to the forefront of global competition. Thus, organizations needed to work interdisciplinary and, the Fourth Industrial

Revolution (Industry 4.0), in which all objects interact with each other through the internet tools, has emerged (Yang, 2017).

The new approach, called Industry 4.0, includes a structure and system that will entirely alter the production and consumption linkages. Furthermore, it illustrates the production systems and techniques that instantly adapt to the changing requirements of the clients, and on the other hand, the automation procedures that are in constant coordination and transmission with each other (Sinan, 2016),and it encourages and facilitates close collaboration between various disciplines in product development (Herter and Ovtcharova, 2016).

The concept of Industry 4.0, which can be outlined as the computerized automation of production processes, first emerged with a project that the German government included among its high-tech investment strategies. Mrugalska and Wyrwicka (2017) describe the concept of industry 4.0 as "the integration of complicated physical devices and machines with networked sensors and software utilized to better predict, monitor, and plan organizational and societal outcomes."

The concept of Industry 4.0 is seen as an crucial and the most fundamental strategy to survive in a challenging competitive environment in the future. Mentioned strategies include the configuration and implementation of competitive services and products, as well as flexible and adaptable logistics services and production methods.Organizations are underlining the industry 4.0 practices to overcome challenges such as; increased product customization, increased resource efficiency, and shortened time to market (Rennug *et al*, 2016).

Despite the advancing complexness of the Industry 4.0 scenario, it also has the advantages summarized below: Increasing competition and adaptablity arising from the dynamic nature of business procedures (time, price, quality, risk, and environmental-friendliness), Eliminating malfunctions in the demand chain, Optimizing decision-making with real-time end-to-end visibility, Provide raised resource productivity (which delivers the highest output from a given volume of resources) and efficiency (uses the lowest possible amount of resources to achieve a given output), Building value opportunities (innovation, new forms of

work, advancement opportunities for start-ups and SMEs), Decreasing personal and energy costs (Mrugalska and Wyrwicka, 2017).

Through Industry 4.0, the systems will create continuous effective and sufficient production. Possible faults in the processes will be detected and the deficincies will be corrected as soon as possible. Efficiency ratio is relatively high and production will be concentrated on quality rather than quantity. Through to flexible and adaptable production, methods will be altered, and services and products will be effortlessly modified. With the advancement of technology, expenditures will be minimized, and large plants will be replaced by small producers and equipment. Plants and houses will be capable to be observed, monitored and processed remotely.

2.1.1 Machine Learning

The concept of Machine learning is an application of artificial intelligence and computer science that progressively raises its accuracy by concentrating on the usage of algorithms and data to imitate and mimic the way human beings learn. IBM has a rich history of machine learning. It is known that the term "machine learning" was first used by one of the IBM insiders, Arthur Samuel, in his research on the game of checkers. Expert checkers player Robert Nealey played the match on an IBM 7094 computer in 1962 and failed against the computer.

Considering what is possible today, this achievement may seem almost trivial; however, it is seen as an important turning point in the field of artificial intelligence.In the next few decades, technical and digital advances in storage and processing power will allow some innovative developments we know and love today, such as Netflix's recommendation engine or self-driving vehicles.

Machine Learning (ML), used by giants technology based organizations such as Google, Amazon, IBM, and Microsoft, is a subgroup of artificial intelligence and operate it to reach the unknown by inferring from the consumed data. Machine Learning, which is a collection of algorithms, focuses on building systems and procedures that will boost the performance and uses both mathematical and statistical methods for

this.Machine Learning is involved in face recognition, online shopping, spam filtering, use of social media, network security threat detection, fraud detection, document classification, or contacting banks.

Machine Learning algorithms are generally divided into two main branches: Supervised Machine Learning and Unsupervised Machine Learning. Supervised Machine Learning algorithms learn what results to reach with labeled datasets with defined outputThe related training is given by the data specialist. These algorithms are used more frequently compared to Unsupervised Machine Learning algorithms.

Unsupervised Machine Learning algorithms learn using unlabeled data with no defined output.In these algorithms, the data expert does not act as a guide. Since there are no tags, algorithms are expected to behave like an explorer.

Machine Learning is used in many fields todayThe number of organizations aiming to reach the right result by using these algorithms for different purposes in business life is increasing day by day. Some of the points that the business world benefits from Machine Learning:

Understanding and Retaining Customers; Many organizations from different sectors, especially e-commerce companies, use Machine Learning algorithms to identify their target customers, understand them correctly, and not lose them.By evaluating the customer data, it becomes possible to predict the revenue within a certain period in the future, and this gives direction to the marketing efforts.

Determining Customer Loss; Machine Learning algorithms are also used to reveal the possible loss of existing customers and the reasons for this loss. Getting help from algorithms is very helpful because retaining loyal customers is inexpensive compared to acquiring new customers. The results obtained also shape the marketing efforts (such as discounts, special offers) to prevent the loss of customers (Kaynar, 2017).

Determining the Premise in CRM Systems; CRM, namely the Customer Relationship Management system, is the approach responsible for the management of the interaction of businesses with their current and potential customers (Helfert, 2003).Machine Learning algorithms are used in CRMs to analyze a subject and get ahead of the others. For

example, when the word "malfunction" is mentioned in an e-mail, this e-mail is given priority to be answered (Farquad, 2014).

Making Dynamic Pricing; The business world needs to be able to make dynamic pricing since the options that the consumer segment can evaluate through different channels to meet any need are much more today than in the past. In this direction, dynamic pricing is made by using Machine Learning algorithms to determine demands, interest levels and attitudes towards campaignsDemand pricing (dynamic pricing), which is made according to the results achieved by using large amounts of data, taking into account the variability according to the situations, also contributes to the rapid adaptation of the related business to the market (Gür Ali and Arıtürk, 2014).

Customer Classification; The intuitions used for marketing strategies to achieve the purpose of the business have left their place to Machine Learning algorithms today. As the data stored increases and the algorithms become more complex, the most accurate result is approached. Identifying the customer segment contributes to personalized marketing and thus increases sales (Farquad, 2014).

In Human Resources; Based on the fact that the success of businesses is based on employee success, Human Resources systems determine the characteristics of an effective employee by using Machine Learning algorithms in new employee recruitment and reflect the data obtained to the recruitment criteria.

Estimating; It is Machine Learning algorithms that pave the way for making predictions by making use of more realistic existing data, not experience while making business decisions. These algorithms, which benefit from rich analytics, are very functional because they present the risks to be encountered in the future today. It increases the uptime of assets, improves their performance, makes their lifespan longer. On the other hand, it also maximizes productivity. Also, predicting when which parts of the machines installed in a factory will fail can be given as an example (Durga, *et al.* 2006).

As it is understood, it is possible to benefit from Machine Learning for different purposes in business life.Although the examples given here

can be multiplied, the main thing to know is that despite all the benefits of Machine Learning, it is not enchanted. In other words, it is not possible to reach the right result in a short time with a single data flow. Time, different data and algorithms are needed to create a model.

2.2 Organizational Agility

The concept of agility was put forward in the early 1990s and was seen as a solution for the company's survival in changing environmental conditions (Nafei, 2016). Organizational agility enables the company to perceive the changes in the environment and respond quickly to these changes, thus adapting to the changing external environment and staying alive.Environmental changes; It covers all changes that concern the company, such as changes in competitors' activities, changes in consumer behavior and preferences, and legal and economic changes (Overby *et al.,* 2006).

Many definitions of organizational agility have been made by researchers in the literature.Organizational agility is the firm's ability to respond quickly to these transformations by utilizing the possibilities and options that will emerge in its environment to meet consumer demands (Akkaya ve Tabak, 2018).

Ravichandran (2018) describes organizational agility as the organization's capacity to respond and react quickly to changes. Nafei (2016) defined agility as the ability to take action efficiently and think quickly. Kanten et al. (2017) explained it as being able to use the resources of the organization immediately when perceiving the opportunities and threats in the market and taking action accordingly.

According to Zitkiene and Deksnys (2018), organizational agility is the capacity and the ability of the organization to notice the transformations in bothexternal an internal environment and to react quickly to these changes by using existing resources, and this process increases the competitiveness of the organization.

As a more broader definition, organizational agility is defined as "being sensitive to changes in the external environment, gaining competitive advantage with new technologies, adopting governance principles

such as cooperation, communication, and transparency, creating a lean and flexible organizational culture in organizations, placing the customer at the center of their work and their desires and needs."(İnanır, 2020).

Organizations with high level of organizational agility; can efficiently detect and predict environmental transformations, decrease costs by reducing unneeded activities and improve investment opportunities, and related organizations are aware of the significance of innovation, can quickly merge resources and capabilities to maintain and increase their competencies (Darvishmotevali and Tajeddini, 2020).

