EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

It's a long way to a dictionary: Towards a corpus-based dictionary of neologisms

Afentoulidou V., Christofidou A.

Academy of Athens, Greece, University of Athens, Greece
vafentoul@phil.uoa.gr, christo@academyofathens.gr

Abstract

In this paper we discuss three main different views on the documentation of neologisms supporting the construction of a corpus-based lexicon of neologisms (as a language resource), which will include those new lexical units that enter the *consolidation stage* (according to certain criteria) before their entering into the *establishment stage* (Kerremans 2015). The documentation (collection and monitoring) of those new lexical units will be both linguistically and lexicographically helpful: a. it provides the linguist with a valuable linguistic information tank (morphology, semantics, morphology-text interface etc.) and b. it facilitates the answer to the desideratum of the dictionary inclusion (or not) of neologisms. We focus on the second issue and show that corpus exploration methods and measurements such as peakedness of distributions and lexical dispersion can be operationalized as tangible criteria to conjointly evaluate the frequency profiles of new formations, and that peakedness is a promising indicator of "lexical sustainability". Drawing examples from a 160-million-word sub-corpus of the *Monitor Corpus of Neologisms* compiled for NEOΔHMIA research project at the Academy of Athens, comprising newspaper discourse spanning 5.4 years, we track the frequency development of selected new formations which emerged during the Greek debt crisis and discuss their evolution in time.

Keywords: neology; dictionary inclusion; corpora; consolidation; peakedness; dispersion

1 Introduction

Living in a period overwhelmed by the pandemic vocabulary, the first question which comes to mind would be: how many and which of these new formations will remain? The old question for linguists and lexicographers arises again: is it possible to establish specific indicators of the evolution, the survival, or the life-cycle of new words? In an attempt to give some answers to this question we chose to look back exploring the behaviour and the evolution of the already fading away neological vocabulary of the Greek debt crisis as witnessed by linguistic evidence in the corpus component of NEOΔHMIA (see section 2) and specifically by the data of one Greek newspaper within the timespan 2015-2020.

In the following (section 2), we highlight the main objectives and methodological commitments adopted for the purposes of the Greek Neology project NEOΔHMIA conducted at the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* concerning Modern Greek Neology. In section 3 we present related work on the topic. In section 4 we discuss the different views on the dictionary inclusion and/or documentation of neologisms. In section 5 a corpus-based analysis is conducted (methodology and results) and a discussion of the corpus findings follows in section 6. The paper ends with concluding remarks and a research outlook.

2 NEOΔHMIA at the Research Centre for Scientific Terms and Neologisms

NEOΔHMIA is an ongoing research programme conducted at the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* (2008-), constantly developed to accomplish the tasks of (semi)automated detection and linguistic analysis of Greek neologisms and terminology, the first of its kind concerning Greek neology (Christofidou et al. 2013). The project concerns the development of an integrated databank, comprising four main modules, constantly updated:

1) Neologism text retrieval component: A custom-made crawler is browsing the online versions of selected Greek newspapers (with the largest circulation). Although accurate text extraction and content representation are adapted to the technical and representational demands of newspaper feeds/webpages, the system is flexible enough to trace other kinds of online sources (i.e., non-press). The texts produced from the crawling module are cleaned and pre-processed and the final output is represented in XML and automatically enriched with metadata following the recommendations of the Text Encoding Initiative (TEI P5 Guidelines@tei-c.org). A document model has been defined as a custom-made TEI schema (Afentoulidou & Christofidou 2017) comprising different annotation layers for the representation of basic metadata and text profiling (genre and topic classification), the structural superordinate divisions of newspaper text (document structure), as well as basic grammatical analysis (POS tagging and lemmatization) and is expandable.

2) Neologism extraction component: A dedicated tool uses the output of the text retrieval component (or accepts any kind of XML file conforming to the TEI-schema of the project) and performs automated detection of "new words" (candidate neologisms) through computational techniques, which make use of the well-documented method of exclusion word lists, as well as named entities stoplists (see Christofidou et al. 2013; Afentoulidou & Christofidou 2017 for details on the processing steps, elimination of noise and methods of updating the exclusion procedure). The system identifies one-word

