Studying the Evolution of Greek Words via Word Embeddings

Vasileios Barzokas
contact@vbarzokas.com
School of Informatics, Aristotle University of Thessaloniki
Thessaloniki, Greece

Eirini Papagiannopoulou
epapagia@csd.auth.gr
School of Informatics, Aristotle University of Thessaloniki
Thessaloniki, Greece

Grigorios Tsoumakas
greg@csd.auth.gr
School of Informatics, Aristotle University of Thessaloniki
Thessaloniki, Greece

ABSTRACT

The meanings of words change over time, reflecting changes in language and society (political, economic or cultural). This study focuses on the more recent form of modern Greek, i.e. Demotic, also known as Dimotiki, aiming to trace semantic shifts of Greek words between consecutive time periods. We develop a systematic framework that gathers free online Greek digitized literature books and analyzes them using natural language processing tools to learn representations of human language. Then, we conduct an experimental analysis to evaluate the quality of our trained models as well as their ability to detect semantic shifts. We present representative results of actual semantic shifts in Greek words in the period 1980 to 2020.

CCS CONCEPTS

• **Computing methodologies** → **Lexical semantics**; • **Applied computing** → **Document management and text processing**.

KEYWORDS

diachronic word embeddings, semantic shifts, k nearest neighbours, Greek language

ACM Reference Format:
Vasileios Barzokas, Eirini Papagiannopoulou, and Grigorios Tsoumakas. 2020. Studying the Evolution of Greek Words via Word Embeddings. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411408.3411425

1 INTRODUCTION

Greek language appeared roughly 3,000 years ago, passing through various changes while being adopted by many population communities [2]. From Ancient Greek to the more recent forms of modern Greek, i.e. Katharevousa and Dimotiki, Greek language expresses philosophical principles and high ideals. Studying the diachronic semantic shifts of words is of interest to several research communities, such as computational linguisticsand political science. Such studies can help researchers discover more about human language and extract time-specific knowledge from documents [12].

SETN 2020, September 2–4, 2020, Athens, Greece

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411425

Word embeddings are a promising machine learning technology for the study of the evolution of the Greek language. In such models, each word in a given language corresponds to a vector representation. These word vectors capture semantic associations between the words. For example, words with vectors that are close to each other, are also close in meaning. Word embeddings are typically trained on large corpora and outperform statistical models based on co-occurrence analysis in various linguistic and natural language processing tasks. Moreover, based on *distributional hypothesis*, words occurring in identical contexts usually imply similar meanings in linguistics. Hence, a change in a word's neighbourhood could indicate a semantic shift [19].

This work develops a systematic framework that uses word embeddings to trace semantic shifts or variations in the context of words over time, aiming to reveal not only linguistic, but also cultural evolution caused by technological and socio-economic developments. We focus on the latest form of the Greek language, after 1970, gathering a time-stamped collection of digitized copyright-free and publicly available Greek literature books from the web. To the best of our knowledge, there is no previous work studying the evolution of the Greek language using distributional methods.

The main contributions of this work are the following: (a) the first study to trace semantic shifts in the context of words over time in the Greek language using diachronic word embeddings, (b) a time-stamped Greek corpus and time-specific embeddings to the research community, (c) appropriate software that collects Greek text documents from the web in PDF format and converts them to plain text, ready for text processing (can be utilised to enrich the corpus with further documents). Contributions (b) and (c) are publicly available on GitHub[1].

The rest of this paper is organized as follows. Section 2 presents related work on the detection of words' semantic shifts using distributional methods. Section 3 describes in detail the methodology we follow, whereas Section 4 demonstrates a qualitative experimental analysis to evaluate the utility of the framework for the analysis of the evolution of the modern Greek language. Finally, Section 5 presents the conclusions and future directions of this work.

2 RELATED WORK

The inspiration for this work comes from a relevant research project by Stanford University, called HistWords [10], which analyzed the semantic change of more than 30,000 words across 7 datasets of 4 languages (English, French, German, and Chinese) over 150 years. The study quantifies semantic change using various embeddings and reveals statistical laws of semantic evolution.

