МРНТИ 16.31.21
УДК 004.912

https://doi.org/10.51889/2938.2022.14.84.020

A TASK OF SYNTHETIC CORPORA GENERATION FOR THE LOW-RESOURCE LANGUAGE

Rakhimova D.[1,3]*, Adali Eşref* [2]*, Shormakova A.* [1]*, Asem Turarbek* [1*]*, Yerkin Suleimenov*[3]

[1]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[2] *Istanbul Technical University, Istanbul, Turkey*
[3] *Institute of Information and Computational Technologies, Almaty, Kazakhstan*
[*]*e-mail: asem.turarbek@kaznu.edu.kz*

Abstract

Recently, various areas of artificial language processing have been actively developing, such as search engines, machine translation technologies, speech technologies, etc. using machine learning technology and non-neural networks. For the implementation and development of these areas, first of all, the task of electronic linguistic resources such as corpora, dictionaries, a set of rules, etc. is acute. These resources should be of a very large volume of good quality. In this article, the problem of shortage of buildings for low-resource languages, which include the Turkic-speaking group, is considered. This is a problem for low-resource languages, such as Kazakh, because there are very few available corpora. This article presents an approach to the creation of synthetic corpora by the method of determining and replacing a candidate word from the list of synonymous dictionary of the Kazakh language.

Test experiments were conducted. As a result, the specified case was enlarged 3.37 times.

Keywords: corpora, Kazakh language, synonyms, linguistic resources.

Аннотация

Д. Рахимова [1,3]*, Эшреф Адали* [2,] *А. Шормакова* [1]*, А.Турарбек*[1]*, Е. Сулейменов* [3]
[1] *Казахский национальный университет имени аль-Фараби, г.Алматы, Казахстан*
[2] *Стамбульский технический университет, г. Стамбул, Турция*
[3] *Институт информационно-вычислительных технологий, г. Алматы, Казахстан*

ЗАДАЧА СОЗДАНИЯ СИНТЕТИЧЕСКИХ КОРПУСОВ ДЛЯ МАЛОРЕСУРСНОГО ЯЗЫКА

В последнее время активно развиваются различные направления обработки искусственного языка, такие как поисковые системы, технологии машинного перевода, речевые технологии и т. д. с использованием технологий машинного обучения и нейронных сетей. Для реализации и развития этих направлений, в первую очередь, решаются задачи электронных лингвистических ресурсов, таких как корпуса, словари, своды правил и т.п. является острым. Эти ресурсы должны быть очень большого объема хорошего качества. В статье рассматривается проблема нехватки корпусов для малоресурсных языков, к которым относится тюркоязычная группа. Это проблема для языков с низким ресурсом, таких как казахский, потому что доступных корпусов очень мало. В статье представлен подход к созданию синтетических корпусов методом определения и замены слова-кандидата из списка синонимического словаря казахского языка.

Были проведены тестовые эксперименты. В результате указанный корпус был увеличен в 3,37 раза.

Ключевые слова: корпусы, казахский язык, синонимы, лингвистические ресурсы.

Аңдатпа

Д. Рахимова [1,3]*, Эшреф Адали* [2,] *А. Шормакова* [1]*, А.Турарбек*[1]*, Е. Сулейменов* [3]
[1] *Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан*
[2] *Стамбул техникалық университеті, Стамбул қ., Түркия*
[3] *Ақпараттық және есептеу технологиялар институты, Алматы қ., Қазақстан*

РЕСУРСТАРЫ АЗ ТІЛ ҮШІН СИНТЕТИКАЛЫҚ КОРПУС ҚҰРУ ЕСЕБІ

Соңғы кездері машиналық оқыту технологиялары мен нейрондық емес желілерді пайдалана отырып, іздеу жүйелері, машиналық аударма технологиялары, сөйлеу технологиялары және т.б. сияқты жасанды тілдерді өңдеудің әртүрлі бағыттары белсенді түрде дамып келеді. Бұл бағыттарды жүзеге асыру және дамыту үшін ең алдымен электронды лингвистикалық ресурстардың корпустар, сөздіктер, ережелер жинағы және т.б. міндеттері шешіледі. өткір. Бұл ресурстар өте үлкен және сапалы болуы керек. Бұл мақалада түркітілдес топты қамтитын ресурсы төмен тілдердің корпусының жетіспеушілігі мәселесі қарастырылады. Бұл қазақ тілі сияқты ресурсы төмен тілдер үшін проблема, өйткені қол жетімді корпустар өте аз.

Бұл мақалада қазақ тілінің синонимдік сөздігінің тізімінен үміткер сөзді анықтау және ауыстыру әдісі арқылы синтетикалық корпус жасау тәсілі берілген. Сынақ эксперименттері жүргізілді. Нәтижесінде көрсетілген корпус 3,37 есеге ұлғайды.

Түйін сөздер: корпус, қазақ тілі, синонимдер, лингвистикалық ресурстар.

