Word2Vec Model Analysis for Semantic and Morphologic Similarities in Turkish Words

Larysa Savytska[a], M. Turgut Sübay[b], Nataliya Vnukova[a], Iryna Bezugla[a] and Vasyl Pyvovarov[c]

[a] *Simon Kuznets Kharkiv National University of Economics, Nauky av. 9a, Kharkiv, 61166, Ukraine*
[b] *Piramit Danismanlik A.S., İstanbul, Kadıköy, Turkey*
[c] *Yaroslav Mudryi National Law University, Pushkinska str. 77, Kharkiv, 61024, Ukraine*

Abstract

The study presents the calculation of the similarity between words in Turkish language by using word representation techniques. Word2Vec is a model used to represent words into vector form. The model is formed using articles from Wikipedia dump Turkish service as the corpus and then Cosine Similarity calculation method is used to determine the similarity value. The open-source Python programming language and Gensim library are used to obtain high quality word vectors with Word2Vec and calculate the cosine similarity of the vectors. Continuous Bag-of-words (CBOW) algorithm is used to train high quality word vectors. The cosine similarity values in the results are derived from the weight (dimension values) of the vector dimensions. The Window size 10 and 300 vector dimension configurations are taken. Increasing the number of cycles contributes to the vectors getting more accurate values. The corpus is trained in five cycles (EPOCH) with the same parameters. The Turkish corpus contains more than one hundred and sixty one million words. The dictionary of words (unique words), obtained from the corpus, is more than three hundred and sixty-seven thousand. Such a big data gives an opportunity to conduct high quality semantic and morphologic analysis and arithmetic operations of the word vectors.

Keywords

NLP, Word2Vec, word vectors, cosine similarity, word embedding, semantic relations, formal (structural) relations, Turkish language

1. Introduction

In today's world automatic analysis is constantly being developed to meet the increasing industrial needs. Thanks to automatic analysis, information access, identifying people or objects from photographs, distinguishing the advertising contents of e-mails, analyzing sentiments in correspondence, translation between languages and many similar needs can be met. Natural Language Processing (NLP) is a general field of computer science, artificial intelligence (AI) and mathematical linguistics [1]. English mathematician Alan Turing asked a question "Can machines think like a human?" This proposal opened the idea of AI and led to discussion [2] that AI technologies can learn like humans and communicate with people.

NLP studies the problems of computer analysis and natural language synthesis. For AI, analysis means understanding the language and synthesis means generating intelligent text. There are different approaches to NLP such as statistical, linguistic, symbolic and etc.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: larisa-savickaya@hotmail.com (L. Savytska); m.turguts@hotmail.com (M. T. Sübay); vnn@hneu.net (N. Vnukova); iryna.bezugla@hneu.net (I. Bezugla); v.pyvovarov@ukr.net (V. Pyvovarov);
ORCID: 0000-0002-9158-6304 (L. Savytska);0000-0002-2967-694X (M. T. Sübay); 0000-0002-1354-4838 (N. Vnukova); 0000-0002-6285-2060 (I. Bezugla); 0000-0001-9642-3611 (V. Pyvovarov)

The linguistic approach to natural language processing consists of four levels: graphematic, morphological, syntactic and semantic [3, 4]. The first level is to identify the individual elements of the text / document, such as sections, paragraphs, sentences, etc. The second level is to determine the morphological characteristics of each word. The third level is responsible for determining the syntactic dependence of words in sentences. The last level is related to the semantic understanding of the text, including developments in the field of artificial intelligence [5].

The clusters and sub-clusters between the vectors obtained by machine learning are parallel in terms of the syntax of words, semantic and formal (structural) relations [6]. These relationships between words find wide application especially in industrial areas such as search engines. In natural language processing, the matching of words with vectors (finding word vectors) techniques are called Word Embeddings (WE) [7, 8]. WE is the collective name for a set of language modelling and features of learning techniques in NLP, where words or phrases are represented in the form of real number vectors [9]. Conceptually, WE involve mathematical formulas. The models used in word embeddings are varied, one is the Word2Vec model. The Word2Vec represents words into vector based on several features they have such as windows size and vector dimensions. Word embedding proved to be an incredibly important method for NLP tasks in recent years, enabling various machine learning models that rely on vector representation as input to enjoy richer representations of text input. These representations preserve more semantic and syntactic information on words, leading to improved performance in almost every imaginable NLP task [10]. One of the reasons for developing word embedding techniques is that it shortens the machine learning training time. The shortening of the training period provides the opportunity to work with more vector dimensions and larger collections in practice. Being able to train machine learning with large corpus and more vector dimensions is shown among the important factors affecting the correct representation of words by vectors.

The research using machine technology Word2Vec is of great practical importance to computerise many areas of linguistic analysis such as

- identifying semantic similarity of words and phrases
- automatic clustering of words according to the degree of their semantic similarity
- automatic generation of thesaurus and bilingual dictionaries
- expanding queries due to associative connections
- constructing semantic maps of various subject areas and so on.

2. Related Works

Learning vector representations of words using neural networks has generated a strong enthusiasm in the NLP research community. In particular, many contributions were proposed after the work of Tomas Mikolov and his team [11, 12, 13] on training word embeddings. The main reasons for this strong interest are: the proposal of a simple and efficient neural architecture to learn word vector representations, the availability of an open source tool Word2Vec and the rapid structuring of a user community. Later on, several contributions have extended the work of T. Mikolov on word vectors to phrases (sequences of words) [12, 14, 15, 16] and T. Luong to bilingual representations [17]. All these vector representations capture similarities between words, phrases or sentences at different levels (morphological, semantic).

T. Mikolov and his team conducted the research on training word embeddings by using Word2Vec model representation from English corpus [11, 12, 13, 14]. We did the research on training word embeddings by using Word2Vec model representation from Ukrainian corpus [18]. D. Chaplinskyi used LexVec, Word2Vec and GloVe model representations to train Ukrainian word embeddings [19]. A. Romanyuk suggests training Ukrainian word embeddings with Word2Vec, FastText and MUSE model representations [20]. V. Vysotska did comparative analysis for English and Ukrainian texts processing based on semantics and syntax approaches [21].