units, although candidate multiword formations are proposed and submitted to human inspection only for n-word grams connected with dashes (Christofidou et al. 2020). Lists of candidate neologisms are produced for manual evaluation. The Neologism extraction component is complemented by a manual selection procedure employing lexicographic criteria, such as (non)occurrence of the candidate neologisms in the reference dictionaries of Modern Greek, as well as the generation of web impact reports (Thelwall 2018) via software[1] or linguistically relevant search engine queries and measurements, Google or Bing-based (*Web as Corpus* methodology, Lüdeling, Evert & Baroni 2007; Christofidou et al. 2013).
3) Neologism classification component: Together with lexical use, morphological (mainly word-formation analysis) and textual information (genre, text type, topic, text structure etc.) is recorded in the Centre's database of neologisms. A system of automatic topic classification of newspaper articles is being developed with the aid of supervised and non-supervised machine learning techniques (see, among others, Hagen 2012) to facilitate contextual analyses of neologisms on a larger scale.
4) Neologism monitoring component: Any further quantitative or qualitative observations regarding the use of the words utterly classified as neologisms make use of the specifically designed corpus of online newspaper discourse mentioned above; the corpus is thus used both for neologism retrieval and monitoring (*Web for Corpus* methodology, see De Schryver 2002). The *Monitor Corpus of Neologisms* (Afentoulidou & Christofidou 2017) nowadays includes more than 400 million words of journalistic discourse and is compiled following international standards (Text Encoding Initiative) to support empirically testable, textually-informed analyses of the morphological tendencies of Modern Greek. Moreover, webometric data (Christofidou et al. 2013; Christofidou, Karasimos & Afentoulidou 2014) are supplied for every neologism for the date of its first recording in the database and on later intervals (on-demand so far, although an annual webometric monitoring is envisaged for all neologisms currently in the database).
The four components offer a dynamic (Renouf 2016; Cartier 2019), unified pipeline of research (although not fully automated yet) and define the Centre's digital infrastructure NEOΔΗΜΙΑ for tracking and classifying neological formations in Modern Greek.

3 Related Work

As far as research on Greek neologisms/neology is concerned, NEOΔΗΜΙΑ is active and constantly updated.[2] We should also mention the prominent research of Professor A. Anastassiadis-Symeonidis, which follows a lexicographic, more qualitatively-oriented database approach (Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009).[3]
The study of lexical innovation with the aid of computer technologies is central to numerous European initiatives and related projects concerning neology. Dedicated research environments based on corpus evidence have been developed for various languages. As a common denominator, they share a dynamic, quantitative orientation and a holistic concern for the balanced development of tools and procedures, not only for the challenging and fundamental task of semi-automatic discovery and linguistic classification of new words and/or new meanings (formal and semantic neology), but also for their monitoring across time, space and contexts. For instance, the French platforms *Néoveille* (Cartier 2016) and *Logoscope* (Gérard, Falk & Bernhard 2014), as well as *Néonaute* (a recent extension of *Néoveille* and *Logoscope* in collaboration with the BnF, see Aubry, Cartier & Stirling 2018); the web interfaces of the worldwide neology networks for Catalan and Spanish, coordinated by the Observatori de Neologia – *OBNEO* at the University Pompeu Fabra in Barcelona[4] (*Antenas Neológicas, NEOROM, NEOROC, NEOXOC*); the German web service *Die Wortwarte*[5] under the umbrella of the Berlin-Brandenburgische Akademie der Wissenschaften and the *Neologismenwörterbuch online*[6] at the Institute for the German Language in Mannheim; the *Neocrawler* (Kerremans et al. 2018) and the English Neologisms Research Group at Ludwig-Maximilians-Universität München; the system for neology extraction and monitoring based on the *Norwegian Newspaper Corpus* (Andersen & Hofland 2012); the special dictionary of neologisms *Neologismenwoordenboek*[7] and the neology section[8] of the online, corpus-based *Algemeen Nederlands Woordenboek* of the Instituut voor de Nederlandse Taal; the pioneering service *APRiL* (Renouf 2007a; Renouf 2013) at the RDUES Birmingham City University; all of them forming, to the best of our knowledge, a representative albeit not exhaustive list of neologism trackers, computational tools and infrastructures with solid online presence and/or dissemination of research outcomes in dedicated bulletins and printed series (see Christofidou et al. 2013; Afentoulidou & Christofidou 2017 for supplementary overviews). Irrespective of possible specific design requirements, this line of research necessarily adopts a database (SQL, no-SQL) approach to data management and requires the manual intervention of the expert-linguist to evaluate and enrich the data collected and classified by the machines. Eventually, such systems as products of technology extend their scope beyond the very study of neology and embrace further empirical and applied objectives both in lexicographic and corpus research. More specifically:
(a) They can natively support dictionary compilation (despite the differences in the degree of lexicographic orientation)