[1] https://github.com/intelligence-csd-auth-gr/greek-words-evolution

2.1 Diachronic Word Embeddings

Diachronic word embeddings are constructed after training models on corpora of different time-periods, which are then aligned over time. Various methods have been used to create word embeddings in the context of this task, such as: (a) Positive Point-wise Mutual Information (PPMI) representations [5]; (b) Singular Value Decomposition (SVD) embeddings [18], [6]; (c) neural embeddings, e.g., Word2Vec [13], fastText [4] etc.; (d) GloVe [16]; and (e) topic modelling [3].

It is possible to train separate word embedding models using corpora from different time-periods. However, it is not sensible to directly calculate cosine similarities between embeddings of the same word from different models. The reason is that modern word embedding models are stochastic, resulting in rotation invariance: the cosine similarity between the vectors of a word from different time-periods can be quite low, even if the word's meaning remains the same, due to the different random initializations of the two models.

A common solution to this problem is the models' alignment using linear transformations to fit them in the same vector space. Then, cosine similarities across models can be used to reveal the semantic shifts of words. Orthogonal Procrustes transformations [11], [17], [8] is a well-known approach for diachronic models' alignment. Given 2 matrices A and B, they find an orthogonal matrix R which most closely maps A to B [17]:

$$R = \arg\min_{\Omega} \|\Omega A - B\|_F \quad \text{subject to} \quad \Omega^T \Omega = I \tag{1}$$

This problem is equivalent to the following one. Given a matrix $M = BA^T$, find its nearest orthogonal matrix R using the SVD:

$$M = U\Sigma V^T \tag{2}$$

$$R = UV^T \tag{3}$$

Another direction is to use second-order embeddings [7], i.e., the vectors of words' similarities to all other words in the shared models' vocabulary. Such approaches do not require any transformations. Moreover, there exist techniques that learn the word embeddings across several time-periods jointly and align them at the same time to put the models in the same vector space [1].

2.2 Quantifying Linguistic/Semantic Change

After the models' alignment, we can compare word vectors across these models. Hamilton et al. [9] distinguish semantic changes caused by cultural shifts (e.g., technological advancements) from those caused by more regular linguistic processes of semantic change (e.g., grammaticalization or subjectification) and propose corresponding local and global measures for quantifying them. Given word embeddings trained on corpora of consecutive decades, they consider the: (a) cosine similarity between the word's second-order vectors (local neighbourhood measure for cultural shifts) and (b) cosine distance between the word's vectors (global measure for linguistic shifts). The appropriateness of the local (global) measure for cultural (linguistic) shifts was confirmed in [10, 12].

3 METHODOLOGY

As there is a lack of publicly available large time-stamped corpora in the Greek language, the challenge was to develop software that collects Greek books from the web along with their publication date and converts them to a plain text format appropriate for model training. Then, following the paradigm of [10], we used the simple but effective comparison approach of k Nearest Neighbours along with cosine distances to determine the words' semantic change.

3.1 Corpus and Metadata

Our corpus comes from two sources: i) Project Gutenberg[2], a library with over 60,000 free e-books from all over the world, offered in various digital formats, including plain text, *EPUB*, *kindle* and *HTML*, and ii) Open Library[3], an online library with more than 7,000 Greek e-books, most of which are in PDF format.

As we focus on the latest form of the Greek language, we had to filter books that were published before 1970. Each book entry of Open Library contains a specific text field that is referencing the publication date of the book. Hence, extracting the publication date for this source was quite easy. However, Project Gutenberg does not contain the publication year of the book among its metadata. We, therefore, had to parse the downloaded text files, examine the structure of many books and develop appropriate techniques for extracting the book's published date. Since many of its books were from the late decades of the 18th century or the dawn of 19th, they had a similar pattern for the covers' structure that usually contained the publisher's location and year of publication among the other data. Table 1 presents some of these patterns.

Table 1: Sample covers containing publisher information and publication date.