Thus, there are new challenges to conduct the research on training word embeddings by using Word2Vec model representation from different corpuses [22, 23]. In this article we are training Turkish word embeddings by using Word2Vec model representation. The research is examining the semantic clustering of Turkish word vectors, semantic relations between words at arithmetic operations of Turkish word vectors, formal clustering of Turkish word vectors and formal relations between words at arithmetic operations of Turkish word vectors.

3. Methodology and Materials

The open-source Python programming language and Gensim library is used to obtain high quality word vectors with word representation technique Word2Vec model and calculate the cosine similarity of the vectors [24, 25, 26, 27]. Continuous Bag-of-words (CBOW) algorithm is used to train high quality word vectors. The cosine similarity values in the results are derived from the weight (dimension values) of the vector dimensions. The Window size 10 and 300 vector dimension configurations are taken. Increasing the number of cycles contributes to the vectors getting more accurate values. The corpus is trained in five cycles (EPOCH) with the same parameters.

The operation steps are the following:

The latest version of the Python programming language is downloaded and installed.

1. Two libraries NumPy and Scipy are installed using the Python library installer (pip).

2. The "Gensim" library is installed using the Python library installer (pip).

3. To write code in Python, the command window can be used by line or “Pycharm” etc. or an editor can be used.

4. If the corpus is related to the subject or they have general content, the resource on the GitHub site [28] can be used and/or the corpus can be organized using different methods. If the corpus is not available, there is an internet address to access ready-made corpus for this resource. Using the Wikimedia dump Turkish service [29], the corpus can be edited using the library named "corpora.wikicorpus"[25] in "Gensim".

5."models.word2vec" or "models.keyedvectors" libraries available in "Gensim" are used in order to obtain word vectors using Word2Vecmodel.

6. The "keyedvectors" library in "Gensim" is used to calculate the cosine similarity of the vectors.

The Turkish corpus is obtained from Wikipedia dump Turkish service [29] and a source [26]. To clean the corpus, the capital letters were changed to lower letters.

Turkish letters from capital to lower letters mapping:

```
   lowerMap = {ord(u'A'): u'a',ord(u'A'): u'a',ord(u'B'): u'b',ord(u'C'):
u'c',ord(u'Ç'): u'ç',ord(u'D'): u'd',ord(u'E'): u'e',ord(u'F'):
u'f',ord(u'G'): u'g',ord(u'Ğ'): u'ğ',ord(u'H'): u'h',ord(u'I'):
u'ı',ord(u'İ'): u'i',ord(u'J'): u'j',ord(u'K'): u'k',ord(u'L'):
u'l',ord(u'M'): u'm',ord(u'N'): u'n',ord(u'O'): u'o',ord(u'Ö'):
u'ö',ord(u'P'): u'p',ord(u'R'): u'r',ord(u'S'): u's',ord(u'Ş'):
u'ş',ord(u'T'): u't',ord(u'U'): u'u',ord(u'Ü'): u'ü',ord(u'V'):
u'v',ord(u'Y'): u'y',ord(u'Z'): u'z'}
```

Python code example to get the corpus from Wikipedia dump Turkish service is the following:

```
   from __future__ import print_function
   import os.path
   import sys
   from gensim.corpora import WikiCorpus
   import xml.etree.ElementTree as etree
   import warnings
   import logging
   import string
   from gensim import utils

   def tokenize_tr(content,token_min_len=2,token_max_len=50,lower=True):
       if lower:
             lowerMap = {ord(u'A'): u'a',ord(u'A'): u'a',ord(u'B'):
u'b',ord(u'C'): u'c',ord(u'Ç'): u'ç',ord(u'D'): u'd',ord(u'E'):
u'e',ord(u'F'): u'f',ord(u'G'): u'g',ord(u'Ğ'): u'ğ',ord(u'H'):
u'h',ord(u'I'): u'ı',ord(u'İ'): u'i',ord(u'J'): u'j',ord(u'K'):
```

```
u'k',ord(u'L'): u'l',ord(u'M'): u'm',ord(u'N'): u'n',ord(u'O'):
u'o',ord(u'Ö'): u'ö',ord(u'P'): u'p',ord(u'R'): u'r',ord(u'S'):
u's',ord(u'Ş'): u'ş',ord(u'T'): u't',ord(u'U'): u'u',ord(u'Ü'):
u'ü',ord(u'V'): u'v',ord(u'Y'): u'y',ord(u'Z'): u'z'}
            content = content.translate(lowerMap)
      return [
      utils.to_unicode(token) for token in utils.tokenize(content,
lower=False, errors='ignore')
      if token_min_len <= len(token) <= token_max_len and not
token.startswith('_')
      ]

  if __name__ == '__main__':

      if len(sys.argv) < 3:
            print("Example command: python3 preprocess.py trwiki-
articles.xml.bz2 outPutWiki.txt")
            sys.exit()

      logging.basicConfig(level=logging.INFO,
  format='%(asctime)s %(levelname)s %(message)s')


      inputFile = sys.argv[1]
      outputFile = sys.argv[2]

      wiki = WikiCorpus(inputFile, lemmatize=False,tokenizer_func =
tokenize_tr)
      logging.info("Wikipedia dump is opened.")
      output = open(outputFile,"w",encoding="utf-8")
      logging.info("Output file is created.")

      i = 0
      for text in wiki.get_texts():
            output.write("".join(text)+"\n")
            i+=1
            if (i % 10000 == 0):
                  logging.info("Saved " +str(i) + " articles.")

      output.close()
```

Not: Because of page regulations some Python code indentations maybe lost

The Turkish corpus contains more than one hundred and sixty one million words. The dictionary of words (unique words), obtained from the corpus, is more than three hundred and sixty-seven thousand. Such a big data gives an opportunity to conduct high quality semantic and morphologic analysis and arithmetic operations of the word vectors.

