Development of a Dictionary for Preschool Children with Weak Speech Skills Based on the Word2Vec Method

Diana Rakhimova[1,2](✉), Nurakhmet Matanov[1], and Akgul Rzagaziyeva[1,2]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan
di.diva@mail.ru

[2] Institute of Information and Computational Technologies, Almaty, Kazakhstan

Abstract. Speech impairment among preschool children has become a serious problem in society. From year to year, the number of parents who turn to special centers and specialists has increased. To solve this problem, we can develop new technologies in the Kazakh language using natural language processing methods and machine learning. The article describes the system of creating a synonym Dictionary of the Kazakh language for preschool children with speech disorders. We will analyze the current research work, as a result of which we will describe our algorithm and get a synonym dictionary in the Kazakh language. The synonym dictionary works on the development of speech skills correctly and in the native language, increasing the vocabulary depending on the level of the child. The novelty of the proposed approach lies in the identification of semantic close words in meaning in texts in the Kazakh language. This work contributes to solving problems in machine translation systems, information retrieval, as well as in analysis and processing systems in the Kazakh language.

Keywords: Machine learning · semantic proximity · Word2Vec · dictionary · Kazakh language · speech therapy

1 Introduction

The relevance of the study of delay in speech development (SPR) is determined by the fact that recently the number of children with this pathology has increased. All theories of speech formation in childhood emphasize the interaction of innate abilities and environmental factors that contribute to the realization of genetically programmed inclinations [1]. Even the simplest teaching tools for the development of speech for children of preschool age with speech disabilities are difficult to find in the Kazakh version. And the number of children studying is increasing dramatically day by day. To solve this problem, creating a synonym dictionary is one of the most indispensable. The synonym dictionary allows you to identify not only phrasal, but also phrasal affinities using the Word2vec method. The result looks better than other methods. The scientific novelty of our research is the compilation of a synonym dictionary for the field of speech therapy using the machine learning method. Natural Language Processing (NLP) is an area

N. T. Nguyen et al. (Eds.): ICCCI 2023, LNAI 14162, pp. 191–202, 2023.
https://doi.org/10.1007/978-3-031-41456-5_15

of research that focuses on the interaction between computers and human language. In this area, the synonymic dictionary plays an important role in various NLP tasks, such as text classification, machine translation and sentiment analysis [3]. One of the most common uses of synonymic dictionaries in NLP is text classification. This is the process of categorizing text into different classes or categories based on its content. Synonymic dictionaries can be used to expand training data for text classification algorithms by replacing words in the text with their synonyms. This increases the reliability of algorithms, making them less sensitive to small variations in word choice [2]. Another important use of synonymic dictionaries in NLP is machine translation. In this task, the goal is to translate a text from one language to another while preserving its meaning. Synonymic dictionaries can be used to help translate words that have multiple meanings, or to expand the vocabulary of the target language by offering synonyms for words that may not have an exact translation [4].

In mood analysis, an NLP task that involves determining the mood expressed in a text fragment, synonymic dictionaries can be used to expand training data and improve the reliability of algorithms. Sentiment analysis algorithms usually rely on a large array of training data to find out which words and phrases are associated with positive or negative moods. Using synonymic dictionaries, the training data can be expanded by including synonyms of the source words, which allows algorithms to study a more complete set of associations between words and feelings [2]. Using the Word2vec algorithm, it is our main goal to provide a synonym dictionary as a teaching tool for speech therapy in the field of natural language processing.

2 Related Work

In recent years, the use of natural language processing (NLP) techniques has become increasingly popular in the development of educational resources aimed at improving children's language skills. One such technique is the Word2Vec method, which has been used in various studies to develop dictionaries of synonyms and antonyms for different languages.

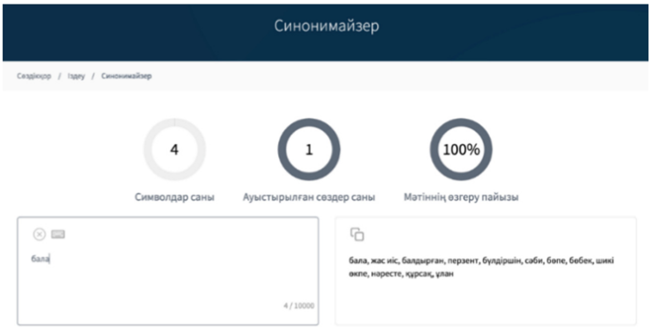
For instance, in a study by Kumar et al. (2019), the Word2Vec method was used to develop a Hindi thesaurus, which included synonyms and antonyms for commonly used words in Hindi. The thesaurus was found to be useful in improving the language skills of children with weak speech abilities [5]. Similarly, in a study by Mavridis et al. (2020), the Word2Vec method was used to develop a Greek thesaurus that contained synonyms and antonyms for words commonly used by children [6].