ΕΝ ΑΘΗΝΑΙΣ ΕΚΔΟΤΙΚΟΣ ΟΙΚΟΣ ΓΕΩΡΓΙΟΥ ΦΕΞΗ 1912
ΑΔΕΛΦΟΙ ΔΕΠΑΣΤΑ ΒΙΒΛΙΟΠΩΛΑΙ ΚΑΙ ΕΚΔΟΤΑ1 ΕΝ ΚΩΝΣΤΑΝΤΙΝΟΥΠΟΛΕΙ ΕΝ ΑΘΗΝΑΙΣ, 1876.
ΑΔΕΛΦΟΙ ΔΕΠΑΣΤΑ ΕΝ ΑΘΗΝΑΙΣ, ΤΥΠΟΙΣ ΔΙΟΝΥΣΙΟΥ ΚΟΡΟΜΗΛΑ. 1864
ΕΚΔΟΣΙΣ ΤΙΜΗΤΙΚΗ ΕΠΙ ΤΗ ΠΕΝΤΗΚΟΝΤΑΕΤΗΡΙΔΙ ΤΟΥ ΣΥΓΓΡΑΦΕΩΣ ΑΘΗΝΑΙ — ΕΚΔΟΤΗΣ ΙΩΑΝΝΗΣ Ν. ΣΙΔΕΡΗΣ — 1921

Our approach regarding the extraction of the publication year from such covers was to get the first 100 lines of text, where usually the book's publication year appears, and use the following regular expression to detect a specific format of text (sequence of words/characters):

'(?:ΑΘΗΝΑ␣ΑΘΗΝΑΙ␣ΑΘΗΝΗΣΙ␣ΠΕΙΡΑΙΕΥΣ␣ΖΑΚΥΝΘΟ␣ΣΜΥΡΝΗ␣ΠΑΡΙΣΙΟΙΣ)(?:[\s\D]\d{0,3})*(\d+(?:\s+\d+){3})'

[2] https://www.gutenberg.org/
[3] https://www.openbook.gr/

In cases where the search did not give any results, we had to estimate the publication year by averaging the authors' birth and death dates, which were available in the metadata of Project Gutenberg. Following this process, we managed to gather 219 digitized Greek literature books from *Project Gutenberg* and 1648 from *Open Library* (i.e., 1867 books in total).

3.2 Parsing and Preprocessing

The parsing of Open Library's PDF files was conducted with a Python port of the Apache Tika library[4]. Manual inspection of the extracted text revealed that a significant amount was gibberish and could not be read in 359/1648 files from *Open Library*. This is a common issue during the conversion of a PDF file to plain text, since PDF is a proprietary format without mark-up tags to determine the document's appearance that uses various types of fonts, each one with its distinct glyph for the same character. Hence, in cases where there is no information within a PDF file's metadata regarding the fonts, i.e., the glyphs are not accurately mapped to the ASCII or UTF-8 tables, PDF tools extract text that does not correspond to the actual file content. We experimented with a few other popular tools regarding extraction of text from PDF[5,6,7], but due to those limitations none of them managed to properly extract the text from those files. Furthermore, PDF files coming from scanned books (i.e., images) are also discarded, as their conversion to plain text fails. This occurred in 218/1648 files from *Open Library*. Also many books, mainly from Project Gutenberg were written in older forms of the Greek language (Ancient or Katharevousa), which would not help in the comparison with books written in modern form (Dimotiki). To automatically detect the problems mentioned above, we follow the procedure described in Algorithm 1 that utilizes a spell checking library, called *pyEnchant*, for the modern Greek language. Finally, we end up with 1071 files from *Open Library* and 71 from *Project Gutenberg* for a total of 1142 books. Table 2 presents the number of books per source and the results after conversion, while Table 3 presents the number of books and tokens per period and Table 4 presents the number of books and tokens per genre for the top 10 genres, which are those extracted from the online sources during our web scrapping process.