Word vectors training Python code example for "Gensim" library is the following:

```
from __future__ import print_function
import logging
import sys
import multiprocessing

from gensim.models import Word2Vec
from gensim.models.word2vec import LineSentence

if __name__ == '__main__':
  if len(sys.argv) < 3:
  print("Please provide two arguments, first one is path to the revised
corpus, second one is path to the output file for model.")
          print("Example command: python3 word2vec.py wiki.tr.txt
trmodel")
          sys.exit()

  inputFile = sys.argv[1]
  outputFile = sys.argv[2]
  logging.basicConfig(level=logging.INFO,
  format='%(asctime)s %(levelname)s %(message)s')
  model = Word2Vec(LineSentence(inputFile), size=300, window=5,
min_count=10, workers=multiprocessing.cpu_count())
  model.wv.save_word2vec_format(outputFile, binary=True)

  Not: Because of page regulations some Python code indentations maybe
lost
```

4. Turkish Corpus trained using Word2Vec model: Experiment and Results

4.1. Semantic clustering of Turkish word vectors

Word vectors obtained from the general content Turkish corpus using Word2Vec model are clustered and related in terms of semantic relations of Turkish words.

The first example is the word "Elma". The first five word vectors with the closest cosine similarity to ('elma') vector are shown below.

```
  [('çilek', 0.7261281609535217),
  ('vişne', 0.6900818943977356),
  ('armut', 0.6884721517562866),
  ('dut', 0.6787133812904358),
  ('şeftali', 0.6731953024864197)]
```

The word "Elma" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun, botanical Rose; a tree (Pyrusmalus) with pink or white flowers.

2. Noun, the bark of the tree is bright, hard; red, yellow and green in colour; pleasant smell; sour or sweet taste; crisp texture, stone fruit [30, 31].

Among the vectors obtained from the Turkish corpus, the vector ('çilek') is the closest cosine vector to ('elma'). The word "Çilek" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun, botanical Rosa; a plant; stems creeping, flowers white.

2. Noun, fragrant; pink, red coloured fruit [30, 31].

These vectors are clustered together referring to semantic relations between words they belong to.

Among the vectors obtained from the Turkish corpus, the second closest cosine-like vector to ('elma') is ('armut'). The word "Armut" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun, botanical Rose; its flowers are white; it’s a tree (Pyruscommunis) that grows all over Turkey [30, 31].

These vectors are clustered together referring to paradigmatic relationsbetween words they belong to.

The other results, obtained from the Turkish corpus, are vectors clustered together referring to lexical paradigm of the words representing the names of fruit trees, related to the meaning of the word "Elma".

As a result of training the word "İstanbul", the city name, the first five word vectors with the closest cosine similarity to the vector ('istanbul') are shown below.

```
[('ankara', 0.6938591599464417),
('bursa', 0.6174916625022888),
('trabzon', 0.591408371925354),
('üsküdar', 0.581426739692688),
('yenibosna', 0.5711308121681213)]
```

The word "İstanbul" in Turkish Language Academic dictionary of proper names is defined as

1. One of the provinces of Turkey in the Marmara Region [30].

Among the vectors obtained from the Turkish corpus, the vector ('ankara') is first closest cosine-like vector to ('istanbul'). The word "Ankara" in Turkish Language Academic dictionary of proper names is defined as

1. One of the provinces located in the Central Anatolian Region of Turkey, the capital of Turkey [30].

These vectors are clustered together referring to semantic relations between words "İstanbul" and "Ankara", two important cities of Turkey.

As for the vectors such as 'bursa' and 'trabzon', they are clustered together referring to lexical paradigm of the words representing the other important cities names of Turkey. The word vectors ('üsküdar') and ('yenibosna') are clustered together with word vector ('istanbul'), because words "Üsküdar" and "Yenibosna" represent names of two important districts of Istanbul.

As a result of training the word "Ahmet", a proper name, the first five word vectors with the closest cosine similarity to the vector ('ahmet') are shown below.

```
[('osman', 0.6758742332458496),
('muhittin', 0.6753208637237549),
('niyazi', 0.6559439897537231),
('halit', 0.6479822993278503),
('mehmet', 0.6463955044746399)]
```

The word "Ahmet" in Turkish Language Academic dictionary of proper names is defined as

Origin: Arabic. Gender: Male.

1. Praised [30].

Among the vectors obtained from the Turkish corpus, the vector ('osman') is the first closest cosine-like vector to ('ahmet'). The word "Osman" in Turkish Language Academic dictionary of proper names is defined as

Origin: Arabic. Gender: Male.

1. A type of bird or dragon.
2. Saint Mohammed’s son-in-law, the third caliph.
3. Founder and first ruler of the Ottoman Empire [30].

When the vectors similar to the cosine-like vector ('osman') are examined, the vectors belonging to words/proper names representing male names as the word "Osman" are investigated. It is semantic cluster related to the area of the word "Ahmet".

As a result of training the word "Ayşe", a proper name, the first five word vectors with the closest cosine similarity to the vector ('ayşe') are shown below.

```
[('melike', 0.796585202217102),
('cemile', 0.7877158522605896),
('merve', 0.7801972031593323),
('hatice', 0.7799881100654602),
('zeynep', 0.7753742933273315)]
```

The word "Ayşe" in Turkish Language Academic dictionary of proper names is defined as

Origin: Arabic, Gender: Female.

1. Living comfortably and peacefully [30].

Among the vectors obtained from the Turkish corpus, the vector ('melike') is closest cosine-like vector to ('ayşe'). The word "Melike" in Turkish Language Academic dictionary of proper names is defined as

Origin: Arabic, Gender: Female.

1. A female ruler.

2. The sultan's wife [30].

The word "Ayşe" is used as a female name in Turkish. When the vectors similar to the closest cosine vector ('ayşe') are examined, the vectors belonging to words/proper names representing female names as the word "Melike" are investigated. It is a semantic cluster related to the area of the word Ayşe".

The word vectors ('ahmet') and ('ayşe') are in the same semantic cluster related to the proper noun-meaning relationship and differentiated according to gender characteristics.