Introduction

The development of a linguistic corpus is one of the main tasks for processing text data and solving various problems of applied artificial intelligence tasks. The corpus contains special markup, which is additional information about the properties of the texts included in it. The label is the main characteristic of the corpus; it distinguishes the corpus from simple collections (or «libraries») of texts. The richer and more diverse the markup, the higher the scientific and educational value of the corpus [1]. Currently, many languages have their own corpora. This underlines the importance of the case.

The main tasks of the corps include:

1) Providing scientific research of vocabulary and grammar of the language;

2) Providing open access to the case;

3) Provision of a representative building, that is, a sufficiently large building;

4) Providing a balanced corpus, i.e. such a corpus that would reflect the real correlation of genres in the language;

5) The opportunity to get acquainted with the language and its features;

6) The ability to search by morphological parameters;

7) The opportunity to learn a language using word translations.

A corpus is constructed primarily to represent language use in a balanced manner in order to study language empirically on the basis of real data. The role and function of corpora in linguistic analyses can be viewed from different perspectives, depending on the research questions at hand. Lüdeling and Kytö (2008, p. ix) summarize the use of corpora in linguistic analyses for three major purposes: (1) empirical support, (2) frequency information, and (3) meta-information.

The corpus query tools help researchers in finding examples of real language use that are relevant to their questions, that is what they now have as an example is a citation of actual language use rather than the alternative—a made-up example or a sample derived by chance and most often de-contextualized. Providing evidence for language structure and use from corpora is not limited to a specific level of linguistic analysis but works at all levels, from sound to form and to function. The data in corpora are tagged and annotated and thus provide the exact type of sampling that empirically supports the hypotheses. As a repository of real language samples, a corpus query returns citations of language use that had not been envisaged before. Additionally, the empirical nature of corpora makes it possible to replicate the analysis conducted, which is not possible with data based on introspection.The language use captured in linguistic corpora further incorporates “meta” information for its users in terms of major participants or components of acommunication event. These include the gender of the participants, their age as wellas their dialectical background, the medium of the text and its specific genre, among others, all of which provide significant information to a linguist in an analysis of natural language use in context.

When we narrow down the actual corpus linguistic work conducted over the years, we observe that they cover major areas. Meyer (2004) lists these general areas which further include many other subfields of linguistics: Grammatical studies of specific linguistic constructions, lexicography, language variation, historical linguistics, contrastive analysis and translation theory, natural language processing, language acquisition, and language pedagogy.

The ever-growing number of publications and the appearance of special journals in the field clearly underline the increasing importance of corpora in linguistics. It is evident that linguists with different interests will continue to build and use corpora in the future. As before, contributions from neighboring disciplines like computational linguistics and natural language processing research will continue to play a significant role in the future of corpus linguistics. As observed by Sampson (2013), there is currently a rising trend in linguistic analyses to adopt empirical approaches.

Related works

The first large computer corpus is considered to be the Brown Corpus (BC, English Brown Corpus, BC), which was created in the 1960s at Brown University and contained 500 fragments of texts of 2 thousand words each, which were published in English in the USA in 1961. As a result, he set a standard of 1 million word usage for creating representative corpora in other languages [2]. The British National Corpus (BNC from the English British National Corpus) is a corpus of texts of 100 million words containing samples of written and spoken British English from a wide range of sources[1],[3],[4]. The corpus covers British English of the late 20th century, represented by a wide variety of genres, and is conceived as an example of the typical spoken and written British English of that time. [http://www.natcorp.ox.ac.uk/corpus/index.xml]

The National Corpus of the Russian Language with a volume of more than 600 milion word uses can serve as an example of a well-developed corpus. It is divided into such half-corps as: basic, syntactic, newspaper, parallel, educational, dialect, poetic, oral, accentological, multimedia, multipark, historical. The national corpus of the Russian language covers, first of all, the period from the middle of the XVIII to the beginning of the XXI century [1].

The existing corpus of Turkic languages include:

1) The Turkish national Corpus of 50 million word usage, which is a balanced and representative corpus of the modern Turkish language. It consists of samples of textual data in a wide variety of genres, covering a period of 20 years (1990-2009) [3].

2) Bashkir poetic corpus with a volume of more than 1.8 million word usage. It is the second poetry corpus in the world. Its peculiarity lies in the fact that the corpus consists of works by Bashkir poets of the XX and the beginning of the XXI century [4].

3) Tatar National Corpus «Tugan tel» with a volume of more than 26 million word usage. The corpus contains texts of various genres, such as fiction, texts THEMSELVES, texts of official documents, educational literature, scientific publications, etc. [5].

4) The written corpus of the Tatar language with a volume of more than 116 million words, with the number of different word forms - about 1.5 million [6].

5) Almaty corpus of the Kazakh language with a volume of 20 million word usage. It is a linguistically representative corpus.

The following is an overview of the existing marked-up corpora for the Kazakh language:

1) NCKL (National Corpus of Kazakh Language) with a volume of 200-250 million words is one of the possible versions of the National Corpus of the Kazakh language as a reference system based on an extensive fund of marked texts of literary Kazakh, the state language of the Republic of Kazakhstan. Main characteristics of NCKL [7]:

1. a convenient tool for scientific research, development of textbooks and workbooks of the Kazakh language, self-study of the Kazakh language, providing most of the word forms with lexico-morphological analysis and Russian-English translation equivalents;
2. annotated corpus with grammatical and bibliographic markings;
3. linguistic representative corpus;
4. the case, which is in the public domain;
5. the balance of the corpus, which includes literary, scientific, journalistic texts.
6. The texts are composed of 5 styles of the Kazakh language (artistic style, scientific style, journalistic style, paper style, speech style).
7. The user can search by word, word form (word variant) and see a list of sentences that use the searched word and their source.