Algorithm 1 Clear malformed text

```
Remove all non-Greek characters from text
Remove multiple whitespaces and multiple new lines
Tokenize the text
if Number of tokens is less than 100 then
    Discard the file, tag it as "no contents" and don't use it
end if
Pickup 100 random tokens
for All the random tokens do
    if Spell checker identifies <80 words as correct then
        Discard the file, tag it as "malformed" and don't use it
    end if
end for
```

[4] https://pypi.org/project/tika/
[5] https://github.com/kermitt2/grobid
[6] https://github.com/mstamy2/PyPDF2
[7] http://www.xpdfreader.com/

Table 2: Filtered data per source.

Source	Total	Malformed	No Content	Parsable
gutenberg.org	219	148	0	71
openbook.gr	1.648	359	218	1.071
Total	1.867	507	218	1.142

Table 3: Tokens and books per period.

Period	Books	Tokens
<1800	49	1.502.698
1800-1900	107	2.766.183
1900-1910	38	1.459.252
1910-1920	81	2.651.672
1920-1930	35	1.344.457
1930-1940	2	207
1940-1950	0	0
1950-1960	3	26.793
1960-1970	3	93.192
1970-1980	19	93.181
1980-1990	17	118.966
1990-2000	49	1.180.326
2000-2010	170	3.820.727
2010-2020	1218	19.831.060
Total	1791	34.888.714

Table 4: Tokens and books per genre for the top 10 genres.

Genre	Books	Tokens
Unknown	172	4.205.178
Poetry	145	592.180
Story	138	1.850.631
Fiction	116	6.212.488
Fairy tale	75	171.125
Essay	67	1.642.666
Theatrical	60	614.123
Novel	38	503.727
Study	34	896.737
Children's	19	65.930

We also performed the following preprocessing steps for both online sources to avoid misleading the algorithm by providing erroneous or non-parsable data:

(1) lowercasing all words: we had to normalize the words by converting them to lowercase.
(2) removing stopwords: we decided to remove stopwords, as they added useless information in our word vectors.
(3) removing all characters that were not letters or numbers: we eliminate any special characters/symbols, as they did not contain any useful information.

3.3 Studying the Semantic Evolution of Words

3.3.1 Diachronic Word Embeddings. We use the metadata earlier extracted and divide the available years of publication into n equal time-intervals (the number of intervals is a user-defined option, e.g., 5/10/20-year intervals). Then, we assign each document to the corresponding time-period based on its publication year. In this way, we have a different corpus for each time-interval, ready for the training of embeddings.

For our purpose of learning efficient word representations we decided to use *fastText*, which is an extension to *word2vec* and some relevant research studies regarding the performance of different word embedding methods over the Greek language [14, 15] showed that it produces slightly better results compared to *word2vec*.

After the model's training on each corpus, n distinct sets of word embeddings are available.

3.3.2 Qualitative Evaluation of the Diachronic Word Embeddings. To check the quality of our trained embeddings, we utilise the pre-trained fastText word vectors for the Greek language[8] (CBOW with position-weights, 300 dimensions, with character 5-grams, a window of size 5 and 10 negatives). For a given set of words, we find their corresponding nearest neighbours using: (a) our models and (b) the pre-trained embeddings. Then, we manually examine/compare the quality of the nearest neighbours in both cases. Despite the limited size of the Greek corpus compared to Common Crawl and Wikipedia used for the pre-trained fastText embeddings, we didn't detect any notable difference in the quality of our models in comparison with the pre-trained one. Moreover, sometimes, the pre-trained model returns slightly worse results with stopwords or metacharacters as neighbours.

As an example we will consider the results returned when querying for the nearest neighbours using those pre-trained word vectors for the word *tomato (ντομάτα)* compared to our vectors produced from the combined corpus. In the first case (Table 5), we get back results that are variations of the queried word, most of which make no sense, while in the latter case (Table 6) we get back various different words related to food products, vegetables and cooking methods, which are closer to the queried word's context.

Table 5: Example low quality result from the pre-trained word vectors.