As a result of training the word "Okul", the first five word vectors with the closest cosine similarity to the vector ('okul') are shown below.

```
[('okulun', 0.7467690110206604),
('ilkokul', 0.6787807941436768),
('dershane', 0.6465392708778381),
('lise', 0.6133290529251099),
('ortaokul', 0.6094698905944824)]
```

The word "Okul"in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun;the place where all kinds of education and training are held collectively [30, 31].

Among the vectors obtained from the Turkish corpus, the vector ('ilkokul') is the second closest cosine-like vector to ('okul'). The word "İlkokul" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun;a four-year school, primary school opened or allowed by the government to provide the basic education and training of girls and boys at the age of compulsory education [30, 31].

These vectors are in a semantic cluster referring to educational place.

The other vectors similar to the cosine-like vector ('okul') are ('dershane'), ('lise'), ('ortaokul'). The word "Dershane" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; classroom.

2. Noun; private institution that gives money to students outside of school [30, 31].

The word "Lise" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; secondary education institution that prepares you for life or higher education with at least four years of education after eight years of primary education.

2. Noun, secondary education institution that prepares you for life or higher education with at least three years of education after three years of secondary school[30, 31].

The word "Ortaokul" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun, generally three-year secondary school that prepares (middle school) students for life on the one hand, and high school on the other, through general education [30, 31].

These vectors are in a semantic cluster referring to educational place belonging to lexical paradigm of the words representing the educational place, related to the meaning of the word "Okul". The first closest cosine-like vector to ('okul') is ('okulun'), obtained from the word "Okul". The word vectors ('okul') and ('okulun') are clustered together representing formal relation. It is formal derivation of the noun root "okul" with the suffix "- in".

According to the results obtained from training the Turkish corpus using Word2vec modal, it is proved that the vectors are clustered in terms of semantic relations of Turkish words.

4.2. Arithmetic operations of word vectors and semantic relations between words

New vectors can be obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Turkish corpus.

The first example is similar to the English example, showed by T. Mikolov [11] obtained from the English corpus when the cosine analogues of the new vector are obtained by adding and subtracting the vectors.

('king') - ('man') + ('woman') = ('queen')

The first five word vectors with the closest cosine similarity to the result vector of ('kral') - ('erkek') + ('kadın') operation are shown below.

```
[('kraliçe', 0.5500485897064209),
('prens', 0.5298552513122559),
('kralın', 0.514844536781311),
('kralı', 0.49624234437942505),
('kraliçenin', 0.46907928586006165)]
```

The result obtained from the Turkish corpus is similar to the result obtained from the English corpus. The vector ('kraliçe') belonging to the word "Kraliçe" is the Turkish equivalent of the word "Queen" and the closest cosine-like vector to the result vector from the operation.

The ('kral') - ('erkek') + ('kadın') operation is the replacement of the gender characteristic in the word "Kral", which expresses nobility. In terms of the word meaning, the result of the process is the word "Kraliçe". It is seen that the word meaning is compatible with the result of adding and subtracting the vectors. The word "Kraliçe" is defined as "the wife of the king or the woman who rules the kingdom" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) [30, 31].

Another example below are the first five word vectors with the closest cosine similarity to the result vector of ('ingiltere') - ('londra') + ('ankara') operation.

```
[('türkiye', 0.6439434885978699),
('kırıkkale', 0.5729399919509888),
('niğde', 0.5030767917633057),
('eskişehir', 0.4853522777557373),
('tbmm', 0.4850592315196991)]
```

The operation ('ingiltere') - ('londra') + ('ankara') is the transaction of the relationship between countries and cities (or their capitals). The vector obtained as a result is ('türkiye'), the first vector among the cosine-like vectors. The process and result vectors are compatible with the result of adding and subtracting vectors.

The first six word vectors with the closest cosine similarity to the result vector of ('finans') - ('para') + ('altın') operation are shown below.

```
[('bankacılık', 0.439474880695343),
('gayrimenkul', 0.4268363118171692),
('kuyumculuk', 0.4161675274372101),
('mücevherat', 0.41351592540740967),
('mücevher', 0.3932022750377655,
('sigortacılık', 0.3760865330696106)]
```

The word "Finans" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA) is defined as

1. Commercial activity to raise funds and capital.

2. A sub-branch of economics that studies the management of money and other assets.

3. Management of money, credit, banking and investments [30, 31].

The word "Para" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA)is defined as

1. Noun; a mean of payment made by paper or metal with the value is written on it. It is printed by the state, cash [30, 31].

The word "Altın" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA)is defined as

1. A precious metal that is used as money or stored by governments in exchange for money due to its scarcity in nature [30, 31].

The closest cosine-like vector obtained from the ('finans') - ('para') + ('altın') operation is ('bankacılık').

The word "Bankacılık" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA) is defined as

1. Noun; all transactions made in the bank.

2. Noun; the job of the banker [30, 31].

The second cosine-like vector obtained from the ('finans') - ('para') + ('altın') operation is ('gayrimenkul').

The word "Gayrimenkul" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA)is defined as

1. Adjective; immovable.

2. Noun; law, house, field, etc. immovable property, real estate [30, 31].

In the sixth row, the closest cosine-like vector obtained from the ('finans') - ('para') + ('altın') operation is ('sigortacılık').

The word "Sigortacılık" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA) is defined as

1. Noun; bilateral connection agreement made with the organization dealing with the business in return for the premium paid in advance to compensate for the future damage if something or someone may encounter in the future [30, 31].

According to the results of operations the word vectors ('bankacılık'), ('gayrimenkul'), ('sigortacılık') are in semantic relations. The process and result vectors are compatible with the result of adding and subtracting vectors.

The first five word vectors with the closest cosine similarity to the result vector of ('spor') - ('futbol') + ('yüzme') operation are shown below.

```
[('olimpik', 0.5659219026565552),
('havuzu', 0.524342954158783),
('sporları', 0.5239308476448059),
('havuzları', 0.5116350650787354),
('binicilik', 0.49981582164764404)]
```

The word "Spor" in the current Turkish Language Academic dictionaryof science and art terms, represented by Turkish Language Association (TLA)is defined as

1. Noun; all the actions performed according to some rules, individually or collectively, with the aim to improve body or mind.