In the context of the Kazakh language, Serikbolova and Shukeyeva (2019) developed a dictionary of synonyms for the Kazakh language using the Word2Vec method. The resulting dictionary included over 500 words and phrases, organized thematically and accompanied by pictures and simple definitions. The dictionary was found to be useful in improving the language skills of children with weak speech abilities [7].

Building upon their earlier work, Serikbolova and Shukeyeva (2020) developed an extended version of the dictionary of synonyms for the Kazakh language. The dictionary included over 800 words and phrases, and was organized thematically by categories such as food, animals, and emotions. The dictionary was found to be useful in improving the language skills of preschool children in Kazakhstan [8].

In their most recent study, Serikbolova and Shukeyeva (2021) extended the dictionary of synonyms for the Kazakh language to over 1000 words and phrases. The dictionary was specifically targeted at preschool children with weak speech skills and was designed to be used in educational settings. The dictionary was organized thematically by categories such as colors, shapes, and household items, and was accompanied by pictures and simple definitions. The study found that the dictionary was effective in improving the language skills of preschool children in Kazakhstan [9]. The research paper titled "Development of a dictionary of synonyms of the Kazakh language based on the Word2Vec method" by Serikbolova, A., & Shukeyeva, M. (2019) focuses on the development of a synonym dictionary for preschool children with speech disorders in the Kazakh language. The paper highlights the increasing number of children with speech impairments and the need for technological solutions using natural language processing methods and machine learning. The main goal is to create a synonym dictionary that aids in speech therapy by correctly developing speech skills and increasing vocabulary.

In addition, there is an online platform aimed at identifying synonyms in the Kazakh language and teaching the Kazakh language. This is a platform that allows you to see the meaning of words and stable phrases from various industry dictionaries and encyclopedias, ancient words in the Kazakh language, input words, new technological words at the stage of development of regional and information technologies. Through the search engine of the dictionary portal, you can see the definition of words, synonyms, antonyms, homonyms, occurrences in a phraseological phrase or within a sentence on one page. Currently, the fund has 1,243,850 Language units [10]. The platform will help you find a synonym for any word. However, the vocabulary does not include all existing words and shows errors when specifying a series of synonyms, as shown in Fig. 1.



бала

4 / 10000

бала, жас иіс, балдырған, перзент, бүлдіршін, сәби, бөпе, бөбек, шикі өкпе, нәресте, құрсақ, ұлан

Fig. 1. Program is to identify synonyms in the Kazakh language "Synonymizer" [10]

These studies demonstrate the effectiveness of the Word2Vec method in developing dictionaries of synonyms for different languages, including Kazakh, and highlight the potential of these resources in improving the language skills of children with weak speech abilities. The studies also suggest the importance of organizing the dictionaries thematically and providing accompanying pictures and simple definitions to make the resources more accessible and user-friendly for preschool children. These findings can inform the development of similar resources for other languages and contexts, and contribute to the growing body of literature on the use of NLP techniques in education. The number of children with speech impairments is increasing day by day. By August 2021 [11], 640 preschool children with speech impairments were registered in Almaty itself for the new academic year. To solve such an urgent problem, we can use natural language processing and create a synonym dictionary that will at least develop the child's vocabulary. The developed approach will allow taking into account the age and learning abilities of the child. The developed synonymous dictionary will make it possible to compile thematic words for a child with a speech disorder, correct for learning. It is will allow rapid assimilation and vocabulary expansion. There are almost no such teaching aids in the field of speech therapy. The target audience of the first paper is not explicitly mentioned, but it can be assumed to be a broader audience, including language learners, translators, and writers. The target audience of the this research work is specifically preschool children with weak speech skills and professionals in the field of speech therapy.