There are many studies in the literature about the characteristics of agile organizations.Sherehiy et al. (2007) analyzed past research and summarized the main characteristics that determine organizational agility in their study.In their work, (1) flexibility, (2) responsiveness, (3) a culture of change, (4) speed, (5) integration and low complexity, (6) quality and customized products, and (7) mobilizing core competencies are described as the key factors that determine organizational agility.

The concept of organizational agility, which was put forward in the early 1990s, is the ability of organizations to use their resources efficiently and effectively to create value by considering both internal and external conditions (Teece et al., 2016), thus responding quickly to changes in the environment. (Zitkiene and Deksnys, 2018).

3. Conclusion

With the 3rd Industrial Revolution, the use of information systems and technologies by organizations started and this situation made significant contributions to the agility of organizations.As a matter of fact, there are many studies conducted in the international literature on this subject and these studies show that information technologies increase organizational agility (Lu and Ramamurthy, 2011; Chakravarty et al 2013; Mao et al.2015; Ravichandran, 2018).

With Industry 4.0, autonomous systems and digitalization have begun to enter organizations in every sector.These systems provide advantages such as; speed, flexibility, quality, minimum use of resources, and continuous interaction within and outside the organization.Therefore,

it offers advantages to organizations far beyond the benefits of classical information technologies and systems and contributes to the agility of organizations in many ways.

In constantly changing, competitive and dynamic market conditions, agility is much more important for organizations.Because under these conditions, agile organizations are much more successful than others (Mao et al., 2015). The reason for this is that organizational success in dynamic environmental conditions depends on speed, responsiveness and adaptability. In this direction, organizations have to focus on constantly improving themselves in order to increase their agility, be able to respond to changes more rapidly in the actions they develop, respond to changes more easily, and adapt to dynamic conditions more easily.

In short, organizations must be dynamic in their structures, decisions, processes and activities.Organizations can achieve this by being innovative and interacting with the environment and using their existing resources efficiently for their purposes (Harraf et al., 2015).For this reason, the technologies and applications they use can provide businesses with significant advantages in this direction.

Businesses that adapt to Industry 4.0 and its components and use it within the organization are the enterprises that can benefit most from this advantage in today's conditions.Industry 4.0 applications, in terms of their structure, interact more continuously and more effectively (Chen *et al*, 2014), resource use much more efficient (Vaidya *et al*., 2018) production and activities much faster (Serinikli, 2018), and, naturally it makes organizations much more dynamic and agile.Industry 4.0 has emerged as a strategic initiative based on the digitalization of production systems. (Rojko, 2017).

It is these three integrations that enable digital transformation in the company. For this reason, the use of Industry 4.0 offers a system in which all elements are integrated with each other, and the organization shows agile features both in its internal structure and in its activities and relations with the external environmentAs a matter of fact, Industry 4.0 applications provide companies with flexibility, efficiency, speed, agility, and quality increase (Serinikli, 2018).

Industry 4.0 can be considered a relatively new concept and the relevant literature is still insufficient as it is still developing.For this reason, it is thought that both theoretical and empirical research about Industry 4.0 and the concept of machine learning will contribute to the development of the related literature.

REFERENCESS

Akkaya, B., ve Tabak, A. (2018), Örgütsel Çeviklik Ölçeğinin Türkçeye Uyarlanması: Geçerlik ve Güvenirlik Çalışması, İş ve İnsan Dergisi, 5(2), 185-206.

Atlı, D. (2013). Yetenek Yönetimi. 2.Baskı. İstanbul: Crea Yayıncılık.

Can, A.V.andKıymaz, M. (2016). "Bilişim teknolojilerinin perakende mağazacılık sektörüne yansımaları: muhasebe departmanlarında endüstri 4.0 etkisi", Sosyal Bilimler Enstitüsü Dergisi, CİEP Özel Sayısı, pp. 107-117.

Chakravarty, A., Grewal, R., ve Sambamurthy, V. (2013), Information Technology Competencies, Organizational Agility, and Firm Performance: Enabling and Facilitating Roles, Information Systems Research, 24(4), 976-997.

Chen, Y., Wang, Y., Nevo, S., Jin, J., Wang, L., ve Chow, W. S. (2014), IT Capability and Organizational Performance: The Roles of Business Process Agility and Environmental Factors, European Journal of Information Systems, 23(3), 326-342.

Darvishmotevali, M., ve Tajeddini, K. (2020), Understanding Organizational Agility, Evidence from the Hotel Industry in Iran, K. Tajeddini, V. Ratten ve T. Merkle içinde, Tourism, Hospitality and Digital Transformation: Strategic Management Aspects- Innovation and Technology Horizons (s. 73-87), New York: Routledge.

Durga L. Shrestha, Dimitri P. Solomatine, (2006). Machine learning approaches for estimation of prediction interval for the model output,

Neural Networks, Volume 19, Issue 2, 2006, Pages 225-235,

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing, 19, 31–40.

Gür Ali, Ö.,& Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systemwith Applications, 41(17), 7889–7903.

Harraf, A., Wanasika, I., Tate, K., ve Talbott, K. (2015), Organizational Agility, Journal of Applied Business Research, 31(2), 675-686.

Helfert, M.,& Heinrich, B. (2003). Analyzing Data Quality Investments in CRM: A Model-Based Approach (pp. 80–95). Presented at the Proceedings of the Eighth International Conference on Information Quality, Massachusetts: ACM Digital Library.

Herter, J. and Ovtcharova, J. (2016). "A model based visualization framework for cross discipline collaboration in industry 4.0 scenarios", Procedia CIRP, vol. 57, pp. 398-403, 2016.

İnanır, A. (2020), Örgütsel Çeviklik, M. Sağır içinde, Modern İşletmecilikte Yönetsel Konular (s. 71- 80), Konya: Eğitim Yayınevi.

Kanten, P., Kanten, S., Keceli, M., ve Zaimoglu, Z. (2017), The Antecedents of Organizational Agility: Organizational Structure, Dynamic Capabilities and Customer Orientation, PressAcademia Procedia, 3(1), 697-706.

Kaynar, O., Tuna, M. F., Görmez, Y., & Deveci, M. A. (2017). Makine öğrenmesi yöntemleriyle müşteri kaybı analizi. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 18(1), 1-14.

Lu, Y., ve Ramamurthy, K. R. (2011), Understanding the Link Between Information Technology Capability and Organizational Agility: An Empirical Examination, MIS Quarterly, 35(4), 931-954.

Makers Turkey Report, (2021) Makine Öğrenmesi Nedir, İş Hayatında Nasıl Faydalanır?

Mao, H., Liu, S., ve Zhang, J. (2015). How the Effects of IT and Knowledge Capability on Organizational Agility are Contingent on Environmental Uncertainty and Information Intensity, Information Development, 31(4), 358-382.

Mohamed, M. (2018), Challenges and Benefits of Industry 4.0: An Overview, International Journal of Supply and Operations Management, 5(3), 256- 265.

Mrugalska, B. and Wyrwicka,M.K. (2017) "Towards lean production in industry 4.0.", Procedia Engineering, vol. 182, pp. 466- 473.

Nafei, W. A. (2016), Organizational Agility: The Key to Organizational Success, International Journal of Business and Management, 11(5), 296-309.

Nagy, J., Oláh, J., Erdei, E., Máté, D., ve Popp, J. (2018), The Role and

Impact of Industry 4.0 and the Internet of Things on the Business Strategy of the Value Chain—The Case of Hungary, Sustainability, 10(10), 3491.

Overby, E., Bharadwaj, A., ve Sambamurthy, V. (2006), Enterprise Agility and the Enabling Role of Information Technology, European Journal of Information Systems, 15(2), 120-131.

Ravichandran, T. (2018), Exploring the Relationships Between IT Competence, Innovation Capacity and Organizational Agility, The Journal of Strategic Information Systems, 27(1), 22-42.

Rennung,F., Luminosu, C.T. and Draghici,A.(2016) "Service provision in the framework of industry 4.0.", ProcediaSocial and Behavioral Sciences, vol. 221, pp. 372-377.

Rojko, A. (2017), Industry 4.0 Concept: Background and Overview, International Journal of Interactive Mobile Technologies, 11(5), 77-90.

Sayer, S.andÜlker,A. (2014) "Ürün yaşam döngüsü yönetimi", Mühendis ve Makina, cilt. 55, no. 657, pp. 65-72.

Serinikli, N. (2018), Endüstri 4.0'ın Özel, Kamu ve Kooperatif Sektörlerine Etkisi, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 23, 1607-1621.

Sherehiy, B., Karwowski, W., ve Layer, J. K. (2007), A Review of Enterprise Agility: Concepts, Frameworks, and Attributes, International Journal of Industrial Ergonomics, 37(5), 445-460.

Sherehiy, B. ve W. Karwowski. (2014). The Relationship Between Work Organization and Workforce Agility in Small Manufacturing Enterprises. International Journal of Industrial Ergonomics. 44. (3), s: 466-473.