2. Adjective; easy to use [30, 31].

(The meaning of the word related to body movements examined in the process. The meaning of the word related to Plant science and Animal science is not found in the process).

The closest cosine-like vector obtained from the ('spor') - ('futbol') + ('yüzme') operation is ('olimpik').

The word "Olimpik" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Related to the Olympics, with Olympic dimensions [30, 31].

In the second and the fourth rows, the cosine-like vectors obtained from the ('spor') - ('futbol') + ('yüzme') operation are ('havuzu') and ('havuzları'), belonging to the words "Havuzu" and "Havuzları" derived from the word "Havuz" and formed by the suffixes "- u" and "- ları". It is formal derivation of the noun root "havuz" with the suffixes "- u" and "- ları".

The word "Havuz" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; water accumulation, swimming, beautifying the environment, etc.

2. It is generally an open place where the bottom and sides are made of things like marble or concrete and filled with water for swimming purposes [30, 31].

In the third row, the cosine-like vector obtained from the ('spor') - ('futbol') + ('yüzme') operation is ('sporları'), belonging the word "Sporları" derived from the word "Spor" and formed by the suffix "- ları". It is formal derivation of the noun root "spor" with the suffix "- ları".

In the fifth row, the cosine-like vector obtained from the ('spor') - ('futbol') + ('yüzme') operation is ('binicilik').

The word "Binicilik" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; state of being a rider.

2. Noun; horse riding sport [30, 31].

The word vector ('binicilik') is the result of two sport branches displacement in the vector process.

Semantic relations between Turkish words build clusters in the vectors. It is proved that semantic results obtained by addition and subtraction operations on vectors obtained from the English corpus can be also obtained from the Turkish corpus.

4.3. Formal clustering of Turkish word vectors

Word vectors obtained from the general content Turkish corpus using Word2Vec model are clustered in terms of formal (structural) relations of Turkish words according to Turkish-specific suffixes.

Turkish language is an agglutinative language. The general feature of agglutinative languages is that word roots are kept constant, suffixes and inflections with various functions are added to the roots. By adding different suffixes to the roots of the word, new words are derived and the vocabulary of the language is formed in this way. All changes and developments in Turkish are based on root suffix combinations [32]. We should not expect only similar words to come close to each other, as there may be similarities in more than one way. These similarities may also occur according to the suffixes taken in inflected languages. When searching similar words by using word vectors, the words ending with similar suffixes can also be reached [11].

The first word to be examined is the word "Gitmek". The first five word vectors with the closest cosine similarity to vector ('gitmek') are shown below.

```
[('dönmek', 0.7897772192955017),
('yetişmek', 0.7705608606338501),
('götürmek', 0.7535400390625),
('inmek', 0.7440905570983887),
('yerleşmek', 0.7398502230644226)]
```

The word vectors clustered like a cosine are vectors belonging to the verbs in the form of infinitive. The clustering of word vectors is related to the formal feature infinitive suffix "- mek". It is formal derivation of the verb root with the suffix "- mek".

Another example, the first five word vectors with the closest cosine similarity to the vector ('gittim') are shown below.

```
[('gitmiştim', 0.8377403616905212),
('gittiğimde', 0.8276962637901306),
('gidiyordum', 0.7992637753486633),
('gidiyorum', 0.7966102361679077),
('gideceğim', 0.7883756160736084)]
```

The word "Gittim" is derived by taking the past tense singular affix "- tim" to the verb root "git". The word vectors, obtained from the words "Gitmiştim", "Gittiğimde", "Gidiyordum", "Gidiyorum", "Gideceğim" are vectors of inflected words derived by adding the first person singular suffix to the verb root "git". The clustering of word vectors is related to the formal feature the first person singular.

The word "Elma" was discussed while analyzing the semantic relations between vectors. For the word "Elmalı" in the sentence "Elmalı turta severim" the first five word vectors with the closest cosine similarity to the vector ('elmalı') are shown below.

```
[('kumluca', 0.7562471628189087),
('akseki', 0.7351764440536499),
('ibradı', 0.7255643606185913),
('karacaören', 0.7211636304855347),
('akçapınar', 0.7149443626403809)]
```

The word "Elma" in the sentence "Elmalı turta severim" is derived by adding the suffix "- lı", which builds the word "Elma" into an adjective with the word meaning apple fruit. The word "Elmalı" also refers to the district of Antalya city. Among the vectors obtained by training from Turkish corpus the closest cosine-like vectors are ('kumluca'), ('akseki'), ('ibradı'), representing districts of Antalya city. The clustering of word vectors is related to the semantic feature being a district of Antalya city. It takes place according to the semantic relation with the word "Elmalı".

The first five word vectors with the closest cosine similarity to vector ('ağaçlık') are shown below.

```
[('ormanlık', 0.8628451228141785),
('çalılık', 0.7839390635490417),
('sazlık', 0.7809475660324097),
('makilik', 0.7765018939971924),
('otluk', 0.772311270236969)]
```

The word "Ağaçlık" is derived by taking the suffix "- lık", to the verb root "ağaç". The place name is derived from the noun describing the item. The word vectors clustered like a cosine are vectors of inflected words derived by adding the suffixes "- lık", "- lik", "- luk" to the noun root. Clustering of word vectors takes place within the relationship of form and meaning of the word. It is related to the formal derivation of the noun root with suffixes"- lık", "- lik", "- luk" and the semantic feature, building a place name from the item.

The first five word vectors with the closest cosine similarity to the vector ('avukatlık') are shown below.

```
[('muhasebecilik', 0.6621850728988647),
('doktorluk', 0.6428958177566528),
('hakimlik', 0.6376224160194397),
('yargıçlık', 0.635696530342102),
('memurluk', 0.593788206577301)]
```

The word "Avukatlık" is derived by taking the suffix "- lık" to the noun root "avukat". The job name is derived from the noun describing a profession name. The word vectors clustered like a cosine are vectors of inflected words derived by adding the suffixes "- lık", "- lik", "- luk" to the noun root. Clustering of word vectors takes place within the relationship of form and meaning of the word. It is related to the formal derivation of the noun root with suffixes "- lık", "- lik", "- luk" and the semantic feature, building a job name from profession name.