2) KLC (Kazakh Language Corpus) is a large-scale corpus containing more than 135 million words and containing five main stylistic genres (areas): literary, journalistic, official, scientific and informal. KLC contains an annotated corpus for reading and speech (RSC), which includes audio recordings of words, phrases, sentences (from all genres), news articles and excerpts from books that have been carefully selected from the primary part of the corpus. Each audio file is accompanied by a label file and a corresponding text transcript. In addition, some of the transcripts were grammatically annotated, that is, part of the data has several levels of an-notation: audio (word segmentation), lexical and morphosyntactic. In total, the RSC contains 10 GB or more than 40 hours of speech [8].

In recent years, the Kazakh language has become an object of major concern to linguists. Several scholars have attempted to make up a corpus of the Kazakh language that would be of great significance in natural language processing tasks, including information retrieval and machine translation. One of the first attempts to compile a Kazakh corpus was made in 2013; it contained more than 135 million words belonging to five different stylistic genres. Rakhimova and Zhumanov (Rakhimova, D., & Zhumanov, Z. (2017). Complex technology of machine translation resources extension for the Kazakh language. Advanced Topics in

Intelligent Information and Database Systems, Studies in Computational Intelligence, 710, 297–307.) created a multilingual parallel corpus of general words in the Kazakh, Russian, and English languages using the Bitextor application and demonstrated how dictionaries could be enriched with new words without special linguistic knowledge.

The above mentioned developments were implemented with the help of modern tools and technologies in IT. When creating a marked-up corpus, special procedures and functions performed on the text are used, such as text pre-processing, text collection and markup (tagging), etc. But unfortunately, these resources are not available to researchers as source material. The above platforms are not user-defined and do not allow you to get full resources. Collecting and processing data for the development of enclosures takes a lot of time and effort. There is also an acute problem of lack of electronic linguistic data for low-resource languages, which include a group of Turkic languages (Kazakh, Kyrgyz, Tatar, Uzbek, etc.) with complex agglunative morphology and syntactic structural form.

Methodology

The authors present an approach for the development of a synthetic corpus for the Kazakh language. This approach consists of two main parts:

Part 1 - Preprocessing and analysis of the syntactic structure of sentences based on the linguistic rules of the Kazakh language. In this part, text preprocessing is performed. The syntactic analysis of the text allows you to identify candidate words for which alternative synonyms will be determined.

Part 2- For candidate words , a set of synonyms is determined from the catalog and further probability subtraction is performed . Further, the work of this approach will be described in more detail.

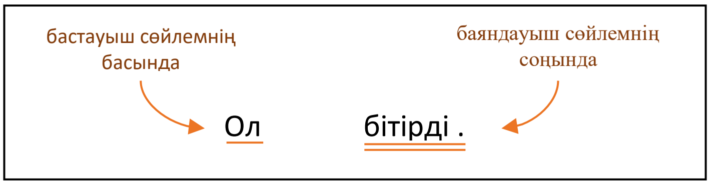
Part 1

The correct structure of the general sentence and the correct transfer of meaning directly depend on the correct location of the members of the sentence in the sentence. How the order of words in a sentence should be is not a matter of the will of the writer or the speaker. Each language has its own internal laws and rules about it. When making sentences, it is better to take them into account and use each word in its place.

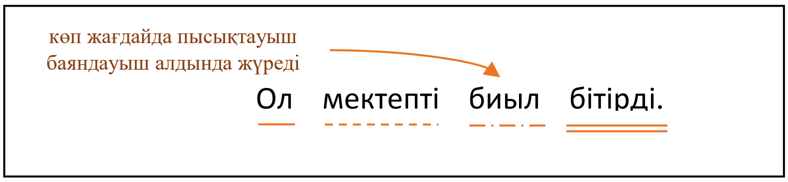
As a general rule, the verb is at the end of the sentence, the subject precedes it, the determinant precedes the defining word, the complement and the finisher precede the words to which they relate.

If we show it using examples.