Word	Distance
τομάτα	0.789549
ντομάτα.	0.778906
ντομάτα1	0.745006
ντομάτα2	0.727504
1ντομάτα	0.725901
ντομάταΠίτσα	0.704069
ντομάταλίγα	0.687392
ντομάτας2	0.680488
αγγουροντομάτα	0.670166
ντομάταςΠίτσα	0.669205

[8] https://fasttext.cc/docs/en/crawl-vectors.html

Table 6: Example of a better quality result from our word vectors.

Word	Distance
σνίτσελ	0.846922
ρίγανη	0.840451
μπακαλιάρο	0.835745
πιπεριά	0.832499
πασπαλίζουμε	0.830127
αλεσμένο	0.828932
μπρόκολα	0.828713
τυροκαυτερή	0.827923
τηγανητά	0.826654
κρίθινο	0.824639

To further evaluate the quality of our fastText models, we merged the corpora of the two datasets (i.e., Project Gutenberg and OpenBook) into one and trained an additional fastText model on this larger corpus. Then, we manually evaluated the resulting model by getting the nearest neighbours for a set of common words that we arbitrarily defined based on our experience and everyday usage of the Greek language, like some common fruits (e.g. *potato (πατάτα), tomato (ντομάτα), apple (μήλο)* etc), animals (e.g. *dog (σκύλος), cat (γάτα), mouse (ποντίκι), lion (λιοντάρι)* etc), items (e.g. *car (αυτοκίνητο), door (πόρτα), table (τραπέζι)* and concepts (e.g. *work (δουλειά), god (θεός), friendship (φιλία)* etc). In Figure 1, we give representative visualizations of the evaluation process (along with English translations of the words to facilitate the understanding).

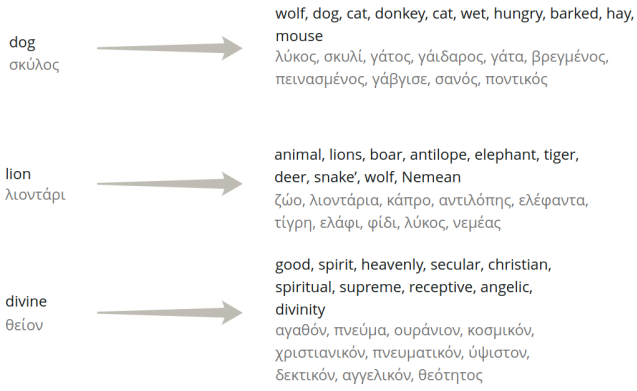

Figure 1: The 10 nearest neighbours of the words: *dog (σκύλος), lion (λιοντάρι), divine (θείον).*

3.3.3 Discovering Shifts from Data. We use the cosine distance metric to detect the words with high semantic shift within two distinct time-periods. First, for each word, we compute the cosine distance between its corresponding embeddings existing in both corpora, i.e., the diachronic embeddings of the word. Then, we keep the top N words with the highest semantic change in time, i.e., the highest cosine distance values. Before any computation/comparison between word vectors, we employ the "orthogonal Procrustes transformations" approach to ensure the alignment in one vector space

(see Equation 1). For each word from those with the highest cosine distance values, we find the 20 nearest neighbours in the corpus of each time-period and keep them separately to perform manual evaluation detecting significant semantic shifts. Based on the distributional hypothesis mentioned previously, words occurring in identical contexts usually imply similar meanings. Hence, a change in a word's neighbourhood could indicate a semantic shift [19].

4 EMPIRICAL STUDY

We repeated the procedure described in Section 3.3 several times to determine the right number of time-intervals that give meaningful results, due to the small size of the Greek corpus and the methodology's requirements to split it into consecutive time-periods creating even smaller training corpora. This section presents the experimental setup of the study (Subsection 4.1) and indicative results with remarkable semantic evolution in time (Subsection 4.2).