The first five word vectors with the closest cosine similarity to the vector ('temizlik') are shown below.

```
[('temizleme', 0.5787885785102844),
('temizliği', 0.5512673258781433),
('banyo', 0.5476162433624268),
('kumlama', 0.5201424360275269),
('tamirat', 0.5156590342521667)]
```

The word "Temizlik" is derived by taking the suffix "- lık" to the adjective root "temiz". The nounis derived from the adjective. In the first and the second rows, the cosine-like vectors obtained from the vector ('temizlik') are ('temizleme') and ('temizliği').The word "Temizlik" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; state of being clean, purity, chastity, kindness.
2. Noun, the state of standing or keeping clean.
3. Noun; the cleaning job.
4. Noun; (slang) eliminate, destroy, kill [30, 31].

The word "Temizleme" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; the cleaning job.
2. Noun; removing stains and dirt, adhering to surfaces, transferring them into a solution or suspension[30, 31].

The first two word vectors clustered like a cosine are vectors of inflected words derived by adding the suffixes "- leme", "- liği" to the adjective root. Clustering of word vectors takes place within the formal relations.

In the third row, the cosine-like vector obtained from the vector ('temizlik') is ('banyo'). The word "Banyo" in Turkish dictionary of the Turkish Language Association is defined as

1. Noun; the part in buildings, where everything is washed.
2. Noun; bathing in the bathtub [30, 31].

The word vectors ('temizlik') and ('banyo') are in semantic relations.

In the fourth row, the cosine-like vector obtained from the vector ('temizlik') is ('kumlama'). The word "Kumlama" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; sandblasting the surface using air pressure to indicate more the visual difference between the growth rings of pine trees[30, 31].

The word vectors ('temizlik') and ('kumlama') are in semantic relations.

The word vectors obtained from the word "Temizlik" clustered like cosine vectors according to the semantic and/or formal relationships.

Formal relations between Turkish words build clusters in the vectors. It is proved that the examined Turkish word vectors are clustered and related according to Turkish-specific suffixes.

4.4. Arithmetic operations of word vectors and morphology between words

New vectors can be also obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Turkish corpus by examining the formal clustering.

The first five word vectors with the closest cosine similarity to the result vector of ('gitmek') - ('git') + ('götür') operation are shown below.

```
[('götürmek', 0.7065088748931885),
('yetişmek', 0.5844410061836243),
('götürülmek', 0.5795775651931763),
('binmek', 0.5781220197677612),
('uğurlamak', 0.561299204826355)]
```

According to the results of operations ('gitmek') - ('git') + ('götür'), the obtained vectors, clustered like a cosine, referring to the formal feature infinitive suffixes "- mek" and "- mak".

The first five word vectors with the closest cosine similarity to the result vector of ('çiçekli') - ('çiçek') + ('yaprak') operation are shown below.

```
[('yapraklı', 0.6582359671592712),
('dallı', 0.6488081812858582),
('dişbudak', 0.6367601752281189),
('otu', 0.624358594417572),
('yapraklar', 0.61882483959198)]
```

The word "Yapraklı" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Adjective, with leaves [30, 31].

The word "Çiçekli" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Adjective, with flowers or pictures of flowers [30, 31].

According to the results of operations ('çiçekli') - ('çiçek') + ('yaprak'), the vector ('yapraklı'), clustered like a cosine, referring to the formal feature noun rooted adjectives derived with similar suffixes "- li", "- lı".

The first five word vectors with the closest cosine similarity to the result vector of ('tazelik') - ('taze') + ('saydam') operation are shown below.

```
[('saydamlık', 0.4215427339076996),
('opak', 0.3784925937652588),
('erçivan', 0.3675283193588257),
('görüntüleme', 0.36332154273986816),
('tipindedir', 0.3586195111274719)]
```

The word "Tazelik" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; state of being fresh, young.
2. Noun; (metaphor) a state of cheerfulness, liveliness [30, 31].

The word "Saydamlık" in the current Turkish dictionaries, represented by Turkish Language Association (TLA) is defined as

1. Noun; state of being transparent, transparency [30, 31].

According to the results of operations ('tazelik') - ('taze') + ('saydam'), the vector 'saydamlık' clustered like a cosine, referring to the formal feature noun rooted nouns derived with similar suffixes"- lik", "- lık".

The vectors obtained from the Turkish corpus are clustered considering the formal relations between the words they belong to. It is proved that the formal results obtained by addition and subtraction on vectors are clustered and related according to Turkish-specific suffixes.

5. Discussions

Word vectors obtained from the general content Turkish corpus using Word2Vec model are clustered and related in terms of semantic relations and formal (structural) relations with Turkish words they belong to or both simultaneously according to Turkish-specific suffixes.

The word vector ('elma') is clustered together with the other word vectors, obtained from the words belonging to lexical paradigm representing the fruit names. The word vector ('istanbul') is clustered together with the other word vectors, obtained from the words "Ankara", "Bursa", "Trabzon", representing the city names. The word vectors ('üsküdar') and ('yenibosna') are clustered together with word vector ('istanbul') because words "Üsküdar" and "Yenibosna" represent names of two important districts of Istanbul. The word vectors ('ahmet') and ('ayşe') are in the same semantic cluster related to the proper noun-meaning relationship and differentiated according to gender characteristics. The word vector ('okul') is clustered together with the other word vectors obtained from the words "İlkokul", "Dershane", "Lise", "Ortaokul", representing the educational place. The word vectors ('okul') and ('okulun') are clustered together representing formal relation. It is formal derivation of the noun root "okul" with the suffix "- in".