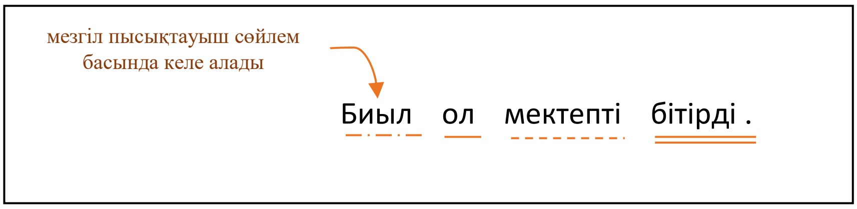
Let's consider a sentence consisting of the initial and the narrator:



We can place the object and adverbial modifier between subject and verb. Adverbial modifier is in fron of the verb in most cases. Lets add adverbial modifier and object to the initial sentence

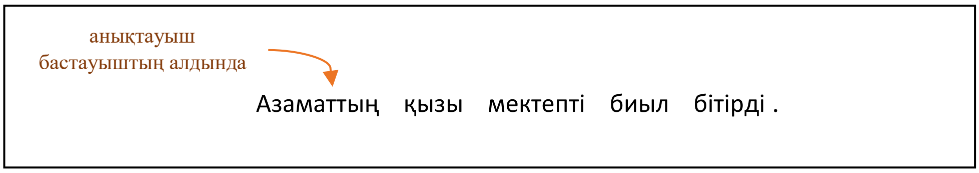


If adverbial modifier means the season, it can be replace the subject in the beginning of the sentence.



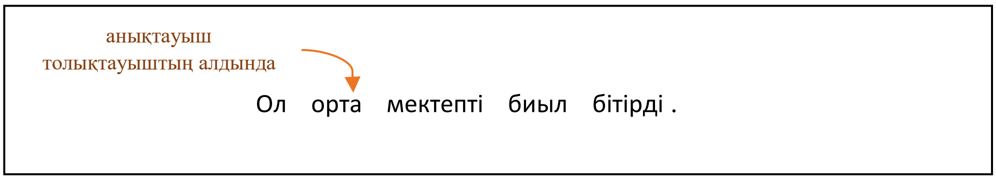
Attribute is placed in front of the member of sentence which is being described. If it’s describing the subject, then it goes in front of the subject, same as for objects and adverbial modifiers.

Let’s attribute the subject. Firstly, replace “He/She” by the word “Daughter”, then add attribute for the word “Girl” itself



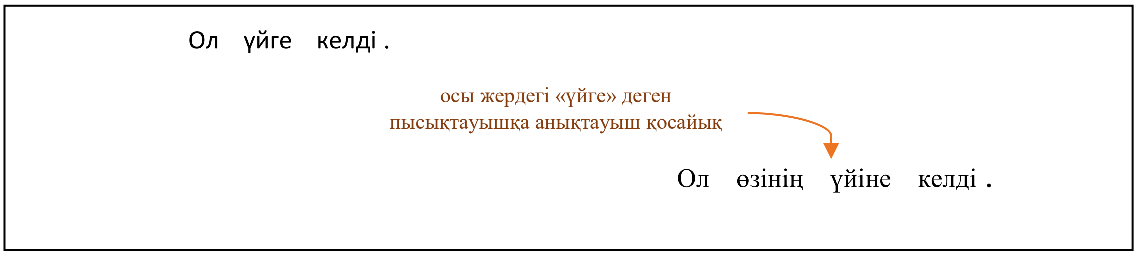
Whose daughter: Azamat’s daughter.

Now lets find the object.



What school: middle school.

We can put an attribute to the adverbial modifier if it’s a noun. For example:



Subject can be represented as a noun, adjective, numeral, pronoun, dead-end verb, interjection, adverb and imitative word in the sentence. For example,

Numeral: **Adults (who?)** sat on the bench

Dead-end verb: **Polite speech (what?)**- a sign of decency

Imitative word: **Dabyr-dubur (what?)** began to be heard up close.

The verb can work as verb, numeral, noun, adjective, pronoun, auxiliary verb and imitative verb. For example:

Noun: Today's youth is a happy **youth**.

Auxiliary verb and imitative word: Кенет әлдене **тарс етті**. Қуанғаннан жүзі **күлмің-күлмің етеді**. Suddenly there was **a bang**. He has a lot of **fun and fun**.

The object can be represented as a noun or pronoun, numeral, adjective, dead-end verb, adverb that means a noun. For example:

Noun: A person who respects his **parents** (**who?**) will not be a fund.

The numeral: One hundred-**twenty-four (what?)** is not divided without a residue.

Parts of speech that can be a adverbial modifier: adverb, adjective, the numeral, noun, adverbial verb. For example:

Adjective: The speaker spoke for a **long time**.

Gerund: Aigerim smiled at this word **with a smile**.

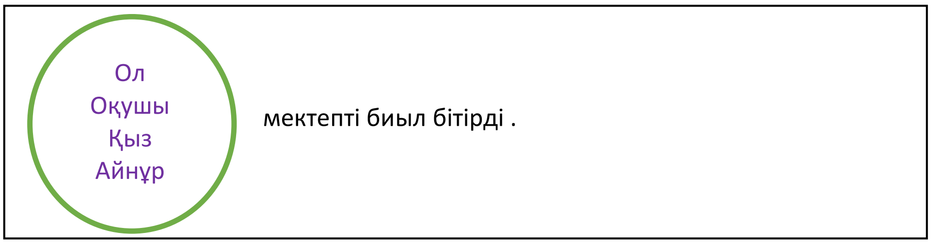
Noun, adjective, numeral, pronoun and participle can work as an attribute in the sentence

The numeral: There are **seven** wonders that have gone down in history on the land of Kazakhstan.

Pronoun: **This** book educates a person to virtue and common sense.