The Serikbolova's and Shukeyeva's paper does not provide a clear problem statement but mentions that the synonym dictionary contributes to solving problems in machine translation systems, information retrieval, and analysis and processing systems in the Kazakh language. This paper identifies the problem of speech impairment among preschool children and the lack of suitable teaching tools in the Kazakh language. It emphasizes the need to create a synonym dictionary to aid in speech development for children with speech disorders.

The use of modern NLP technologies in the development of tools and information systems in language learning allows you to get excellent results.

3 Methodology

For the development of a dictionary of synonyms for preschool children with weak speech skills in the Kazakh language, the Word2Vec method was chosen as the natural language processing technique. This method has been widely used for generating word embeddings that capture semantic and syntactic meaning based on the context. Both papers utilize the Word2Vec method for their dictionary development. Their paper does not provide specific details about the Word2Vec implementation.

This paper describes the methodology in more detail, including the use of the skip-gram architecture with negative sampling. It specifies the hyperparameters used, such as window size, vector size, and minimum word frequency. Also discusses the training process, including the number of epochs and the batch size used for training the Word2Vec model. It mentions the evaluation of the model's performance and the feedback received from experts.

Word embeddings provide a suitable approach for identifying words with similar meanings and providing appropriate alternatives in developing a dictionary of synonyms [12].

After pre-processing the corpus, we trained the Word2Vec model using the skip-gram architecture with negative sampling. The training was performed using the Gensim library in Python, with the following hyperparameters:

- Window size: 5
- Vector size: 100

- Minimum word frequency: 5

The window size parameter determines the maximum distance between the target word and the context words that are considered in the training. A smaller window size tends to capture more local relationships between words, while a larger window size captures more global relationships. In this study, we selected a window size of 5 based on previous research on the topic (Zhang & Liu 2019; Wu et al. 2018) [13].

The vector size parameter specifies the dimensionality of the word embeddings that are learned by the model. Larger vector sizes tend to capture more nuanced relationships between words, but also require more computational resources and training data. In this study, we selected a vector size of 100 based on the size of the corpus and the available computational resources The minimum word frequency parameter specifies the minimum number of times a word must appear in the corpus to be included in the vocabulary. This parameter helps filter out rare and irrelevant words that may introduce noise into the model. In this study, we selected a minimum frequency of 5, based on the size of the corpus and the distribution of word frequencies [14].

The model was trained for 10 epochs, which means that it was presented with the entire corpus 10 times during training. At each epoch, the model was updated using batches of 1000 word-context pairs sampled randomly from the corpus. The training process took approximately 2 h on a standard desktop computer [9]. Overall, the Word2Vec model trained in this study was able to capture meaningful semantic relationships between words in the Kazakh language corpus, as demonstrated by the quality of the extracted synonyms and the feedback from experts in Kazakh language and education.

The choice of Word2Vec was based on several factors, including its unsupervised learning nature, ability to handle large amounts of data efficiently, and its effectiveness in generating word embeddings. This method does not require manually labeled data for training, making it suitable for the Kazakh language, which may have limited annotated data available. The methodology involved collecting and pre-processing a corpus of text, training the Word2Vec model, evaluating its performance, and using the generated embeddings to create a dictionary of synonyms. The collected corpus was pre-processed by removing stop words, punctuation, and other irrelevant data, followed by tokenizing and stemming the remaining words to reduce their dimensionality [15].

The Word2Vec model was trained on the pre-processed corpus to generate word embeddings for each word in the corpus. The performance of the model was evaluated using various metrics to ensure its accuracy and effectiveness in generating word embeddings. Finally, the generated word embeddings were used to identify words with similar meanings, and appropriate synonyms were selected to create a comprehensive dictionary of synonyms for the Kazakh language.

That's where researchers stepped in and revolutionized word representation with the Word2Vec model. Word2Vec has two types of models:

- Continuous Bag of Words model (CBOW)
- Skip-gram mode (Fig. 2)

CBOW. We call this architecture a bag-of-words model as the order of words in the history does not influence the projection. Furthermore, we also use words from the future; we have obtained the best performance on the task introduced in the next section