[1] Webometric Analyst Web Impact Reports (http://lexiurl.wlv.ac.uk/index.html).
[2] The *Néoveille* platform (Cartier 2016) has the potential to track and monitor Greek neologisms; the public and the guest interface, however, seem currently not updated for Greek.
[3] The *electronic database of Modern Greek Neologisms* is characterized by its creators as a "lexicographic product" (Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009: 419) and was also used to enrich the macrostructure of the *Reverse Index of Modern Greek Vocabulary* (https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/reverse/index.html).
[4] https://www.upf.edu/web/obneo
[5] https://wortwarte.de
[6] https://www.owid.de/docs/neo/start.jsp
[7] http://neologismen.ivdnt.org/search
[8] http://anw.inl.nl/neologismen

by constituting specialized lexicographic resources themselves[9] or by affiliation or contribution to larger dictionary projects and corpus lexicography.[10] In any case, every research infrastructure implements its own inclusion and exclusion criteria in filtering neologisms for final presentation in a dictionary or collection of neologisms.[11]

(b) Such infrastructures utterly make a strong impact in computational corpus-linguistic research, because of their reliance on monitor corpora or at least very large textual sources to automatically extract, classify and systematically monitor neologisms. The domain of the Press (online versions of newspapers or news feeds) has traditionally provided the starting point for neologism tracking for most of the projects[12] for technological and linguistic reasons: RSS protocols permit easier harvesting of data for corpus compilation and journalistic discourse has better chances of representing institutionalized usage, thus the written standard. Despite the limitations in genre representativeness and balance, nowadays, the multifaced character of online journalism encompasses much more genres and topics than in the past and extends to a wider range of documents (supplements, magazines) supporting the web edition of a newspaper, besides the prototypical news categories. At the same time, we can observe a growing tendency to document neologistic use beyond morphological and word-formation taxonomies, with an emphasis on the otherwise neglected (con)textual variables (genres, topics) and the development of relevant annotation schemes, so valuable in corpus-linguistic research. Finally, a reconciliation of the *web as/for corpus* methodologies is witnessed in newest platforms, such as the *Néonaute*, in line with NEOΔHMIA's (see section 2) early commitment in doing both (Afentoulidou & Christofidou 2017), since there are inherent epistemological disadvantages in assessing linguistic usage both through search engine results and big data architectures. In any case, all projects share a mutual concern in developing more intelligent methods for automated neologism detection – for instance by machine learning – as well as corpus resources with richer annotation layers and open science data policies. In the ever-growing digital landscape of linguistic resources and services, the focus in documenting language has now shifted from the lexicographic exclusion principles to inclusion possibilities or rather "prioritization policies" (Connor Martin 2019) and data analytic frameworks,[13] with a parallel concern to eliminate true data noise and balance recall and precision.

NEOΔHMIA, in a parallel line of research (see section 2) embraces all those challenges and methodological commitments in data collection, has a strong orientation towards morphology and textlinguistics with a view to incorporate semantic/pragmatic approaches and computationally unify, in a single digital ecosystem, research on neology with terminological research.[14]

4 Approaches to the Dictionary Inclusion of Neologisms

There are numerous descriptions and definitions of neologisms and/or neology. According to more strict approaches, a neologism is defined as any lexical unit being classified as a member of the active vocabulary of the speech community (institutionalized or lexicalized lexical units, see below, cf. Hohenhaus 2005: 359-365), based on certain criteria, which would allow its inclusion in general-purpose dictionaries (for criteria see among others Teubert 1998: 131ff.; Herberg, Kinne & Steffens 2004: XII; cf. Klosa-Kückelhaus & Wolfer 2019). According to more open approaches a neologism is defined as any new lexical unit occurring in the oral or written discourse of a certain period (see definition in Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009: 420, cf. Nikolaou & Anastassiadis-Symeonidis 2017: 272). Following an intermediate approach, under the notion of neologism we understand any new lexical unit which exhibits a recent repeatability of use[15] within the speech community, even if it does not meet (yet) the condition of establishment and institutionalization (see below; also, Renouf 2007b). Thus, a decisive factor for the treatment of neology seems to depend on the research objective: a research for the sake of lexicography should follow the first more strict approach while a broader, linguistically driven research should rather follow the other two approaches.