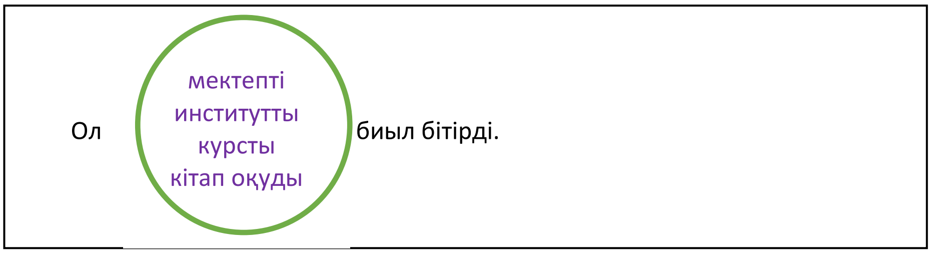
Once we know parts of speech of the members of sentence, we can generate different sentences by substituting them within the same sentence.

Generating new sentences by changing the subject:



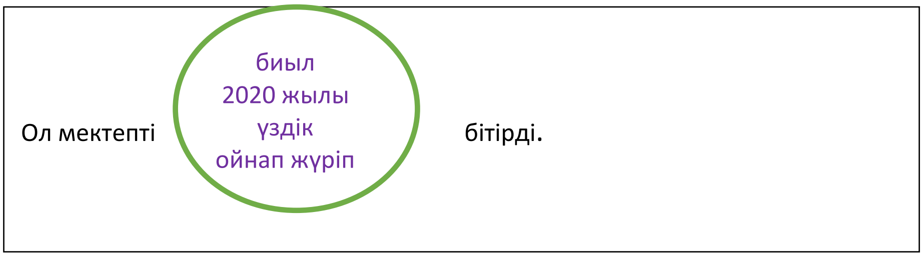
Ол
Оқушы
Қыз
Айнұр
мектепті биыл бітірді .

Generating new sentences by changing the object:



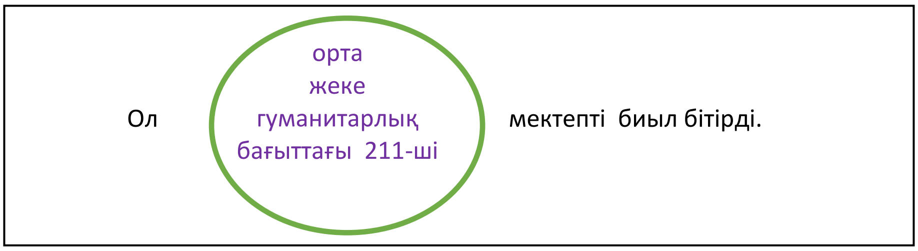
Ол
мектепті
институтты
курсты
кітап оқуды
биыл бітірді.

Generating new sentences by changing the adverbial modifier:



Ол мектепті
биыл
2020 жылы
үздік
ойнап жүріп
бітірді.

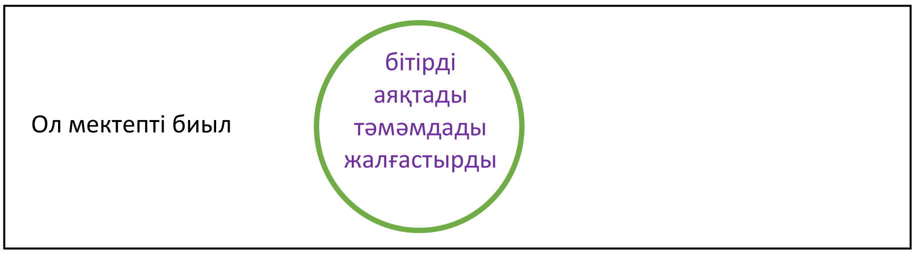
Generating new sentences by changing the attribute:



Ол
орта
жеке
гуманитарлық
бағыттағы 211-ші
мектепті биыл бітірді.

The verb change depends on the complement that precedes it. Because in this sentence, the word «finished» requires that the complement be in the income clause. If I put another verb» came», it requires the complement to stand in the leopard and eastern declensions.

However, it is precisely by giving synonyms for the verb that we can generate new sentences:



Ол мектепті биыл
бітірді
аяқтады
тәмәмдады
жалғастырды

The basic rules of syntactic structures of sentences for the Kazakh language are presented above. And I take into account the linguistic rules and properties of the Kazakh language, the main parts of speech and candidate words have been identified, which can be replaced with various other words, it is possible to increase the number of correct sentences in the corpus. Next, the compilation of synonyms for candidate words will be performed.

Part 2

It is necessary to create synonyms for the Kazakh language. Synonyms when creating a catalog, an online dictionary was used. The thesaurus.com from the site dictionary.com recommended by the world's largest and most reliable free online thesaurus.