4.1 Setup

The option given by fastText of character n-grams did not help our study, i.e., we set the corresponding parameters *minn* and *maxn* equal to 0. We also used the maximum CPU resources of our system to speed up the training time of fastText. The parameter *minCount* (i.e., the minimum occurrences of a word) is equal to 5 to avoid the return of embeddings for infrequent terms that do not provide useful insights regarding semantic shifts. We have also experimented with various vector dimension sizes (values for the parameter *dim* from 100 to 300) getting the same quality of embeddings. Hence, we keep the default size of dimensions, i.e., 100. Concerning the number of epochs (*epoch* parameter) for the training, we tried several values ranging from 5 (default) to 50. Generally, more epochs create better embeddings in terms of quality but need more execution time. We concluded that values higher than 25 were not worth the effort. For the rest of the parameters, we used the default values. Table 7 shows some indicative results for the impact of each fastText parameter to execution time.

Table 7: Impact of each fastText parameter on the training time of our models (*ws* stands for the size of the context window; *neg* is the number of negative sampled instances; *defaults* refers to the default parameters' values).

Parameter	Value	Execution time
defaults	defaults	2m 30s
dim	200	4m 20s
dim	300	5m 10s
dim	400	8m 30s
epoch	10	5m 10s
epoch	20	10m 20s
epoch	50	28m 20s
minCount	15	2m 30s
minCount	50	2m 30s
neg	10	3m 30s
neg	20	3m 30s
ws	10	4m 00s
ws	20	6m 50s

4.2 Representative Results Per Period

We focus our analysis on terms related to the fields of politics and technology/commercial products, as socio-political progress (changes) and technological advances usually cause semantic shifts in words [12]. We present and discuss the most representative examples identified during our study. However, similar results, in the same spirit, are obtained for other words too.

First, we trace the semantic evolution of the term "κρίση" (crisis). The word means an event that often leads to an unstable situation affecting individuals, groups, etc. We choose to study the semantic shifts in 5-year time intervals. Hence, based on the methodology outlined in Section 3.3, we capture the general contexts of the word in 5 consecutive periods from 1995 to 2020, highlighting the main concerns of Greek society. Figure 2 visualizes the diverse concepts of crisis based on the resulting nearest words. In the analysis that follows, in this paragraph, the words that appear as nearest neighbours of the corresponding periods are in italics. Some major issues of concern to Greek society during the period 1995-2000 are the military crisis of Imia between Greece and *Turkey* in 1996 as well as the impact of the Soviet regime overthrow and the end of the Cold War on the social-democratic/*Left* political parties. However, there is a shift in the meaning closer to *financial* concepts, indicating the *violent* transition to the deep *economic* recession and the small-scale humanitarian crisis (*empathy* issues) of Greece since 2009 as well as the intense feelings (*depression*) arisen in people due to *imperative* mass *dismissals*, living problems, etc. (2005-2015). Phrases such as *austerity*, *neoliberal* politics, *fiscal* adjustment, the economic blow to the *middle class* were in the daily discussions' vocabulary. Finally, the nearest neighbours of the term crisis based on the embeddings coming from the corpus comprise the most recent documents (2015-2020) reflect the economic instability and uncertainty due to referendum for the approval or rejection of the *Troika's* bailout terms. The country's possible *bankruptcy*, the poverty *plaguing* the masses, the negative *stock market indices* as well as the Greece's imminent economic *recovery* (exit from the bailout on August 20, 2018) cause intense *skepticism*.

Then, we present the semantic evolution of the words' contexts "υπολογιστής" (computer) and "ποντίκι" (mouse), where the rapid technological progress causes the shifts (Figures 3 and 4, respectively). A computer is a device instructed to perform sequences of operations automatically via executable programs, whereas a mouse can be either a small rodent or the computer's pointing device. In this case, we study the semantic shifts in 20-year time intervals, capturing the general contexts of the words in two consecutive periods from 1980 to 2020. Before the wide use of personal computers (PC) by the public (1980-2000), the majority of books and articles use terms that highlight the main features of a PC, such as the *electronic mail*, the concept of a *user*, the computer *graphics* and *software*, etc. However, after 20 years from the IBM's first PC launch (1981) and 10 years from the public availability of the World Wide Web (1991), modern contexts replace the previous ones reflecting the fast evolution of telecommunications (*broadcast*), the contemporary needs for portability (*portable*, *mobile*, *tablet*) as well as the widespread use of computers and peripherals in work and home (*keyboard*, *printer*). In this vein, the corresponding nearest neighbours in the first period (1980-2000) for the term mouse are