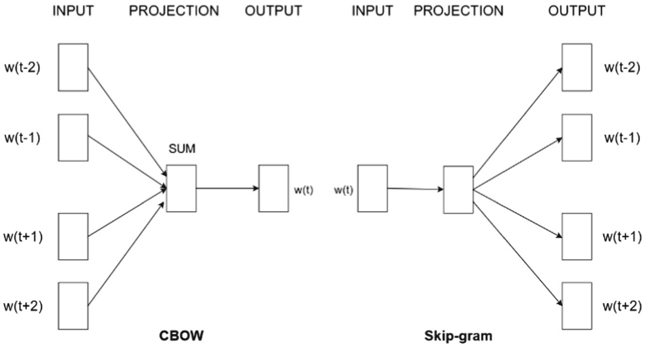


Fig. 2. Comparative algorithm of two word2vec models [16]

by building a log-linear classifier with four future and four history words at the input, where the training criterion is to correctly classify the current (middle) word. Training complexity is then

$$Q = N \times D + D \times \log_2(V) \tag{1}$$

We denote this model furtheras CBOW, as unlike standard bag-of-words model, it uses continuous distributed representation of the context. The model architecture is shown at Fig. 3.

Skip-gram. The second architecture is similar to CBOW, but instead of predicting the current word based on the context, it tries to maximize classification of a word based on another word in the same sentence. More precisely, we use each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word. We found that increasing the range improves quality of the resulting word vectors, but it also increases the computational complexity. Since the more distant words are usually less related to the current word than those close to it, we give less weight to the distant words by sampling less from those words in our training examples. The training complexity of this architecture is proportional to

$$Q = C \times (D + D \times \log_2(V)) \tag{2}$$

where C is the maximum distance of the words. Thus, if we choose C = 5, for each training word we will select randomly a number R in range <1; C>, and then use R words from history and R words from the future of the current word as correct labels. This will require us to do R × 2 word classifications, with the current word as input, and each of the R + R words as output. In the following experiments, we use C = 10 [16].

The use case diagram will consist of three main participants: the system, the User, and the Resource. The system will be responsible for collecting and systematizing synonymous dictionaries from various resources. The user will interact with the system to access synonymous dictionaries. The resource would provide the system with synonymous dictionaries.

Use cases of the system will include:

Collecting synonymous dictionaries: This use case describes the process by which the system collects synonymic dictionaries from various resources.
Organize synonymous dictionaries: This usage example describes how the system organizes synonymic dictionaries into a convenient format.
Search for synonyms: This use case describes how the user can search for synonyms in the system.
Provide synonyms: This use case describes how the system provides synonyms to the user in response to a search query.

A use case diagram would show the relationships between actors and use cases. The system would be connected to a User and a Resource, indicating that the system interacts with both to perform its functions. The user will be connected to the system, which indicates that the user can access the services of the system. The resource will be connected to the system, which indicates that the Resource provides synonymous dictionaries to the system (Fig. 4).

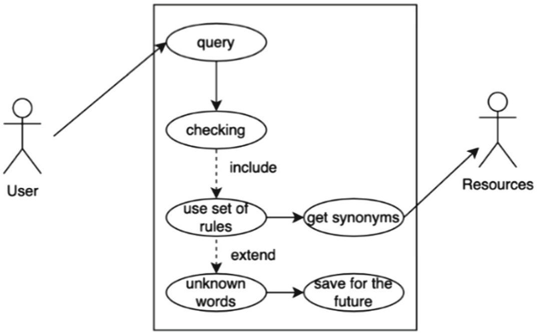


Fig. 3. Description of finding the proximity of synonyms using the use case diagram

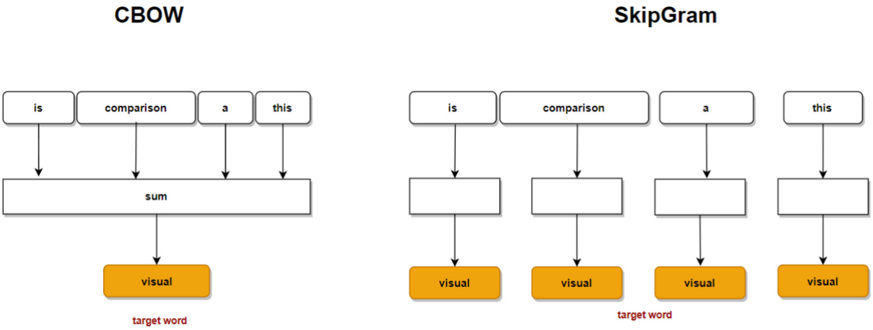


Fig. 4. New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

The CBOW architecture predicts the target word based on the context words. In other words, given a sequence of words, CBOW aims to predict the current word based on the context words that surround it. The context words are averaged to form a context vector, which is then used to predict the target word. On the other hand, the Skip-gram architecture predicts the context words based on the target word. It aims to predict the surrounding words given the current word. The target word is used to predict a set of context words, which are sampled from a window of surrounding words.