Dictionary.com is the world's leading online source for definitions, word origins and more tools. Dictionary.com reveals the secrets of the English language for millions of people. Dictionary.com it is a tool that tries to inspire connections, communication, learning, creativity and things in many areas in a world that works with words. Dictionary.com - the world's leading digital dictionary. This dictionary offers millions of English definitions, spellings, sound pronunciations, sample sentences and word origin. The main proprietary source for this site is the Random House Unabridged Dictionary, which is regularly updated by a team of experienced lexicographers and supplemented by reliable, recognized sources, including American Heritage and Harper Collins, to support various language needs. It also offers a translation service, a crossword solution and a lot of editorial content that will be useful for advanced word lovers and English language learners.

For more than 20 years thesaurus.com it has been helping millions of people improve their English proficiency and find a specific word with more than 3 million synonyms and antonyms.

The Beautiful Soup library was used for using synonyms from thesaurus.com Beautiful Soup is a Python library used to extract data from HTML and XML files, or parser for syntactic analysis of HTML/XML files written in the Python programming language [9].

After finding the wrong words in the initial first task in the research work, thesaurus.com is used to find synonyms of English versions of the same identified wrong words . All English synonyms found are translated into Kazakh using Google translator and saved in a file. Then a catalog of translated Kazakh synonyms will be created. Automatic directory creation is described in the following algorithm.

Algorithm. The algorithm for finding synonyms for a word consists of the following steps:

1. *Incoming data:* $w_j^{каз}$ *- translated words in the Kazakh language*

$w_i^{анг}$ *- English word according to the Kazakh language.*

2. *Find* $w_j^{каз}$ *in the synonym catalog.*

3. *If* $w_j^{каз}$ *is in the catalog, show the list of synonyms of* $w_j^{каз}$ *, go to the 8th step, else do the 4th step.*

4. $w_j^{каз}$ *- choosing a synonym with a high probability for replacement (go to the same module).*

5. *Find* $w_i^{анг}$ *from English synonym words. (3,5 million words)*

6. *Translate synonyms of* $w_i^{анг}$ *into Kazakh language by Google Translate*

7. *On the new line, replenishment (expansion) of the catalog with Kazakh synonyms of* $w_j^{каз}$*. Release the list of synonyms*

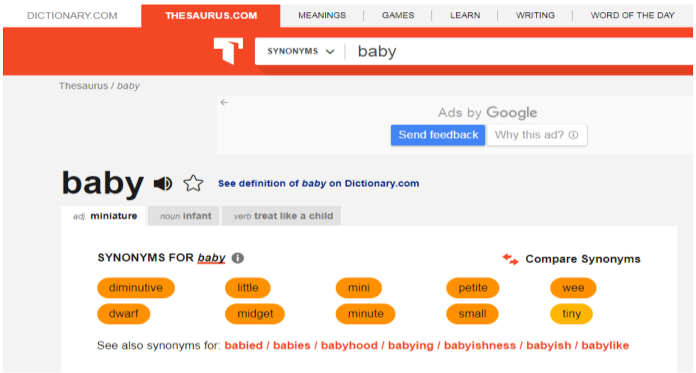
8. *End.*

Following these steps, a list of translated synonyms is created. Then the directory will be created automatically.

The algorithm for creating an automated catalog of synonyms for words is carried out as follows:

- After finding the translated words from the target sentence T, the English version of this word is taken from the S sentence. Then a list of synonyms for the word S* is created from the site [thesaurus.com] and recorded in a file.
- Then, using Google machine translation, each synonym was translated and a list of synonyms in the Kazakh language was compiled.
- In the next step, the found synonyms in the Kazakh language are copied to the directory in the form of an entry with a list of all possible alternatives.

In this way, these steps are applied to the found words. Applicable online dictionary on thesaurus.com can be seen in the following figure 1.



baby

Thesaurus / *baby*

baby

See definition of *baby* on Dictionary.com

adj. **miniature** *noun* **infant** *verb* **treat like a child**

SYNONYMS FOR *baby*

diminutive, little, mini, petite, wee, dwarf, midget, minute, small, tiny

See also synonyms for: **babied / babies / babyhood / babying / babyishness / babyish / babylike**

Figure 1. Synonyms of word "baby" in thesaurus.com

The words shown in Figure1 are all possible equivalent words, for example this catalog shows English versions of the wrong words. The *thesaurus.com* site was used in software development using the Beautiful Soup library. For example, the mistranslated word *baby* has synonyms such as:

Baby: *diminutive, dwarf, little, midget, mini, minute, petite, small, wee, tiny* and etc.

Found incorrectly translated synonyms in English words are translated into Kazakh using Google translator and recorded in the catalog. Each new word in the Kazakh language and its equivalents are recorded in a new line of the catalog. The short snippet type of the directory entry looks like this:

1. Less, not enough, slight, limited

2. …

40. Respectful, good, justice, kind, charming, favorable, wonderful, unusual

41. baby, dwarf, very small, small … etc.

The vocabulary used in the catalog can be seen in the following table 1.