Sinan A. (2016)., "Üretim için yeni bir izlek: sanayi 4.0", Journal of Life Economics, no. 8, pp. 19-30

Teece, D., Peteraf, M., ve Leih, S. (2016), Dynamic Capabilities and Organizational Agility: Risk, Uncertainty, and Strategy in the Innovation Economy, California Management Review, 58(4), 13-35.

Vaidya, S., Ambad, P., ve Bhosle, S. (2018), Industry 4.0–A Glimpse, Procedia Manufacturing, 20, 233-238.

Xu, L. D., Xu, E. L., ve Li, L. (2018), Industry 4.0: State of the Art and Future Trends, International Journal of Production Research, 56(8),

2941-2962.

Yang, L.U. (2017) "Industry 4.0: a survey on technologies, applications and open research issues. Journal of Industrial Information Integration, in press.

Zitkiene R., ve Deksnys, M. (2018), Organizational Agility Conceptual Model, Montenegrin Journal of Economics, 14(2), 115-129.

Study and development of a post-editing model based on machine learning for English-Kazakh and Russian-Kazakh translation

Diana Rakhimova, Kamila Sagat, Kamila Zhakypbaeva
Al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru, kamilasagat@gmail.com,
zhakypbaeva18@gmail.com

Abstract.

In this paper mentions the importance of machine translation and post-edited machine translation in nowaday, and also describes the use of NMT for post-processing of output data obtained using existing machine translation systems. We investigated the post-editing model for Russian-Kazakh and English-Kazakh translation. The study revealed common mistakes when translating from English to Kazakh and Russian to Kazakh and compared the results of the online translation systems sozdik.kz, yandex, Google translate, webtran.ru. With the help of the obtained results, we identified the comparative characteristics of online machine translation systems when translating from English to Kazakh and Russian to Kazakh, in addition to this, were given the main reasons for errors when translating from English to Kazakh and Russian to Kazakh.

In the course of the research, the post-editng machine translation model and the full post-editng machine translation architecture were built, the materials necessary for the implementation of machine learning were collected, and the results of the experiment were presented.

Keywords:

Neural Machine Translation (NMT), Machine translation, Kazakh language, post-editing machine translation (PEMT), machine learning.

11. Introduction

In recent years, the amount of data required for translation has been growing rapidly, and this factor has contributed to the growing demand for machine translation. As demand grows, so does the quality and speed of machine translation. For example, in 2010 the Executive Directorate of the European Translation Commission decided to develop a machine translation system that would be available not only to professional translators, but also to independent publishers who provide unprocessed translations. In addition, the integration of machine translation with various translation systems has allowed it to become more widespread in a professional context. Of course, it seems impossible to completely replace translators with machine translation, but working with a machine has become part of the work of professional translators today [2].

The post-edited text must be clear and concise, as well as grammatically and stylistically appropriate. With this in mind, the International Standard (ISO / CD 18587: 2017) [24] developed requirements for the process of post-editing, results and competencies of translators engaged in post-editing [3].

Nowadays, it is even more important to translate data from one language to another. About three-quarters (3/4) of Internet users use free translation tools due to the availability and integration of machine translation solutions. More than 90% of English speakers use machine translation software to translate accessible English websites [4].

Despite the significant development of MT systems, the output material often requires human revision. This process is called post-processing machine translation (PEMT, post-editing) and means making corrections to the text of the machine translation in accordance with the pre-established requirements. Unlike post-editorial editing, in the first case the source text is a machine translation, and in the second case it is a human translation. The text that has passed the post-processing stage can be further submitted to the editor to correct stylistic, grammatical and lexical errors and the functional orientation of the text to the target reader. The editing specialist should be more qualified than the translator or editor. Statistics also show that interaction with MT becomes one of

the most important components of the work of a professional translator [5].

12. Motivation and related work

Machine translation is constantly improving, faster, cheaper and more accurate. However, we still cannot compare the results of machine translation with the results of human translation. Post-editing machine translation helps to combine the best of both parties (machine translation and translator): the speed, ability and skill of trained linguists can help to process large volumes of text quickly and efficiently [7].

As a result of various experiments, scientists have come to the conclusion that post-editing takes less time than translating the whole text. However, time savings vary depending on the type of text [8]. The results of comparisons and identification of previous experiments (O 'Brien 2007; Guerberof 2009; Plitt and Masselot 2010 [22]) and recent experiments (Daems et al. 2017; Carl et al. 2011; Screen 2017 [23]) refute the notion that post-editing takes less time. However, it is important to be fair in making such statements. This is due to the fact that some experiments involved very strong translators, and some involved people with little experience in the field of translation [8]. The existence of such two-sided views is due to the linguistic differences (styles) of the texts. Often, scientific, formal, and literary texts take significantly more time than translating / post-editing a simple oral text. This is because it may take time to search for specific terms and words in specific online resources [8].

Numerous studies and experiments show that the use of the NMT method is now more successful than other methods of machine translation. Neural networks allow more accurate translation of text by context [6]. In addition, the use of special textbooks containing various scientific terms and professional words in the field related to the field in which the text to be translated in the implementation of machine learning helps to improve the quality of machine translation [6].

13. Analysis of the translation of the Kazakh language in online machine translation systems

The Kazakh language is an agglutinative language with a complex nominative (morphological and syntactic) participation of polysynthesis. The various machine translation systems still cannot translate completely correctly. The analysis compared various parameters - lexical, grammatical and stylistic correctness of the translation, and also revealed spelling errors, narrowing of the context and distortion of meaning. In order to assess the quality of the translation and the amount of required post-editing, an experiment was carried out to compare the translations of the most popular electronic online MT systems (translators). A comparative analysis of the translation of complex sentences was performed to identify errors that occur during machine translation. Table 1 shows the results of the translation from Russian to Kazakh. Table 2 shows the results of the translation from English to Kazakh. A total of 4 machine translation systems were selected to analyze the results [10,11,12,13].

Table 1. Results of the translation from Russian to Kazakh.

Source text	*Names of machine translation systems and translation results*			
	Google translation	*Yandex translation*	*Sozdik.kz*	*Webtran.ru*
Ударение — выделение каким-либо акустическим средством одной из сегментных единиц языка — слога, слова или синтагмы [Accent is the selection of one of the segmental units of a language-a syllable, word, or syntagma — by some acoustic means]	Стресс - бұл кез-келген акустика-лық құралдар арқылы тілдің сегменттік бірліктерінің бірін-буынды, сөзді немесесинтагманы таңдау [Stress-bul kez-kelgen akýstıka-lyq quraldar arqyly tildiń segmenttik birlikteriniń birin-býyndy, sózdi nemesesıntagmany tańdaý]	Екпін-тілдің сегменттік бірліктері-нің бірін — буынды, сөзді немесе синтагманы акустика-лық құралдармен оқшаулау [Ekpin-tildiń segmenttik birlikteri-niń birin — býyndy, sózdi nemese sıntagmany akýstıka-lyq quraldarmen oqshaýlaý]	Екпін — бөлу қандай акустическим құралы бірі сегментных бірлік тілі — буынды сөздер немесе синтагмы [Ekpin — bólý qandaı akýstıcheskım quraly biri segmentnyh birlik tili — býyndy sózder nemese sıntagmy	Стресс - бұл кез-келген акустикалық құралдар арқылы тілдің сегменттік бірліктерінің бірін - буынды, сөзді немесе синтагманы таңдау [Stress - bul kez-kelgen akýstıkalyq quraldar arqyly tildiń segmenttik birlikteriniń birin - býyndy, sózdi nemese sıntagmany tańdaý]
Я им как младшая сестра [I'm like a little sister to them]	Мен олар үшін кішкентай әпкем сияқтымын [Men olar úshin kishkentaı ápkem sıaqtymyn]	Мен оларға kishi қарындас ретінде [Men olarǵa kishi qaryndas retinde]	мен оларға қарындасы [MEN olarǵa qaryndasy]	Мен олар үшін кішкентай әпкем сияқтымын [Men olar úshin kishkentaı ápkem sıaqtymyn]
Лист бумаги по утрам, на квартиру носят нам [A sheet of paper in the morning, they bring us to the apartment]	Таңертең бір парақ, олар бізді пәтерге апарады [Tańerteń bir paraq, Olar bizdi páterge aparady]	Таңертең қағаз парағы, біз пәтерге киеміз [Tańerteń qaǵaz paraǵy, biz páterge kıemiz]	Бір парақ қағаз таңертең, пәтерге тағады бізге [Bir paraq qaǵaz tańerteń, páterge taǵady bizge]	Таңертең бір парақ, олар бізді пәтерге әкеледі [Tańerteń bir paraq, olar bizdi páterge ákeledi]

Table 2. Results of the translation from English to Kazakh.