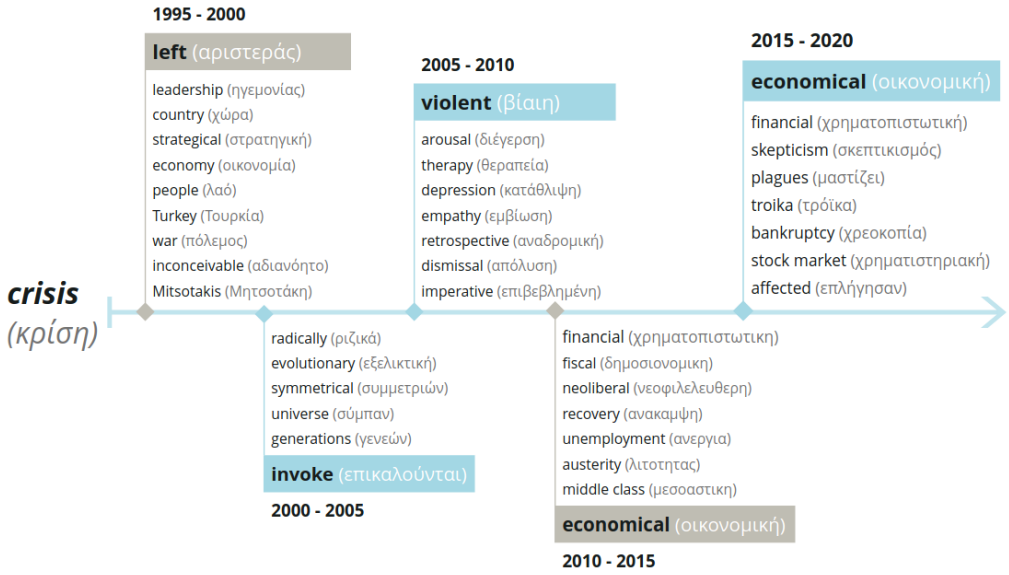

Figure 2: Nearest neighbours per 5 years to word *crisis* (κρίση) from 1995 to 2020

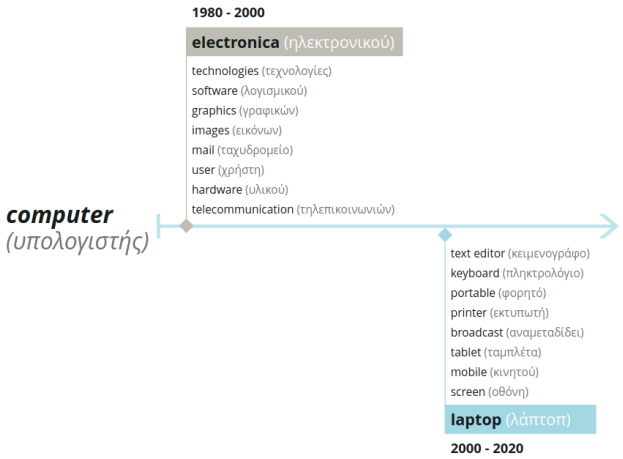

Figure 3: Nearest neighbours per 20 years to word *computer* (υπολογιστής) from 1980 to 2020.

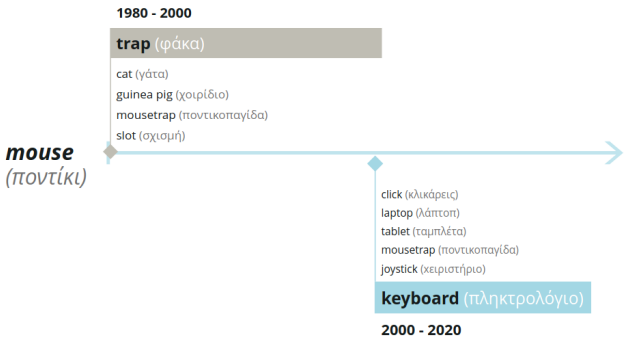

Figure 4: Nearest neighbours per 20 years to word *mouse* (ποντίκι) from 1980 to 2020.

relevant to the actual rodent animal (*mousetrap*, *guinea pig*, *cat*). However, in the second period (2000-2020), the corresponding contexts reveal the semantic transition from rodent to electronic device (*laptop*, *tablet*, *click*, *joystick*).