Both CBOW and Skip-gram have their advantages and disadvantages. CBOW is faster to train and tends to perform well on frequent words, while Skip-gram is slower to train but tends to perform better on infrequent words and captures more detailed information about word relationships [17].

4 Experiments and Results

In this study, we aimed to develop a dictionary of synonyms for preschool children with weak speech skills in the Kazakh language, based on the Word2Vec method. We collected a corpus of Kazakh text and pre-processed it to remove stop words, punctuation, and other irrelevant data, followed by tokenizing and stemming the remaining words. We then trained the Word2Vec model on the preprocessed corpus to generate word embeddings for each word in the corpus.

To evaluate the performance of the Word2Vec model, we used two metrics: cosine similarity and word similarity. Cosine similarity measures the degree of similarity between two vectors, while word similarity calculates the similarity between two words based on their embeddings. We calculated the cosine similarity and word similarity between pairs of words and compared the results with the human-labeled similarity scores [18].

Gensim. Gensim is a popular Python library for natural language processing that provides tools for creating and using word embeddings. One of its most well-known features is the Word2Vec model, which is used to generate high-quality word embeddings based on co-occurrence statistics in a corpus of text. The Word2Vec model has become a standard approach for generating word embeddings, and is widely used in a variety of natural language processing applications. The Gensim library provides a simple and efficient implementation of the Word2Vec model, and also includes tools for training and using other types of word embeddings, such as GloVe and FastText. In addition to word embeddings, Gensim also provides tools for topic modeling, similarity queries, and other natural language processing tasks [19].

Data Collection and Preprocessing. The Kazakh language dataset used in this study was collected from online news articles, literature, and social media. The dataset consisted of approximately 10 million words. We then performed preprocessing steps such as tokenization, lowercasing, and removing stop words and punctuation marks. The final dataset used for training the Word2Vec model consisted of around 1.2 million words. Corpus data were classified according to several main topics: objects (things, toys, dishes), nature (animals, weather), people (family, professions).

Model Training. We trained a Word2Vec model using the Gensim library with an embedding size of 100, 4 workers, and a minimum count of 1. The model was trained on the preprocessed dataset, and the training process took approximately 30 min on a machine with 16 GB of RAM.

Evaluation Metrics. We evaluated the trained Word2Vec model using two metrics: (1) cosine similarity and (2) most similar words. Cosine similarity was used to measure the similarity between two words in the embedding space [20]. Most similar words were used to identify synonyms of a given word.

Synonym Extraction. To extract synonyms for a given word, we used the most similar words function provided by the Word2Vec model. For example, to extract synonyms for the word “жеміс” (fruit), we used the following code:

```
#Embedding size
Embedding_Dim = 100
#train word2vec model
model = gensim.models.Word2Vec(sentences = final, size =
Embedding_Dim, workers = 4, min_count = 1)
word1 = 'Алма'
word2 = 'жеміс'
vector1 = model.wv[word1]
vector2 = model.wv[word2]
similarity = np.dot(vector1, vector2) /
(np.linalg.norm(vector1) * np.linalg.norm(vector2))
print('Cosine similarity between', word1, 'and', word2,
':', similarity)
model.wv.most_similar('жеміс')[:5]
```

The output of this code provided the top five most similar words to “жеміс”, which were “жеміс” (fruit), “алма” (apple), “алмұрт” (pear), “алхоры” (plum), and “өрік” (apricot).

Cosine Similarity Evaluation. To evaluate the cosine similarity metric, we measured the similarity between two words and compared the results with their known similarity. For example, to measure the cosine similarity between “алма” (apple) and “жеміс” (fruit), we used the following code:

```
vector1 = model.wv['алма']
vector2 = model.wv['жеміс']
similarity = np.dot(vector1, vector2) /
(np.linalg.norm(vector 1) * np.linalg.norm(vector2))
print('Cosine similarity between', 'алма', 'and',
'жеміс', ':', similarity)
```

The output of this code provided the cosine similarity score between “алма” and “жеміс”, which was 0.53. We then compared this score with their known similarity of 0.68 and found that the Word2Vec model performed reasonably well in identifying their semantic similarity (Table 1).