Source text	*Names of machine translation systems and translation results*		
	Google translation	*Yandex translation*	*Sozdik.kz*
When I woke up, I felt pain all over my body	Мен оянған кезде бүкіл денем ауырды [Men oıanǵan kezde búkil denem aýyrdy]	Оянғанда бүкіл денем ауырды [Men oıanǵanda búkil denem aýyrdy]	Менің оянған кезде бүкіл денем ауырды [Meniń oıanǵan kezde búkil denem aýyrdy]
Arman and Bakhyt are playing in the field	Далада Арман мен Бақыт ойнап жүр [Dalada Arman men Baqyt oınap júr]	Арман мен Бақыт алаңда ойнайды. [Arman men Baqyt alańda oınaıdy]	Арман мен Бахыт ойнады далада [Arman men Bahyt oınady dalada]
The old men sat gloomy from morning to evening without saying anything to each other	Қарттар таңертеңнен кешке дейін бір-біріне ештеңе айтпастан қабағын түйіп отырды [Qarttar tańerteńnen keshke deıin bir-birine eshteńe aıtpastan qabaǵyn túıip otyrdy]	Қарттар таңертеңнен кешке дейін бір-біріне ештеңе айтпастан күңгірт отырды [Qarttar tańerteńnen keshke deıin bir-birine eshteńe aıtpastan kúńgirt otyrdy]	Қарттар таңертеңнен кешке дейін бір-біріне ештеңе айтпастан қабағын түйді [Qarttar tańerteńnen keshke deıin bir-birine eshteńe aıtpastan qabaǵyn túıdi]

Machine translation systems have advantages and disadvantages in translation. Analyzing the work of each of them, we identified possible errors in the translation from Russian to Kazakh and from English to Kazakh [14]. They are shown in Table 3:

Table 3. Comparative characteristics of online machine translation systems when translating from Russian to Kazakh and from English to Kazakh.

Machine translation systems	Disadvantage	Advantage
Google translation	In some cases, there are some inconsistencies in large sentences	High quality translation of complex scientific and technical texts with minor problems of complex
Yandex translation	When translating large texts, parts of a sentence are not broken or translated	Translates very large sentences, complex texts with high quality
Sozdik.kz	Translation of large scientific texts does not give good results	Good translation results for words and short sentences
Webtran.ru	During translation, some words remain unchanged. In large sentences, the connection is incorrect	Good translation results for short sentences and phrases

In addition, the characteristic of the Kazakh language significantly affect the occurrence of errors in translation. Here they are:

- proximity of the lexical structure;
- The law of harmony;
- agglutination-a series of applications;
- no category;
- Absence of auxiliary words (prepositions);
- special word order [19].

14. Model of post-editing of the Kazakh language in the MTS

Fig. 1 shows the complete machine translation - based full post-editing architec-ture. Full post-editing is the process of processing the results of machine translation and producing meaningful translations. It must produce accurate translations that do not have stylistic inconsistencies and consistently use correct and approved terminology, avoiding any grammatical errors.

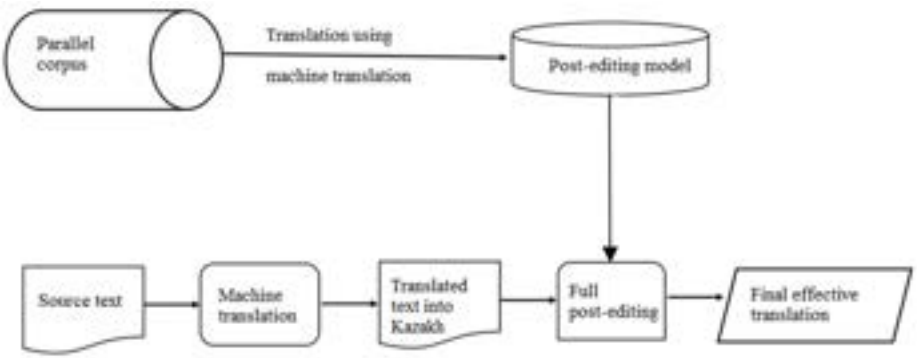

Fig. 1. Full post-editing architecture based on machine translation.

The full post-editing architecture, based on machine translation, describes in detail how a step-by-step translation from English or Russian into Kazakh using post-editing processing is done.

15. Practical results

On the website www.akorda.kz, a parallel corpus of 80022 sentences in Russian and English was assembled using code written in the Python programming language. The assembled corpus was translated into the Kazakh language (into the Cyrillic alphabet of the Kazakh language) applying Google Translate. However, due to the fact that the Kazakh language belongs to the group of agglutinative languages, the translation quality was low. That is, the sentences are incomplete, the word order is incorrect, etc. This translation was summarized in the document «kaz». A high-quality and error-free translation of this document was created. This revised translation is summarized in the «kazedit» document. We processed the corpus, tokenized the corpus, separated the punctuation from the words.

1. train.kazedit, train.kaz - list of sentences for training;
2. tst2021.kazedit, tst2021.kaz - list of sentences submitted for validation;
3. tst2022.kazedit, tst2022.kaz - list of sentences submitted for testing;
4. vocab.kazedit, vocab.kaz- list of dictionaries.

For the Russian-Kazakh language, 97.35% of the total corpus was allocated for training, 1,34% - for validation, 1.31% - for testing. The number of words and sentences in these documents is shown in Table 4.

Table 4. The amount of information in the documents for machine learning for the Russian-Kazakh language.

Document name	*kaz*	*kazedit*
train	112728- sentences	112728- sentences
tst2021	1556- sentences	1556- sentences
tst2022	1501- sentences	1501- sentences
vocab	71523- words	70899- words

115785 For the English-Kazakh language, 96.24% of the total corpus was allocated for training, 2.1% - for validation, 1.66% - for testing. The number of words and sentences in these documents is shown in Table 5.

Table 5. The amount of information in the documents for machine learning for English-Kazakh language.

Document name	*kaz*	*kazedit*
train	84963 - sentences	84963 - sentences
tst2021	1860 - sentences	1860 - sentences
tst2022	1452 - sentences	1452 - sentences
vocab	69103- words	71221- words

88275 We are training this document using the seq2seq model for NMT (Neural Machine Translation).

The Seq2seq model is a model that takes sentences, words as input and returns another sequence of elements. The following example is given for the studied model:

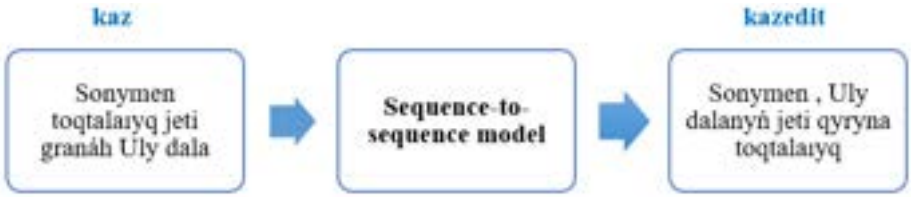

Fig. 2. Example of a sequence-to-sequence model.

The creation of the NMT model is associated with the attention mechanism, which helps to remember long sentences. The current target latent state is compared with all baseline states to obtain attention weights. Based on the attention weights, we calculate the context vector as a weighted average of the initial states and combine the context vector with the current target hidden state to get the final attention vector. This vector is supplied as input for the next time step [9].

The analysis of the results of improving machine translation was carried out using the BLEU metric. BLEU (bilingual evaluation understudy) is an algorithm for evaluating the quality of text translated from one natural language into another [15].

The main purpose of post-editing is to improve translation performance. When comparing the original translation, the results of the BLEU metric are as follows:

Table 6. Results of BLEU metrics.

	Google translation	*Yandex translation*	*Post-editing*
Russian-Kazakh	8.3	6.6	10.7
English-Kazakh	7.7	7.1	9.4

The main goal of post-editing is to improve the indicator of this metric. The BLEU metric value when comparing the original enclosures is 6.6. After the training stage for post-editing into the Kazakh language, a BLEU metric value of 10.7 was obtained. Due to this, the result of translating texts has been improved.