The word vectors ('dönmek'), ('yetişmek'), ('götürmek'), ('inmek'), ('yerleşmek') are found in the cosine similarity of the word vector ('gitmek'). They are in formal relations, representing the word vectors referring to the verbs in the form of infinitive. It is formal derivation of the verb root with the infinitive suffix "- mek". The word vectors ('gitmiştim'), ('gittiğimde'), ('gidiyordum'), ('gidiyorum'), ('gideceğim') are found in the cosine similarity of the word vector ('gittim'). They are in formal relations representing the word vectors belonging to the inflected words derived by adding the first person singular suffix to the verb root "git". The word vectors ('kumluca'), ('akseki'), ('ibradı') are found in the cosine similarity of the word vector ('elmalı'). The clustering of word vectors is representing the semantic feature being a district of Antalya city. It takes place according to the semantic relations with the word "Elmalı". The word vectors ('ormanlık'), ('çalılık'), ('sazlık'), ('makilik'), ('otluk') are found in the cosine similarity of the word vector ('ağaçlık'), representing the names of the place. The clustering of word vectors is related to the formal feature inflected words derived by adding the suffixes "- lık", "- lik", "- luk" to the noun root and takes place within the semantic and formal relations with the words. It is formal derivation of the noun root with suffixes"- lık", "- lik", "- luk" and semantic feature, building a place name from the item. The word vectors ('muhasebecilik'), ('doktorluk'), ('hakimlik'), ('yargıçlık'), ('memurluk') are found in the cosine similarity of the word vector ('avukatlık'), representing building a job name from profession name. The clustering of word vectors is related to the formal feature inflected words derived by adding the suffixes "- lık", "- lik", "- luk" to the noun root and takes place within the semantic and formal relations with the words. It is formal derivation of the noun root with suffixes"- lık", "- lik", "- luk" and semantic feature, building a job name from profession name. The first two word vectors ('temizleme') and ('temizliği') in the cosine similarity of the word vector ('temizlik') are clustered like word vectors of inflected words derived by adding the suffixes "- leme", "- liği" to the adjective root. These are formal relations. Word vectors ('banyo') and ('kumlama') are clustered like vectors, representing the semantic relations with the word "Temizlik". The first five word vectors obtained from the word "Temizlik" clustered like cosine vectors according to the semantic and/or formal relationships.

New vectors are obtained as a result of adding and subtracting (arithmetic operations) the word vectors obtained from the Turkish corpus.

The vector obtained as a result of ('kral') - ('erkek') + ('kadın') operation is ('kraliçe'), the first vector among the cosine-like vectors. It is the replacement of the gender characteristic in the word "Kral", expresses nobility. The vector obtained as a result of ('ingiltere') - ('londra') + ('ankara') operation is ('türkiye'), the first vectors among the cosine-like vectors. It is the transaction of the relationship between countries and cities (or their capitals). The vectors obtained as a result of ('finans') - ('para') + ('altın') operation are word vectors ('bankacılık'), ('gayrimenkul'), ('sigortacılık'). They are in semantic relations compatible with the result of adding and subtracting vectors.The closest cosine-like vector obtained as a result of ('spor') - ('futbol') + ('yüzme') operation is ('olimpik'). It is in semantic relation with word vectors. The word vectors ('havuzu'), ('havuzları') and ('sporları') are in formal relations. It is formal derivation of the noun roots "havuz" and "spor" with the suffixes "- u" and "- ları". The word vector ('binicilik') is the result of two sport branches displacement in the vector process. According to the results of operations ('gitmek') - ('git') + ('götür'), obtained vectors clustered like a cosine, referring to the formal feature infinitive suffixes "- mek" and "- mak". According to the results of operations ('çiçekli') - ('çiçek') + ('yaprak'), the vector ('yapraklı'), clustered like a cosine, referring to the formal feature noun rooted adjectives derived with similar suffixes "- li", "- lı". The vector obtained as a result of ('tazelik') - ('taze') + ('saydam') operation is ('saydamlık').The clustering of word vectors is related to the formal feature inflected words derived by adding the "- lik", "- lık" to the noun root.

Our previous research was conducted on training word embeddings by using Word2Vec model representation from Ukrainian corpus. In this paper we took Turkish language to analyse word embeddings by using Word2Vec model representation from corpus belonging to other language family, Turkic.

6. Conclusions and Future Work

The research analyses regarding to the clustering of word vectors obtained from Turkish corpus of general subject content (using Word2Vec model) are made considering the two sub-branches of linguistics, semantics and morphology. The research analyses made in terms of semantics proved that the word vectors' accuracy could represent clusters according to semantic or formal relations with the words they belong to. The research analyses made in terms of morphology prove that word vectors are clustered and related in terms of morphological features according to Turkish-specific suffixes. It indicates a high structural level of construction of the Turkish language.

The cosine similarities of the vectors obtained by addition and subtraction on vectors are examined in terms of their compatibility with the meaning of the process. It is proved that the semantic results that can be obtained by addition and subtraction on vectors obtained from the English corpus can be also obtained from the Turkish corpus.

Considering the morphological properties of the words, the vectors can be clustered according to the suffixes they take or represent semantic relations between words.

The research analyses made in terms of semantics and morphology prove that vectors are clustered according to semantic or formal relations with the words they belong to. Verb and noun rooted words cause clusters in word vectors according to their semantic or morphological features or a mixture of both, their meanings in the sentence and the suffixes they took.

Our future works on this topic will focus on constructing semantic maps of various subject areas and expanding queries due to associative connections.

7. References

[1] J. F. Allen, Natural Language Processing, in: Encyclopedia of Computer Science, John Wiley and Sons Ltd, 2003, pp. 1218–1222.

[2] A. M. Turing, Computing Machinery and Intelligence, Mind (1950) 433–460. doi:10.1093/mind/lix.236.433

[3] T. Brants, A. Popat, P. Xu, F. Och, J. Dean, Large Language Models in Machine Translation,in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language, EMNLP-CoNLL, Association for Computational Linguistics, 2007, pp. 858–867.