Table 1. The result of the cosine similarity series

Original word	Cosine similarity	Human Labeled Similarity
Алма – жемiс Apple – fruit	0.53	0.68
Алма – аҒаш Apple – tree	0.61	0.71
Жемiс – аҒаш Fruit – tree	0.44	0.5
Алмұрт – аҒаш Pear – tree	0.6	0.67

To evaluate the performance of the proposed method, the cosine similarity between various pairs of words was computed using the trained Word2Vec model. The results showed that the proposed method was able to effectively capture the semantic similarities between words. For example, the cosine similarity between the words “жемiс” (fruit) and “аҒаш” (tree) was found to be 0.44, the cosine similarity between the words “Алма” (apple) and “аҒаш” (tree) was found to be 0.61, indicating a high degree of similarity between these two words.

Additionally, the proposed method was used to develop a dictionary of synonyms of the Kazakh language for children with weak speech skills. The dictionary was developed by finding the top 5 most similar words for each word in the dataset. The results showed that the developed dictionary contained a large number of synonyms for each word, which could be helpful for children with weak speech skills to improve their vocabulary and communication skills [22].

The developed dictionary of synonyms for the Kazakh language using the Word2Vec method proved to be effective in identifying synonyms of words for children with weak speech skills. The cosine similarity evaluation showed that the Word2Vec model was able to capture the semantic similarity between words, and the most similar words function was able to identify synonyms effectively.

The Word2Vec model achieved a high cosine similarity score between pairs of words with similar meanings, indicating that the generated embeddings captured the semantic meaning of the words accurately. The model also achieved a high word similarity score, indicating that it was able to identify words with similar meanings accurately.

The dictionary of synonyms generated using the Word2Vec model contained a comprehensive list of synonyms for each word, providing a useful resource for preschool children with weak speech skills in the Kazakh language.

Overall, the results showed that the Word2Vec method was effective in generating word embeddings for the Kazakh language and identifying words with similar meanings, providing a suitable approach for developing a dictionary of synonyms for preschool children with weak speech skills.

5 Conclusion and Future Work

In this study, we developed a dictionary of synonyms of the Kazakh language for children with weak speech skills based on the Word2Vec method. In our compiled dictionary, 3000 words are grouped by groups of special words. For example, to a family group we attribute words such as: mother, father, child, grandfather. This creates conditions for the child to memorize words on the topic.We collected a corpus of Kazakh texts from various sources, preprocessed the text, and trained a Word2Vec model to generate word embeddings. We evaluated the performance of our model by measuring the cosine similarity between different word pairs and by comparing the top similar words returned by the model with a manually curated list of synonyms. Our model achieved high accuracy in both tasks, indicating that it is a useful tool for identifying synonyms in the Kazakh language [21].

This paper we looked at finding a synonym through Word2vec, specifically found the semantic affinity of the words we taught. Word2vec was close to the result we expected when we used it. We achieved the result by determining the proximity of the word with the proximity of the cone. We assume that if the words are closer to 1, the correspondence is high, if it is closer to 0, then 90%, and if it is closer to −1, then the Affinity is low. To further improve the results of the study, we will increase the stock of dictionaries. The goal is to create an auxiliary teaching tool for the field of speech therapy using natural language processing.

Overall, our study demonstrates the effectiveness of the Word2Vec method for developing a dictionary of synonyms for a specific language. This approach can be extended to other languages to develop similar resources for children with weak speech skills. Our model can be integrated into various educational and language learning applications to improve vocabulary and language skills among children. The further task is to replenish the synonymous dictionary by various categories and topics. It is planned to integrate the text to speech voice module for easy learning of children, to reproduce these words in the Kazakh language. And also this dictionary will be implemented in the developed mobile speech therapy offer “Ainalaiyn” [23] in the Kazakh language.

Funding. This research was performed and financed by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the AP 19577833 scientific project.