An example of some of the sentences that appear as a result of testing is as follows:

Table 7. Results obtained after machine learning.

kaz	*kazedit*
Жыл қорытындысы бойынша мемлекет басшысы бұқаралық ақпарат құралдары өкілдерімен елді [Jyl qorytyndysy boıynsha memleket basshysy buqaralyq aqparat quraldary ókilderimen eldi]	Мемлекет басшысы жыл қорытындысы бойынша еліміздің бұқаралық ақпарат құралдары өкілдерімен кездесті [Memleket basshysy jyl qorytyndysy boıynsha elimizdiń buqaralyq aqparat quraldary ókilderimen kezdesti]
Кездесуде Нұрсұлтан Назарбаев “жыл адамы” мемлекеттік күзет Қызметінің бастығы осы мекеменің Ардақ Ашимбековича [Kezdesýde Nursultan Nazarbaev “jyl adamy” memlekettik kúzet Qyzmetiniń bastyǵy osy mekemeniń Ardaq Ashımbekovıcha]	Кездесуде Нұрсұлтан Назарбаев Мемлекеттік күзет қызметінің басшылық құрамына осы мекеменің жаңа бастығы Ардақ Әшімбекұлын таныстырды [Kezdesýde Nursultan Nazarbaev Memlekettik kúzet qyzmetiniń basshylyq quramyna osy mekemeniń jańa bastyǵy Ardaq Áshimbekulyn tanystyrdy]

Сонымен тоқталайық жеті гранях Ұлы дала [Sonymen toqtalaıyq jeti granáh Uly dala]	Сонымен , Ұлы даланың жеті қырына тоқталайық [Sonymen , Uly dalanyń jeti qyryna toqtalaıyq]
Кездесу барысында тараптар ынтымақтастығының маңызды мәселелерін талқылады және одан әрі дамыту перспективалары туризм саласындағы қатынастарды [Kezdesý barysynda taraptar yntymaqtastyǵynyń mańyzdy máselelerin talqylady jáne odan ári damytý perspektıvalary týrızm salasyndaǵy qatynastardy]	Кездесу барысында тараптар ынтымақтастықтың маңызды мәселелері мен туризм саласындағы қатынастарды одан әрі дамыту перспективаларын талқылады [Kezdesý barysynda taraptar yntymaqtastyqtyń mańyzdy máseleleri men týrızm salasyndaǵy qatynastardy odan ári damytý perspektıvalaryn talqylady]

Analyzing the resulting sentences:

- word and word connection;
- the meaning of the sentence, and etc. can be seen to be well translated.

16. Conclusion and future work

Nowadays, the use of new technology and quality data is the result of machine translation. In order to improve the quality of translation, the system is constantly updated and checked, which simplifies and speeds up the work of translators. We analyzed online machine translation systems and identified common translation errors, as well as the advantages and disadvantages of machine translation systems from Russian into Kazakh and from English into Kazakh. In addition, a post-editing model was created for Russian-Kazakh and English-Kazakh translations, and a bilingual parallel corpus was assembled. The corpus, containing a total of 115785 sentences, was assembled using programming code tailored to the characteristics of each language. Various statistics are collected on the collected corpus and translation errors are grouped. After machine learning, we got the result: best bleu - 11.7.

The main contribution to this work is: automated collection of parallel corpuses for English-Kazakh and Russian-Kazakh translations. Development of a post-editing model for the Kazakh language in machine translation systems, with consider linguistic properties of the language. To improve the quality of translation and the work of the post-edit model, an approach based on a neural network was applied.

In the future, it is planned to increase the quantity of the corpus and use the Latin alphabet of the Kazakh language.

REFERENCES

Chakyrova Iý.I. Postredaktırovanıe v translatologıcheskoı paradıgme // Vestnık PNIPÝ. Problemy ıazykoznanıa ı pedagogıkı. - 2013. - №8. - S. 137-144.

Koponen M. Is Machine Translation Post-editing Worth the Effort? A Survey of Research into Post-editing and Effort // The Journal of Specialised Translation Issue 25 – January 2016, - C. 3-5

The Importance of Machine Translation Post-Editing and Its Application to Translation // WEB- Strastburg University website. [Electronic resource] - 2018. - URL: https://mastertcloc.unistra.fr/2018/12/03/ . - (date of access: 04/04/2021)

ISO 18587:2017 (2017). Translation Services – Post-editing of machine translation output – Requirements. International Organization for Standardization

Nechaeva N.V., Svetova S.Iý. postredaktırovanıe mashınnogo perevoda kak aktýalnoe napravlenıe podgotovkı perevodchıkov v výzah // Voprosy metodıkı prepodavanıa v výze. - 2018. - №7 (25). - S. 64-72

Fully automated machine translation service // WEB-сайт. – [Electronic resource] – 2020. – URL: https://www.semantix.com/machine-translation /. - (date of access: 04.04.2021)

Post-editing Machine Translation Best Practices. By Dan Zdarek // WEB- website. - [Electronic resource] - 2020. - URL: https://www.memsource.com/blog/post-editing-machine-translation-best-practices/ . - (date of access: 09.04.2021)

How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study Yanfang Jia, Hunan University Michael Carl, Kent State University Xiangling Wang, Hunan University // The Journal of Specialised Translation. - 31 January 2019. – C. 61-76.

Thang Luong, Eugene Brevdo, Rui Zhao: Building Your Own Neural Machine Translation System in TensorFlow, https://github.com/tensorflow/nmt.

https://translate.yandex.kz/

https://www.webtran.ru/

https://translate.google.kz/

https://sozdik.kz/

Vychıslıtelnaıa obrabotka kazahskogo ıazyka: sbornık naýchnyh trýdov/ pod redaksıeı Rahımovoı D.R. –Almaty: Qazaq Ýnıversıteti, 2020. -146 s.

BLUE metrics // https://en.wikipedia.org/wiki/BLEU: 04.11.2020.

Rakhimova, D., Turganbayeva, A. Auto-abstracting of texts in the Kazakh language // Proceedings of the 6th International Conference on Engineering & MIS. – 2020. – P. 1-5 // https://doi.org/10.1145/3410352.3410832

Rakhimova D., Shormakova A.: Problems of semantics of words of the Kazakh Language in the information retrieval // Proceedings of the 11-th International Conference, ICCCI 2019, Springer. 2019. Part II. – P. 70-81 (2019)

Rakhimova D. , Zhumanov Zh. Complex technology of machine translation resources extension for the Kazakh language // Studies in Computational Intelligence 2017 №710, 297 - 307 s.

Dujsebekova K.S., Kopbolsyn L.S., (2020) Vychislitel'naja obrabotka kazahskogo jazyka: sbornik nauchnyh trudov[Computational processing of the Kazakh language: collection of scientific papers] / pod redakciej Rahimovoj D.R. – Almaty: Қazaқ universiteti,– C. 123-124

Perehodko I.V., Máchın D.A. Osenka kachestva mashınnogo perevoda // Vestnık Orenbýrgskogo Gosýdarstvennogo ýnıversıteta. – 2017. – № 2 (202). – S. 92-96.

Kazakh-Tatar Machine Translation on the Base of Complete Set of Endings Model

Kamila Sagat[0000-0003-3885-3124], kamilasagat@gmail.com

Yntymak Abdrazakh[0000-0002-7119-1217]
yntymak05099@gmail.com

Ramazan Mukhamatzhanov[0000-0002-0980-1790]
mukhamatzhan.r@gmail.com

Ualsher Tukeyev[0000-0001-9878-981X], ualsher.tukeyev@gmail.com

Aliya Turganbayeva[0000-0001-9660-6928], turganbaeva.aliya@bk.ru

Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract

This article provides an overview of the works and technologies related to machine translation in accordance with a given pair of languages. Based on the model of a complete set of endings, the problem of machine translation for a pair of Tatar–Kazakh languages is considered. For the development of Kazakh-Tatar machine translation, the morphological model CSE (Complete Set of Endings) is taken as a basis. During the research work, the necessary linguistic resources were collected and a program was developed. A complete set of connections of Kazakh-Tatar languages is being created. Morphological analysis of compounds and compilation of morphological tables in accordance with each language. Morphological tables will also be identified; for a pair of Kazakh-Tatar languages, a matching list of a set of Stem words and a matching list of a set of Stop Word words will be formed. A machine translation algorithm for a pair of Kazakh-Tatar languages will be developed based on a model of a complete set of endings. Software based on this algorithm will be developed. The scientific contribution of the article is that new ma-

chine translation technology was created based on the CSE model. Experiments results shoes possibility of proposed technology of machine translation for Turkic languages.

Keywords:
Tatar language, Kazakh language, morphology model, machine translation, morphological analysis.

1. Introduction

The modern world and our near future depend on applied intelligent systems, as new technologies are evolving every day. One of the tasks of intelligent systems is machine (automated) translation from one natural language to another. Machine translation (MT) allows people to communicate regardless of language differences, as it removes the language barrier and opens up new languages for communication. In general, there are several types of machine translation. In current day, state-of art in machine translation is neural machine translation (NMT). Nevertheless, there are low-resource languages for which are not satisfactory volume of parallel corpora for realized of NMT for these languages. Most of Turkic languages are low-resource languages.

The Turkic languages are the largest linguistic group among the Altai languages. The syntax, morphology and phonology of the Turkic languages are similar. Kazakh and Tatar languages belong to the group of Turkic languages. Kazakh is the state language of the Republic of Kazakhstan, as well as the native language of Kazakhs living in China, Uzbekistan, Russia, Mongolia and other countries. The Kazakh language belongs to the Kipchak-Nogai branch of the Kipchak group of Turkic languages [4]. The national language of the Tatar people is the Tatar language. It is the state language of the Republic of Tatarstan and the second language in terms of distribution and number of speakers in the Russian Federation. The Tatar language belongs to the Kipchak group of Turkic languages, within which the Kipchak-Bulgar branch is located [5].