5 CONCLUSIONS, CHALLENGES AND FUTURE WORK

In this work, we develop a systematic framework to study the evolution of the modern Greek language using distributional methods. The Greek corpus, the trained models and our software for the collection of Greek PDF documents from the Web and the conversion to plain texts are available to the research community. Despite the small size of the corpus, we show that using state-of-the-art NLP libraries (e.g. fastText) and comparison techniques (e.g. cosine similarity), we produce meaningful word representations and discover semantic shifts revealing socio-economic or technological changes. The biggest challenges we successfully faced were (a) the lack of freely available digitized Greek books compared to other languages (e.g., English, German or Chinese) and (b) the multiple forms of the Greek language, e.g., Ancient, Katharevousa, Demotic, etc., impeding the comparison of words between different periods.

This study can be the groundwork and a reference for future and more extensive research in various fields, e.g., Linguistics, Sociology, Philosophy and NLP. Furthermore, a corpus that spans evenly across the last century or the enhancement of the documents' collection with texts from newspapers, magazines and online sources (web forums and social media) would be an interesting direction.

ACKNOWLEDGMENTS

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code:T1EDK-05580)

REFERENCES

[1] Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Doina Precup and Yee Whye Teh (Eds.). Proceedings of Machine Learning Research (PMLR), Sydney, Australia, 380–389. http://proceedings.mlr.press/v70/bamler17a.html

[2] Alfred Bammesberger and Theo Vennemann. 2003. *Languages in prehistoric Europe*. Universitätsverlag Winter (UWH), Heidelberg, Germany.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics* 5 (2017), 135–146. https://transacl.org/ojs/index.php/tacl/article/view/999

[5] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguistics* 16, 1 (1990), 22–29.

[6] Susan T. Dumais. 2004. Latent semantic analysis. *ARIST* 38, 1 (2004), 188–230. https://doi.org/10.1002/aris.1440380105

[7] Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*. Association for Computational Linguistics, Berlin, Germany, 52–58. https://doi.org/10.18653/v1/p16-2009

[8] John C. Gower and Garmt B. Dijksterhuis. 2004. *Procrustes problems*. Oxford University Press, Oxford, UK.

[9] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics, Austin, Texas, USA, 2116–2121. https://doi.org/10.18653/v1/d16-1229

[10] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1489–1501. https://doi.org/10.18653/v1/P16-1141

[11] John R Hurley and Raymond B Cattell. 1962. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral science* 7, 2 (1962), 258–262.

[12] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1384–1397. https://www.aclweb.org/anthology/C18-1117

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Workshop Track Proceedings of the International Conference on Learning Representations (ICLR 2013)*. Scottsdale, Arizona, USA, 1–12. http://arxiv.org/abs/1301.3781

[14] Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. 2020. Evaluation of Greek Word Embeddings. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2543–2551. https://www.aclweb.org/anthology/2020.lrec-1.310

[15] Stamatis Outsios, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis. 2018. Word Embeddings from Large-Scale Greek Web Content. http://arxiv.org/abs/1810.06694

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[17] Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (3 1966), 1–10. https://doi.org/10.1007/BF02289451

[18] G. W. Stewart. 1993. On the Early History of the Singular Value Decomposition. *SIAM Rev.* 35, 4 (1993), 551–566. https://doi.org/10.1137/1035134

[19] Menasha Thilakaratne, Katrina Falkner, and Thushari Atapattu. 2020. Connecting the Dots: Hypotheses Generation by Leveraging Semantic Shifts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12085 LNAI. Springer, Singapore, 328–340. https://doi.org/10.1007/978-3-030-47436-2{_}25