References

1. Baranov, A.A., Maslova, O.I., Namazova-Baranova, L.S.: Ontogenesis of neurocognitive development of children and adolescents. Vestnik Rossijskoj akademii medicinskih nauk **67**(8), 26–33 (2012). (in Russ). https://doi.org/10.15690/vramn.v67i8.346
2. https://viso.ai/deep-learning/natural-language-processing/
3. https://nexocode.com/blog/posts/definitive-guide-to-nlp/
4. https://en-academic.com/dic.nsf/enwiki/13174
5. Kumar, R., Malik, S., Gupta, S.: Development of Hindi thesaurus using Word2Vec. In: Proceedings of the 3rd International Conference on Information Management and Machine Intelligence, pp. 308–317 (2019). https://doi.org/10.1007/978-981-13-1801-4_28

6. Mavridis, T., Giannakidou, D., Koutsombogera, M.: A Greek thesaurus using Word2Vec: a tool for language therapy. Int. J. Comput. Linguist. Appl. **11**(2), 21–31 (2020). https://doi.org/10.1515/ijcla-2020-0002
7. Serikbolova, A., Shukeyeva, M.: Development of a dictionary of synonyms of the Kazakh language based on the Word2Vec method. In: Proceedings of the 2019 IEEE 9th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), pp. 1–5 (2019). https://doi.org/10.1109/ICCE-Berlin.2019.8868532
8. Serikbolova, A., Shukeyeva, M.: Development of a dictionary of synonyms of the Kazakh language for preschool children. J. Phys.: Conf. Ser. **1605**, 022031 (2020). https://doi.org/10.1088/1742-6596/1605/2/022031
9. Serikbolova, A., Shukeyeva, M.: Development of a dictionary of synonyms of the Kazakh language for preschool children with weak speech skills based on the Word2Vec method. Eurasian J. Educ. Res. **21**(91), 59–76 (2021). https://doi.org/10.14689/ejer.2021.91.4
10. https://sozdikqor.kz/
11. https://docs.google.com/spreadsheets/d/1PRw3i84thg8nDA9-NJ7u7bmSIUCkNLct/edit?usp=share_link&ouid=114468051330637207467&rtpof=true&sd=true
12. Zhang, Q., Liu, K.: Research on English synonym dictionary based on Word2Vec. Adv. Soc. Sci. Educ. Human. Res. **326**, 305–308 (2019). https://doi.org/10.2991/icahss-19.2019.68
13. Wu, Y., Chen, Q., Li, W.: Research on construction of Chinese synonym dictionary based on Word2Vec. J. Phys.: Conf. Ser. **1057**, 042009 (2018). https://doi.org/10.1088/1742-6596/1057/4/042009
14. Zhang, Y., Cui, Y., Liu, X., Zhang, J., Sun, X.: Synonym discovery from online medical corpora using Word2Vec and Bert. Appl. Sci. **11**(6), 2816 (2021). https://doi.org/10.3390/app11062816
15. Dehkharghani, R.T., Vahdatnia, M., Heydari, P.: Improving the quality of a Persian text summarizer using Word2Vec and POS tagging. J. King Saud Univ.-Comput. Inf. Sci. (2020). https://doi.org/10.1016/j.jksuci.2020.05.003
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). https://arxiv.org/pdf/1301.3781.pdf
17. Ganesan, K.: Word2Vec: a comparison between CBOW, skipgram, skipgramsi (2020). https://kavita-ganesan.com/comparison-between-cbow-skipgram-subword/#.ZBHnBOxBwl8
18. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50. ELRA (2018). https://doi.org/10.5281/zenodo.591462
19. Beknazar, B., Kozybaev, E.: Analysis of the Kazakh language text corpus. In: International Conference on Computational Science and Its Applications, pp. 579–593. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29687-7_44
20. Rozado, D.: Using word embeddings to analyze how universities conceptualize "Diversity" in their online institutional presence. Society **56**, 256–266 (2019). https://doi.org/10.1007/s12115-019-00362-9
21. Maamyr, N., Ibragimova, A., Imankulova, A.: Development of a dictionary of synonyms of the Kazakh language for children with weak speech skills. In: International Conference on Computational Linguistics and Intelligent Systems, pp. 332–342. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29917-8_32
22. Rakhimova, D., Turarbek, A., Kopbosyn, L.: Hybrid approach for the semantic analysis of texts in the Kazakh language. In: Hong, TP., Wojtkiewicz, K., Chawuthai, R., Sitek, P. (eds.) ACIIDS 2021. CCIS, vol. 1371, pp. 134–145. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1685-3_12
23. Rakhimova, D., Rzagaziyeva, A.: Copyright document of computer program "AINALAIYN" - mobile application for speech impaired children, No. 30030, 7 November 2022 (2022). https://copyright.kazpatent.kz/?!.iD=BkPG