Despite the rapid development of machine translation from Kazakh into various languages and vice versa, there are still cases when some translation systems return low-quality translation results. In addition, many Tur-

kic languages are less resource-intensive, and this factor may lead to limited use of Turkic languages in the current era of digitalization. This problem motivates researchers to consider and use different methods and algorithms, and increase the amount of resources required. At the present time of rapid development of information flow, it is very important to develop Kazakh-Tatar machine translation and get quality results from it. Tatar and Kazakh languages are closely related to the Turkic language group and have many grammatical similarities. This similarity allows to use the CSE model for Kazakh and Tatar in language tools development. In order to implement the Kazakh-Tatar machine translation, the morphological model of the CSE is used. The CSE model is an example of a complete set of endings, and the fact that all the endings in it are summarized in tables greatly contributes to the implementation of natural language processing problems. The scientific contribution of this work is that new machine translation technology for Kazakh-Tatar was created based on the CSE model.

2. Related works

This section discusses previous work on the MT of these languages, including available corpus and language resources. Two Turkic languages considered in the study: Tatar and Kazakh.

In general, within the framework of the Apertium project [6], work is underway to create machine translation systems (key components such as rule-based morphological converters) for translation between Tatar-Bashkir, Turkish-Kyrgyz, Azerbaijani-Turkish, etc. Among the released MT systems is Kazakh-Tatar [7]. Kazakh-Tatar languages pair is supported by Google Translate and Yandex Translate. However, because these products are commercial systems, their source data and software are closed, and these products are diffiluclt to embed to new software products. Therefore, in this work is proposed alternative way for creating machine translation, based on using relational decision models based on CSE morphology model. By opinion of authors is actual to create machine translation systems with possibility to realize as small volume applications for speech-to-speech. Proposed MT technology based on CSE morphology model oriented on such tasks, because realized on relational (tabular) data models.

3. The Tatar and Kazakh languages

Tatar and Kazakh languages belong to the Kipchak (or north-western) group of Turkic languages. There is also a large community of native speakers in China, neighboring Central Eurasian republics, and Mongolia, where Kazakh is the main state language in Kazakhstan. The total number of speakers is at least 10 million. Tatar is spoken by about 6 million people in and around Tatarstan. It is an official language of the Russian Federation, along with Russians in Tatarstan.

The Kazakh and Tatar languages are an agglutinative language group with complex nominative and morphological and syntactic rules of polysynthetics. Due to the annual development of the country at the global level and the growth of external relations, various translation programs are widely used in the translation into Kazakh or other languages. In addition, there are a small number of dictionaries, multilingual, bilingual dictionaries, translation systems and other linguistic concepts in both Tatar and Kazakh languages. Therefore, the main goal is to increase these resources. It is known that machine translation from one language to another is one of the tasks of intelligent systems, for example, the quality of machine translation is in some cases lower. However, in most cases it is possible to know the original meaning. Therefore, this is the main problem of the machine drive system.

3.1 Difference of Tatar and Kazakh

Tatar and Kazakh languages, as close languages, combine many phonological processes, including systems of consonant sound. Although both Kazakh and Tatar belong to the group of Turkic languages, including the Kipchak group, which includes thick and hard sounds, there are differences and similarities in the phonetic system. In general, there are 42 letters in the Kazakh language and 39 letters in the Tatar language. There are 12 vowels in Kazakh and 9 in Tatar. Number of consonant sounds in the Kazakh language - 26, in the Tatar language - 28. Specific letters in the Kazakh language: ә, ғ, қ, ң, ө, ү, ұ, і, h. Letters typical of the Tatar language: ә, ө, ү, ж, ң, h. Vowels are divided into 3 parts in Kazakh and 2 parts in Tatar. The sound of [n] in Tatar is used in the same way as

the sound of [n] in Kazakh, which is pronounced through the nose. For example, rain is rain, new is new, and so on. One of the sounds in the Tatar language is represented by the Kazakh sounds "a", "e" and "y". For example, mother is mother, father is father, say is say, mother is mother, and so on.

Kazakh and Tatar are grammatically similar languages in the Turkic group of languages. In general, the grammatical structure of the Tatar language, as in the Kazakh language, is divided into root words, compound words, derivative words, compound words and double words. It is limited to 9 word groups in the grammar of the Kazakh language, and 6 basic word groups in the Tatar language. Word groups in the Tatar language are divided into auxiliary word groups, special word groups, main word groups. If we consider the system of cases of the Kazakh and Tatar languages, there are 7 cases in the Kazakh language, and 6 types of cases in the Tatar language (Table 1).

Table 1. The system of cases in the Kazakh and Tatar languages.

Cases		Endings	
Kazakh	**Tatar**	**Kazakh**	**Tatar**
Nominative case (Atau septigi)	Бас килеше (Bas kileshe)		
Genitive case (Ilik septigi)	Ялек килеше (Ialek kileshe)	-ның/-нің, -дың/-дің, -тың/-тің (nyŋ/niŋ, dyŋ/diŋ, tyŋ/tiŋ)	-ның/-нең
Dative case (Barys septigi)	Юналеш килеше (Iýnálesh kileshe)	-ға/-ге -қа/-ке (ğa/ge, qa/qe)	-га/-гә -ка/-кә
Accusative case (Tabys septigi)	Төшем килеше (Tóshem kileshe)	-ны/-ні -ды/-ді -ты/-ті (na/ne, dy/di, ty/ti)	-ны/-не
Locative case (Jatys septigi)	Чыгыш килеше (Chygysh kileshe)	-нан/-нен -тан/-тен -дан/-ден (nan/nen, dan/den, tan/ten)	-нан/-нән -тан/-тән -дан/-дән
Ablative case (Shyğys septigi)	Урын-вакыт килеше (Úryn-vakyt kileshe)	-да/-де -та/-те (da/de, ta/te)	-да/-дә -та/-тә
Instrumental case (Kómektes septigi)		-мен/-пен/- бен (men, ben, pen)	

As you can see from the table, in the Tatar language, without its auxiliaries, its function is performed by conjunctive pronouns such as birlan, berga, bergalep, berlek. Kazakh linguists also believe that the auxiliary suffix originated from this unit: historically.

Another example of the large-scale morphological differences between Tatars and Kazakhs is that in the Kazakh language there is a 2-sided polite form, and in the Tatar language there is no 2-sided polite form, which is given in Table 2.

Table 2. 1st, 2nd, 3rd pers. pronoun systems of Kazakh and Tatar

Persons	Kazakh language		Tatar language	
	singular	*plurar*	*singular*	*plural*
1st person	men	biz	мин	без
2nd person	sen	sender	син	сез
2nd side polite person	siz	sizder	син	сез
3rd person	ol	olar	ул	алар

In conclusion, the Kazakh-Tatar languages, which belong to the group of Turkic languages, are closely related, although they have differences in grammatical, lexical and phonetic levels.

4. Method of machine translation

The steps for machine translation of the Kazakh-Tatar language pair on the basis of the CSE model are as follows:

1. development of a complete set of endings for Kazakh-Tatar languages using the morphological model CSE
2. conducting a morphological analysis of the connections of Kazakh-Tatar languages, drawing up a morphological table of the connections of the Kazakh language and the Tatar language
3. identification of the morphological table of Kazakh-Tatar languages
4. create a list of stem words bases of the Kazakh-Tatar language
5. Create a list of stop words base of the Kazakh-Tatar language

6. creation of an algorithm for machine translation into the Kazakh-Tatar language based on the CSE model
7. creating software based on an algorithm

The steps listed above are analyzed separately below.

4.1 Development of a complete set of endings for Kazakh-Tatar languages using the morphological model CSE

A complete set of endings has been developed for the Kazakh language using the morphological model CSE[1]. Words in the Tatar language have 3 large conjunctions that are attached to the noun words, which are

Plural endings(K)
Possessive endings(T)
Case endings(C)
Stem words (S)

Let's look at all possible options for placing suffix types: one type, two types, and three types. The number of placements in the Tatar language is 7.

3 locations of one type of endings (K, T, C) are a valid definition in terms of meaning. Two different endings have 3 semantic favorable locations (KT, TC, KC). Of the three types of endings, there is 1 semantic acceptable type of placements (KTC). We have considered endings in Tatar as nominal base words (nouns, adjectives and numerals) and verbal base words (verbs, adverbs, adverbs, Rays). Complete set of Tatar endings: total – 5217.

Inferring of endings for Tatar language different placement types for Tatar endings presents in Tables 3-5.

Table 3. Inferring of endings for placement type KT (Plural-Possessive).