Table 1. Information about the catalog. Catalog consist of kazakh and English synonyms

Vocabulary in English	*Vocabulary in Kazakh*
More than 3.5 million	*More than 7000*

In fact, this is a catalog of synonyms in the Kazakh language, the volume of which is 1000 lines. The directory is indexed by the first word. Each line of the catalog contains synonyms for the wrong words, that is, there are fifteen to Seventeen synonyms in each line of the catalog with 1000 lines of incorrectly translated words. Each new error word and synonyms found are saved to a new line in the directory. If the desired incorrect word is found when searching in the directory, then to use tables related to it, a table with the same number is opened, taking into account the line number of the wrong word in question. For example, the words *polite, good, fair, kind, charming, acceptable, wonderful, unusual* in the above catalog list are located on the fortieth line, that is, the required table number associated with these words is designated as fortieth [10].

Next, using the maximum entropy method, we calculate the probability of the best element from the list of synonyms for various parts of speech for the Kazakh language.

In the maximum entropy model:

$$P(c\,|\,x) = \frac{1}{Z}\exp\sum_j \lambda_j f_j \tag{1}$$

Here Z is the normalizing factor.

Equation that calculates the probability of *X* in Class c given at maximum entropy (equation 2):

$$P_i^e(c\,|\,x) = \frac{\exp\left(\sum_{j=1}^{N^e} \lambda_{ij}^e f_{ij}^e(c,x)\right)}{\sum_{c'\in S^e} \exp\left(\sum_{j=1}^{N} \lambda_{ij}^e f_{ij}^e(c',x)\right)} \tag{2}$$

Where,

$$f_{ij}^{e}(c,x)=\begin{cases} 1, \text{ if } x=z_j^e \,\&\, c=s_i^e \ (x\ c\text{-ның тіркесі}) \\ 0, \text{ кері жағдайда} \end{cases} \tag{3}$$

$\omega^e -$ *polynomial word,*

c $-$ *class of synonyms,*

$z_j^e -$ *j-th property word of Class c (* S_i^e *)*

$x -$ *word under study,*

$N^e -$ *number of features for* ω^e

$\lambda_{ij}^e -$ *the weight of the feature* f_{ij}^e ,

$S^e -$ *set of synonyms for* ω^e

The calculated results according to formula 3 can be seen in table 2.

Table 2. ω^e *word frequency* ω^l

							…	
	ω^e	$z^e{}_1$	$z^e{}_2$	$z^e{}_3$	$z^e{}_4$	$z^e{}_5$	…	$z^e_{N^e}$
S^e	s^e s_1 f_{1j} s_1 g_{1j}	*0* *0*	*1* *5*	*0* *0*	*1* *6*	*0* *0*	…	*0* *0*
	s_2 f_{2j} s^e s_2 g_{2j}	*1* *1*	*0* *0*	*0* *0*	*0* *0*	*1* *6*	…	*0* *0*
							…	

Where g_{ij} – frequency. For the synonym $s_1{}^e$ of the polynomial word ω^e , the frequency g and the property f were used and the weight λ^e was calculated. For this, the data in Table 2 were used:

$$\lambda_{12}=\frac{g_{12}f_{12}}{\sum_{j=1}^{N^e} g_{1j}f_{1j}}=5/11=0,45 \qquad \lambda_{14}=\frac{s_1 f_4}{\sum_{j=1}^{N^e} g_{1j}f_{1j}}=6/11=0,54$$

where, g_{ij} - the frequency of word propery z_j for synonym s_i (phrase frequency for s_i).

The calculation of weight λ^e of polynomial word ω^e and synonym word s_2^e shown below:

$$\lambda_{21}=\frac{g_{21}f_{21}}{\sum_{j=1}^{N^e} g_{2j}f_{2j}}=1/7=0,14 \qquad \lambda_{25}=\frac{g_{25}f_{25}}{\sum_{j=1}^{N^e} g_{2j}f_{2j}}=6/7=0,85$$

The full result with calculation can be seen in Table 3.

Table 3. Calculation of the semantic cube for the word ω^e

ω^e	$z^e{}_1$	$z^e{}_2$	$z^e{}_3$	$z^e{}_4$	$z^e{}_5$	…	$z^e_{N^e}$
s_2 f_{2j} s_2 λ_{2j}	1 0,14	0 0	0 0	0 0	1 0,85	…	0 0
			…				

The calculation of the probability of a class of synonyms by the formula (3) will be as follows:

$$P(s_1 \mid x) = \frac{e^{0,45} * e^{0,54}}{e^{0,45} * e^{0,54} + e^{0,14} * e^{0,85}} = \frac{0,243}{0,243 + 0,119} = \frac{0,243}{0,362} \approx 0.67$$

$$P(s_2 \mid x) = \frac{e^{0,14} * e^{0,85}}{e^{0,45} * e^{0,54} + e^{0,14} * e^{0,85}} = \frac{0,119}{0,243 + 0,119} = \frac{0,119}{0,362} \approx 0,33$$

Then apply the classification formula:

$$\hat{c} = \underset{c \in C}{\arg\max}\, P(c \mid x) \tag{4}$$

Hence, ynonym class s_i is chosen where $P(s_i \mid x)$ is maximum. For example, chosen synonym appears to be s_1, because the value $P(s_i \mid x)$ shows the biggest (maximum) [11-15].

Results

To implement the work and test the approach, it was initially necessary to take a corpus of the Kazakh language in the amount of 120 thousand sentences https://github.com/NLP-KazNU.