[4] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of Machine Learning Research (2003) 1137–1155. doi:10.1162/153244303322533223

[5] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets, L. Chyrun, S. Chyrun, A. Vysotskyi, V. B. Fernandes, The Mobile Application Development Based on Online Music Library for Socializing in the World of Bard Songs and Scouts' Bonfires, in: Advances in Intelligent Systems and Computing IV, CSIT 2019, vol 1080. Springer, Cham, pp. 734–756. doi:10.1007/978-3-030-33695-0_49

[6] Y. Bengio, Y. Lecun, Scaling learning algorithms towards AI, in: L. Bottou, O. Chapelle, D. DeCoste, J. Weston (Eds), Large-scale kernel machines, Mit Press, Cambridge, Mass, 2007. doi:10.7551/mitpress/7496.003.0016

[7] Y. Chen, B. Perozzi, R. Al-Rfou, S. Skiena, The Expressive Power of Word Embeddings, in: Proceedings of the 30 th International Conference on Machine Learning, ICML 2013, Atlanta, Georgia, USA, 2013. arXiv:1301.3226

[8] R. Lebret, R. Collobert, Word Embeddings through Hellinger PCA, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 482–490. doi:10.3115/v1/E14-1051

[9] D. Jatnikaa, M. A. Bijaksanaa, A. A. Suryania, Word2Vec Model Analysis for Semantic Similarities in English Words, in : The Workshop Proceedings of the 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), pp. 160–167. doi:10.1016/j.procs.2019.08.153

[10] Y. Bengio, J.-S. Senecal, Quick Training of Probabilistic Neural Nets by Importance Sampling, in: Proceedings of AISTATS 2003. Society for Artificial Intelligence and Statistics, Florida, USA, 2003.

[11] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proceedings of Workshop at ICLR 2013, Computation and Language Scottsdale, Arizona, USA, 2013. arXiv:1301.3781v3

[12] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting Similarities among Languages for Machine Translation; Computing Research Repository (CoRR), 2013. arXiv:1309.4168v1

[13] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic Regularities in Continuous Space Word Representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia 2013, pp. 746–751

[14] Q. Le, T. Mikolov, Distributed Representations of Sentences and Documents, in: Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014, pp. 2931–2939.

[15] Y. Li, L. Xu, F. Tian, L. Jiang, X. Zhong, E. Chen, Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 2015, pp. 3650–3656

[16] C. Servan, A. Bérard, Z. Elloumi, H. Blanchon, L. Besacier. Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers Osaka, Japan, 2016, pp. 1159–1168 URL: https://aclanthology.org/C16-1110

[17] T. Luong, H. Pham, C. D. Manning, Bilingual word representations with monolingual quality in mind, in: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, Denver, Colorado, 2015, pp. 151–159. doi: 10.3115/v1/W15-1521

[18] L. Savytska, N. Vnukova, I. Bezugla, V. Pyvovarov, M. T. Sübay, Using Word2Vec Technique to Determine Semantic and Morphologic Similarity in Embedded Words of the Ukrainian Language, in: The Workshop Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), Lviv, Ukraine, volume 2870, pp. 235-248. URL: http://ceur-ws.org/Vol-2870/paper21.pdf

[19] A. Rysin, V Starko, D. Chaplynskyi, Slovnyk VESUM ta inshi poviazani zasoby NLP dlia ukrainskoi movy [VESUM dictionary and other related NLP tools for the Ukrainian language], 2007. URL: https://r2u.org.ua/articles/vesum

[20] A. Romanyuk, Vektorni predstavlennia sliv dlia ukrainskoi movy [Vector Representations of Ukrainian Words], Ukraina moderna [Modern Ukraine], No 27, 2019, pp. 46–72. doi: uam.2019.27.1062

[21] V. Vysotska, S. Holoshchuk, R. Holoshchuk, A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach, in: The Workshop Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021), Lviv, Ukraine, volume 2870, pp. 311-356. URL: http://ceur-ws.org/Vol-2870/paper26.pdf

[22] E. Altszyler, M. Sigman, S. Ribeiro, D. Fernández Slezak, Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database, Computer Science 2016. arXiv:1610.01520

[23] E. Altszyler, M. Sigman, D. Fernandez Slezak, Corpus Specificity in LSA and Word2vec: The Role of Out-of-Domain Documents, in: Proceedings of The Third Workshop on Representation Learning for NLP, Melbourne, Australia, Association for Computational Linguistics, 2018, pp. 1-10. doi: 10.18653/v1/W18-3001

[24] Gensim: topic modelling for humans. Word2vec embeddings. URL: https://radimrehurek.com/gensim/models/word2vec.html

[25] Gensim: topic modelling for humans. Corpus from a Wikipedia dump. URL: https://radimrehurek.com/gensim/corpora/wikicorpus.html

[26] A. Aksoy, Word2Vec gibi işlemlerde kullanılmaya uygun Türkçe metin dosyaları [Turkish text files suitable for use in processes such as Word2Vec]. URL: https://drive.google.com/drive/folders/0B_iRLUok9_qqOFozeHNFMjRHTVk

[27] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, D. Pinto (2014) Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model, in: Computación y Sistemas, Thematic issue: Computational linguistics, Vol 18, No 3, 2014, pp. 491–504. doi: 10.13053/CyS-18-3-2043

[28] A. Aksoy, Python ile Türkçe derlem (corpus) hazırlama [Preparing a Turkish corpus with Python], 2021. URL: https://github.com/ahmetax/derlemtr

[29] Wikimedia database dump of the Turkish Wikipedia, URL: https://archive.org/details/trwiki-20190101

[30] Türk Dil Kurumu, Sözlük [Turkish Language Association, Dictionary], 2019. URL: https://sozluk.gov.tr/

[31] Türkçe Kelime Sözlüğü [Turkish Turkish Dictionary], 2022. URL: https://kelimeler.gen.tr/

[32] Z. Korkmaz, Türkiye. Türkçesi grameri: şekil bilgisi, Türk Dil Kurumu [Turkey. Turkish grammar: morphology, Turkish Language Association], Ankara, 2019, 1027 p.