Tatar	Suffixes type K	Suffixes type T		Number
		Singular	Plural	
Examples	-нар- -нәр- -лар- -ләр-	-ым, -ем -ың, -ең -ы, -е	-ыбыз, -ебез -ыгыз, -егез -ы, -е	4*5=20
	-нар-	-ым, -ың, -ы	-ыбыз, -ыгыз	5
	-нәр-	-ем, -ең ,-е	-ебез, -егез	5
	-лар-	-ым, -ың, -ы	-ыбыз, -ыгыз	5
	-ләр-	-ем, -ең ,-е	-ебез, -егез	5

Number of endings fot Tatar language placement KT is 20.

Table 4. Inferring of endings for placement type KC (Plural-Case)

	Suffixes type K	Suffixes type C		Number
Examples	-нар- -нәр- -лар- -ләр-	1. nom. 2. gen. 3. dat. 4. acc. 5. loc. 6. abl. 7. gen.	- -ның, -нең -га, -гә -ны, -не -да, -дә -дан, -дән -белән	4*6=24
	-нар-	-ның, -га, -ны, -да, -дан, -белән		6
	-нәр-	-нең, -гә, -не, -дә, -дән, -белән		6
	-лар-	-ның, -га, -ны, -да, -дан, -белән		6
	-ләр-	-нең, -гә, -не, -дә, -дән, -белән		6

Table 5. Infering of endings for placement type TC (Possessive-Case)

	Suffixes type T		Suffixes type C		Number
	Singular	Plural			
Examples	м-, ым-, ем- н-, ын-, ен- сы-, се-, ы- е-	быз, без, ыбыз, ебез гыз, гез, ыгыз, егез	1. nom. 2. gen. 3. dat. 4. acc. 5. loc. 6. abl. 7. gen.	- ның, нең а, ә, га, гә, на, нә ны, не да, дә нан, нән, дан, дән белән	2*24+2* 18+3*6= 102
	м, ым	быз, ыбыз	-ның, а, га, ны, да, нан, дан, белән		24
	ем	без, ебез	-нең, ә, гә, не, дә, нән, дән, белән		18
	н, ын	гыз, ыгыз	-ның, а, га, ны, да, нан, дан, белән		24
	ең	гез, егез	-нең, ә, гә, не, дә, нән, дән, белән		18
	сы		-ның, на, ны, да, нан, белән		6
	се		-нең, нә, не, дә, нән, белән		6
	е		-нең, нә, н, ндә, ннән, белән		6

4.2 Development of relational decision tables of morphological analysis Tatar and Kazakh languages using the CSE model

Morphological analysis of Tatar endings presents as relational decision tables is shown in the Table 6-9.

Table 6. Morphological analysis of KT endings of the Tatar language

Tatar KT Endings	Morph analysis
нарым	<NB>+нар<pl>+ым<pos><sg><p1>
нарын	<NB>+нар<pl>+ын<pos><sg><p2>
нары	<NB>+нар<pl>+ы<pos><sg><p3>
нарыгыз	<NB>+нар<pl>+ыгыз<pos><pl><p2>
нарыбыз	<NB>+нар<pl>+ыбыз<pos><pl><p1>

Table 7. Morphological analysis of KC endings of the Tatar language

Tatar KC Endings	Morph analysis
нарның	<NB>+нар<pl>+ның<gen>
нарга	<NB>+нар<pl>+га<dat>
нарны	<NB>+нар<pl>+ны<acc>
нарда	<NB>+нар<pl>+да<loc>
нардан	<NB>+нар<pl>+дан<abl>
нар+белән	<NB>+нар<pl>+белән

Table 8. Morphological analysis of TC endings of the Tatar language

Morph analysis	Tatar TC Endings
<NB>+м<pos><sg><p1>+ның<gen>	мның
<NB>+м<pos><sg><p1>+а<dat>	ма
<NB>+м<pos><p1>+ны<acc>	мны
<NB>+м<pos><sg><p1>+да<loc>	мда
<NB>+м<pos><sg><p1>+нан<abl>	мнан
<NB>+м<pos><sg><p1>+белән	м+белән

Table 9. Morphological analysis of KTC endings of the Tatar language

Morph analysis	Tatar KTC Endings
<NB>+нар<pl>+ым<pos><sg><p1>+ның<gen>	нарымның
<NB>+нар<pl>+ым<pos><sg><p1>+а<dat>	нарыма
<NB>+нар<pl>+ым<pos><sg><p1>+ны<acc>	нарымны
<NB>+нар<pl>+ым<pos><sg><p1>+да<loc>	нарымда
<NB>+нар<pl>+ым<pos><sg><p1>+нан<abl>	нарымнан
<NB>+нар<pl>+ым<pos><sg><p1>+белән	нарым+белән

Thus, for 5217 Tatar language endings were constructed relational decision tables for morphological analysis.

Below is an example of one for each group based on the endings of Kazakh nominal and verval base words (Table 10-11).

Table 10. Morphological analysis of endings of Kazakh nominal base words

Types	Endings	Morph analysis
K	dar	<NB>*dar<pl>
T	m	<NB>*m<pos><sg><p1>
C	ğa	<NB>*ğa<dat>
J	myn	<NB>*myn<per><sg><p1>
KT	darym	<NB>*dar<pl>*ym<pos><sg><p1>
KC	dardyŋ	<NB>*dar<pl>*dyŋ<gen>
KJ	darmyz	<NB>*dar<pl>*myz<per><pl><p1>
TC	mnyŋ	<NB>*m<pos>*nyŋ<gen>
CJ	danyn	<NB>*da<loc>*myn<per><sg><p1>
TJ	msyŋ	<NB>*m<pos><p1><sg>*syŋ<per><sg><p2>
KTC	darynnyŋ	<NB>*dar<pl>*ym<pos><sg><p1>*nyŋ<gen>
KTJ	larymsyŋ	<NB>*dar<pl>*ym<pos><sg><p1>*syŋ<per><sg><p2>
KCJ	larğanyn	<NB>*dar<pl>*ğa<dat>*myn<per><sg><p1>
TCJ	manyn	<NB>*m<sg><p1>*a<dat>*myn<pos><sg><p1>
KTCJ	darymanyn	<NB>*dar<pl>*ym<pos><sg><p1>*a<dat>*myn<sg><p1>

Table 11. Morphological analysis of endings of Kazakh verbal base words

Types	Endings	Morph analysis
sprt	myn	<VB>*myn<sprt><per><sg><p1>
cprt	yp otyrmyn	<VB>*yp otyr<cprt>*myn<per><sg><p1>
tprt	amyn	<VB>*a<tpt>*myn<per><sg><p1>
neg tprt	maimyn	<VB>*mai<tpt>*myn<per><sg><p1>
pst	tym	<VB>*ty<pst>m<per><sg><p1>
neg pst	madym	<VB>*ma<neg>*dy<pst>*m<per><sg><p1>
ppt	ganmyn	<VB>*gan<ppt>*myn<per><sg><p1>
apt	yppyn	<VB>*yp<apt>*pyn<per><sg><p1>
neg apt	mappyn	<VB>*ma<neg>*p<apt>*pyn<per><sg><p1>
tpat	atynmyn	<VB>*atyn<tpt>*myn<per><sg><p1>
neg tpat	maitynmyn	<VB>*ma<neg>*ityn<tpt>*myn<per><sg><p1>
aft	arsyn	<VB>*ar<aft>*syn<per><sg><p2>
neg aft	aremespyn	<VB>*ar<aft>*emes<neg>*pyn<per><sg><p1>
ift	maqpyn	<VB>*maq<ift>*pyn<per><sg><p1>
neg ift	maqemespyn	<VB>*maq<ift>*emes<neg>*pyn<per><sg><p1>
tft	amyn	<VB>*a<tft>*myn<per><sg><p1>
neg tft	bamyn	<VB>*ba<neg>*i<tft>*myn<per><sg><p1>
RK	gandar	<VB>*gan<pp>*dar<pl>
RT	ganym	<VB>*gan<pp>*ym<pos><sg><p1>
RC	gannyn	<VB>*gan<pp>*nyn<gen>
RJ	ganmyn	<VB>*gan<pp>*myn<per><sg><p1>
RKT	gandarym	<VB>*gan<pp>*dar<pl>*ym<pos><sg><p1>
RKC	gandardyn	<VB>*gan<pp>*dar<pl>*dyn<gen>
RKJ	gandarmyz	<VB>*gan<pp>*dar<pl>*myz<per><pl><p1>
RTC	ganymnyn	<VB>*gan<pp>*ym<pos><sg><p1>*nyn<gen>
RKTC	gandarymnyn	<VB>*gan<pp>*dar<pl>*ym<pos><sg><p1>*nyn<gen>
RKCJ	gandargamyn	<VB>*gan<pp>*dar<pl>*ga<dat>*myn<per><sg><p1>

Thus, for 5368 Kazakh language endings were constructed relational decision tables for morphological analysis.