First of all, the text was preprocessed:

Text preprocessing translates text in natural language into a format convenient for further work. Preprocessing consists of various stages, which may differ depending on the task and the implementation of subtasks:

- Translation of all letters in the text to lowercase or uppercase;
- Deleting digits (numbers) or replacing them with a text equivalent (regular expressions are usually used);
- Removing punctuation. It is usually implemented as removing characters from a predefined set from the text;
- Removing whitespaces (whitespaces);

Partial markup- morphological and syntactic analysis of the text is performed using the platform (https://github.com/apertium/apertium-eng-kaz).

Next, candidate words are identified for each sentence, in particular, these are parts of speech - nouns, adjectives, pronouns and verbs. Next, a list of synonyms is compiled for a certain word and for each element we subtract the probability of the best element from the list of synonyms for various parts of speech for the Kazakh language. Further, the elements obtained from it were automatically substituted into the text. Using this approach, a syntactic corpus with a volume of 405,612 sentences was obtained.

Conclusion

Based on the results of this work, the following results were obtained: Corpus systems for low-resource Turkic languages were analyzed. Based on the linguistic properties of the Kazakh language, candidate words for replacement have been identified. An algorithm for determining synonyms for the Kazakh language has been developed. An automated catalog of synonyms for the Kazakh language has been developed, in which there are more than 7 thousand entries. From the resulting array of synonyms, words with the highest probabilities for various parts of speech for the Kazakh language were calculated. Test experiments were conducted. As a result, the specified case with a volume of 120 thousand sentences was increased 3.37 times to 405 thousand sentences. This approach is convenient because it does not require large resources and can be adapted for other various low-resource Turkic languages. That will allow you to get high-quality electronic enclosures.

Acknowledgments

This research was performed and financed by the grant Project IRN AP 09259556 and AP08052421of Ministry of Science and Higher Education of the Republic of Kazakhstan.

References:

1 Mohamed, S.A., Elsayed, A.A., Hassan, Y.F., Abdou M.A. Neural machine translation: past, present, and future. Neural Comput & Applic 33, 15919–15931, (2021).

2 Tukeyev, U., Karibayeva, A., Zhumanov Z.: Morphological segmentation method for Turkic language neural machine translation. Cogent Engineering, 7(1), 1-16 (2020).

3 Dabre, R., Chu, Ch., Kunchukuttan A.: A Survey of Multilingual Neural Machine Translation. ACM Comput. Surv, 53(5), pp. 1-38 (2020).

4 Sennrich, R., Haddow, B., Birch A.: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1715-1725 (2016).

5 Zhixing, T., Shuo, W., Zonghan Y., Gang Ch., Xuancheng H., Maosong S., Yang L. Neural Machine Translation: A Review of Methods, Resources, and Tools. Computation and language, pp. 1-20 (2020).

6 do Carmo, F., Shterionov, D., Moorkens, J., Wagner, J., Hossari, M., Paquin, E., Schmidtke, D., Groves, D., Way, A. A review of the state-of-the-art in automatic post-editing. Machine Translation, 35, 101-143 (2021).

7 Pal, S., Naskar S.K., Vela, M., van Genabith, J.: A Neural Network based Approach to Automatic Post-Editing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 281–286. Association for Computational Linguistics, Berlin, Germany (2016).

8 Vu, T., Haffari, G.: Automatic Post-Editing of Machine Translation: A Neural Programmer-Interpreter Approach. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3048–3053. Association for Computational Linguistics, Brussels, Belgium (2018).

9 https://www.crummy.com/ software/BeautifulSoup/]

10 Tebbifakhr, A., Agrawal, R., Negri, M., Turchi, M.: Multi-source Transformer for Automatic Post-Editing. Italian Journal of Computational Linguistics 5(1), 89-103 (2019).

11 Abdulmumin, I., Galadanci, B., Isa, A., Kakudi, H., Sinan, I. A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data. Engineering Letters, 29(4), pp. 1478-1493 (2021).

12 Mengtao, S., Wang, H., Pasquine, M., Hameed, I.A. Machine Translation in Low-Resource Languages by an Adversarial Neural Network. Applied Sciences, 11(22), 10860 (2021).

13 Zhumanov Z., Madiyeva A., Rakhimova D.: New Kazakh Parallel Text Corpora with On-line Access. In: Nguyen, N., Papadopoulos, G., Jędrzejowicz, P., Trawiński, B., Vossen, G. (eds) Computational Collective Intelligence ICCCI 2017, Lecture Notes in Computer Science, vol. 10449. pp. 501-508. Springer Professional (2017).

14 Abdulmumin I., Galadanci B.S., Isa A.: Enhanced Back-Translation for Low Resource Neural Machine Translation Using Self-training. In: Misra S., Muhammad-Bello B. (eds) Information and Communication Technology and Applications. ICTA 2020. Communications in Computer and Information Science, vol. 1350, pp. 355-371. Springer, Cham (2021).

15 Lee, W., Jung, B., Shin, J., Lee, J.: Adaptation of Backtranslation to Automatic Post-Editing for Synthetic Data Generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 3685-3691 (2021).