4.3 Development of corresponding table of morphological analysis of endings, stem words and stop words for Tatar and Kazakh languages using the CSE model

Using the collection of Kazakh-Tatar language endings based on the morphological model of the CSE, a corresponding table of morphological features of the endings of two languages was developed (Table 12). The morphological properties of each addition of the Kazakh language were found in the morphological table of the Tatar language, and the corresponding morphological features were added to a separate table.

Table 12. Morphological table of Kazakh-Tatar endings

Kazakh Endings	Kazakh Morph	Tatar Morph	Tatar Endings
dar	<NB>*dar<pl>	<NB>*нар<pl>	нар
m	<NB>*m<pos><sg><p1>	<NB>*м<pos><sg><p1>	м
ğa	<NB>*ğa<dat>	<NB>*га<dat>	га
myn	<NB>*myn<per><sg><p1>	<NB>* <per><sg><p1>	empty
darym	<NB>*dar<pl> *ym<pos><sg><p1>	<NB>*нар<pl> *ым<pos><sg><p1>	нарым
darymamyn	<NB>*dar<pl> *ym<pos><sg><p1> *a<dat>*myn<sg><p1>	<NB>*нар<pl> *ым<pos><sg><p1> *a<dat>* <per><sg><p1>	нарыма

According to this approach, a complete correspondence to the base of endings of each language has been developed. A total of 5,217 endings were identified for morphological characteristics. The basis of the Kazakh language was the correspondence of the endings of the Tatar language.

Translation into Tatar was carried out using special dictionaries and an online translation system to create a corresponding list of the prepared Kazakh and Tatar stem word base (Table 13).

Table 13. Stems in Kazakh and Tatar

stem in qazaq	stem in tatar
bir	бер
bala	бала
belgı	билге

In addition, a corresponding list of stop words of the Kazakh-Tatar language was compiled (Table 14).

Table 14. Stop words in Kazakh and Tatar

SW in qazaq	SW in tatar
men	мин
jäne	һәм
ne	яки

4.4 Creation of an algorithm for machine translation into the Kazakh-Tatar language based on the CSE model

Based on the CSE model, the algorithm for machine translation into Kazakh-Tatar was constructed as follows:

1. We break down a given sentence (SS) into a separate word (Wi)

   The word (Wi) is first compared with the dictionary of stop words. If the word (Wi) is in the dictionary of stop words, then it is considered a stop word, and a stop word in the language $WiTL{=}StWiTL$ is taken as a translation of the word in accordance with this word. Step 3 is performed if a match is not found in the dictionary of Stop words.
3. Stemming the separate word (Wi). $Stis$ is stem of Wi, Eis is ending of Wi (present in the base of endings).
   $fsteming(Wi){=}Stis{+}Eis$
   If the Wi is in the stop word base, a direct translation is put, you don't need to make stemming

4. We look for the Eis ending from the morphological table shown in Table 3. We find Eit - the ending of the corresponding line Eis.
5. In the stem word base, we find $Sti\ t$ - translation of the stem word $Stis$
6. $Ti = Sti\ t{+}Eit$
7. T = $\Sigma T\ i$

5. Experiments and results

The main advantage of the CSE model is mathematically deducing the legitimacy of the location of connections, using the model you can quickly and easily extract a list of connections, even if you do not fully know the established language. Therefore, in this work, a translation into the Kazakh-Tatar language was made using the CSE model.

In general, the algorithm was implemented in the python programming language, taking four documents as input. They:

- Morphological table of Qaz-tat compounds ("qaz-tat-tab.xlsx ");
- Qaz-tat Stem Dictionary ("qaz-tat-stems.xlsx ");
- Words qaz-tat stopwords ("qaz-tat-stopwords.xlsx ");
- Text in Kazakh ("text-qaz.txt ").

As a result of the program, a translation of the original Kazakh text or sentence into the Tatar language is obtained.

In testing the operation of the MA algorithm, implemented on the basis of a model of a complete set of connections of KAZ-tat languages, a corpus of 45 parallel sentences of KAZ-TAT languages was used. In addition, a model of a set of complete connections was used for the input of the program, consisting of a list of 12 stop words, a dictionary of 538 stems and 5217 connections. The results obtained for the Kazakh-Tatar language pair are presented in Table 15.

Table 15. The results obtained for the Kazakh-Tatar language pair (segment)

Source text in Qazaq	Translated text by the our program	Correct translation of text
Radiotehnika qazırgı kezde adamzat ömırınıŋ barlyq salasynda qoldanady.	радиотехникаеңрә qazırgıeңрә kezdә adamzateңрә ömırınıŋ barlyqeңрә salacыda qoldanачаклар .	Радиотехника хәзерге кайчанда кешелек тормышының барлык тармакта кулланыла .
Mysaly, kez kelgen qaşyqtyqta radiotelefon bailanysyn ornatu, keskın, chertej, suret, gazet matricalaryn taratu, tez әreketti telegraftyq radiobailanys jasau. Kosmos objektileri men jer, kosmos apparattarynyŋ өz arasynda tikelei bailanys ornatylady.	mysalы , hәp кайсы qaşyqtyqta radiotelefonы bailanyacын ornatuеңрә , keskınы , chertejeңрә , suretеңрә , gazeteңрә matricalarыны taratuеңрә , tezeңрә әrekettieңрә telegraftык radiobailanyseңрә jasauеңрә . kosmoseңрә objektileriеңрә men jepe , kosmoseңрә apparattarыныŋ үзеңрә araсыда tikeleeн bailanyseңрә ornatyлачаклар .	Мәсәлән , теләсә кайсы арада радиотелефон элемтәсе урнаштыру , сурәтләрне , сызымталарны , рәсемнәрне , газета матриалларын трансляцияләү , тиз хәрәкәт итүче телеграф элемтәсе . Космос объектлары һәм жир , кешлек аппаратлар үз арасында туры элемтә урнаштырыла .

The metrics TER (Translation Error Rate), WER (Word Error Rate) and BLEU (Bilingual Evaluation Understudy) were used to evaluate the results of the machine translation Kazakh-Tatar based on the CSE model.

Results of experiments presents in Table 16.

Table 16. Result of experiment

Language pair	TER	WER	BLEU
Kazakh- Uzbek	0.61	0.63	17%

The results of the experiment of machine translation Kazakh-Tatar of the proposed technology show a high percentage of errors, the translation itself in meaning turned out to be very close to the translation template.

6. Conclusion

In this research paper, a machine translation into the Kazakh-Tatar language was made based on the CSE model. The article describes the process of performing the steps necessary for machine translation, as well as fragments and examples of accumulated language resources. In addition, an algorithm has been created that is necessary for machine translation into the Kazakh-Tatar language based on the CSE model, its steps and formulas used are shown. The algorithm is implemented in the Python programming language.

Scientific contribution of this article: creation of machine translation technology based on the CSE model and demonstration of its results in experiments. The results obtained in experiments showed a value of 0.61 when evaluating the TER metric from a text composed of 630 words. Although the results of a formal assessment of machine translation using the proposed technology show a high percentage of errors, the translation itself in meaning turned out to be very close to the translation template. This shows that the use of the proposed technology is possible and the proposed machine translation technology does not need large volumes of parallel corpora for training.

In the future, it is planned to improve of quality of proposed machine translation technology, to create speech-to-speech (STS) technology on the cascade model. To do this, needs to create a text-to-speech (TTS) and speech-to-text (STT) parts of cascade model, and using the technology developed in this article for part text-to-text (TTT).

REFERENCES

Tukeyev, U.: Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language. Proceedings of the international conference Turkic languages processing TURKLANG-2015, pp. 91-100, Kazan. Tatarstan. in Russian, September 17–19 (2015)

Turkic languages – 2021. -URL: https://kk.wikipedia.org/wiki/%D0%A2%D2%AF%D1%80%D0%BA%D1%96_%D1%82%D1%96%D0%BB%D0%B4%D0%B5%D1%80%D1%96 – (access data: 20.12.2021)

Qamet, A., Tukeyev, U.: Morphology Model and Stemming of Turkish Words on Complete Set of Inflectional Endings, TURKLANG-2021

Yıldız, O.T., Avar, B., Ercan, G.: An Open, Extendible, and Fast Turkish Morphological Analyzer, ACL Anthology: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), September, 2019, Varna, Bulgaria, INCOMA Ltd., pages 1364–1372 (2019)

Sak H.: Integrating morphology into automatic speech recognition: morpholexical and discriminative language models for Turkish. Ph.D. thesis, PhD thesis, Bogazic¸i University, Istanbul (2011)

Khanna, T., Washington, J.N., Tyers, F.M. et al. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. Machine Trans-lation (2021). https://doi.org/10.1007/s10590-021-09260-6

Ilnar Salimzyanov, Jonathan North Washington, Francis Morton Tyers. A free/open-source Kazakh-Tatar machine translation system. Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.) Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 175–182