From the three different views above (cf. also the presentation of related work in section 3) it becomes obvious that three respective approaches could be discerned concerning the objectives of neologisms' documentation (see also Guerra 2016):

a. Inclusion of only these new formations which "deserve" it, i.e., they meet certain quantitative and qualitative conditions which ensure that the words are institutionalized or lexicalized[16] (with an indication for the neologicity of the entry), see Klosa-Kückelhaus & Wolfer (2019); Freixa & Torner (2020); Cabré & Nazar (2012); Christofidou (2015); Connor Martin (2019)

[9] For example, the *Neologismenwörterbuch* available at the IDS portal OWID (https://www.owid.de/) and the online dictionary of neologisms of different varieties of Spanish *El Antenario* (https://antenario.wordpress.com/presentacion/).

[10] For example, the so far connection of the late version of *Die Wortwarte* (Lemnitzer 2010) to the enrichment of the *Digitales Wörterbuch der Deutschen Sprache* in the DWDS portal (https://www.dwds.de/), of the *Norwegian Newspaper Corpus* (http://avis.uib.no/) to specific dictionary projects of Norwegian (Andersen 2013), of *Neocrawler* to the Oxford English Dictionary Team, of the *Neologismen Database* to the compilation of the *Algemeen Nederlands Woordenboek* (see section 3).

[11] Cf. the maximalistic all-inclusive *in natu* recording policy of *Die Wortwarte vs.* the conservative, with multiple exclusion criteria, filtering of neologisms to-be-included in the *Neologismenwörterbuch* macrostructure (Klosa & Lüngen 2018).

[12] Cf. the *Neocrawler* system with a different, *web as corpus* approach and *OBNEO's* extended methodology of scanning texts also from magazines and spoken sources to trace neologisms.

[13] A converging trend can be witnessed from the part of lexicographic projects, such as the *OED Oxford Labs Initiative*, with the aim to gain richer insights into language change through the lens of large-scale analysis of the *OED* dataset itself.

[14] The Centre's terminological resources make use of thesaurus building systems and classification schemes based on ontologies.

[15] The criteria to establish repeatability differ according to the specific research objectives.

[16] On institutionalization *vs.* lexicalization and their role to neology there are subtle differentiations among the researchers (see Hohenhaus 2005; Klosa-Kückelhaus & Wolfer 2019; Kerremans 2015: 41).

b. Documentation of any attested new coinage (even nonce formations under certain conditions) in electronic (static/dynamic) databases (among others the German neologisms project *Die Wortwarte*)
c. Documentation of new lexical units (even ephemeral ones) which meet a minimum of – mostly – quantitative conditions for their repeatability in a dynamic (electronic) lexicon of neologisms (and further monitoring)[17]

In the following, we will further discuss the third approach in favour of which we would like to argue: Kerremans (2015: 40, cf. Schmid 2008) proposes a model of the establishment process of a neologism, i.e., she defines three *stages* (creation, consolidation, establishment) which a new formation would undergo within three different *perspectives*: lexicalization (structural perspective), institutionalization / conventionalization (socio-pragmatic perspective) and hypostatization / entrenchment (psycholinguistic perspective). Following partially Kerremans' model (2015) we propose that a new lexical unit should be included in a dictionary (see approach a. above), only if it covers all three *stages* (i.e. creation, consolidation and establishment) from at least the first two *perspectives*, that of lexicalization and institutionalization / conventionalization.[18] As far as linguistic research is concerned, we assume that a new formation should be captured and registered in an electronic dictionary or a dynamic base of neologisms already at the beginning of the *stage of consolidation* (i.e., stabilization of form and meaning) within the first *perspective* of lexicalization (see approach c. above). Moreover, nonce formations (or hapax legomena) are ad hoc formations, which still remain at the first *stage of creation*. To our mind, nonce formations – though linguistically very important – should be treated separately, since ad hoc formations behave dramatically different than formations reaching the *stage of consolidation* (in all three *perspectives*). They often consist in multiword expressions, blends, surface analogy[19] or poetic formations and they are the only formations that can be ungrammatical and exhibit an ad hoc (exclusively context-dependent) meaning (see also Renouf 2007b: 8ff.).

The above proposal seems to be based on two different research views: For a lexicographer it is more important to pursue registering only these new formations which meet the level (*perspective*) of institutionalization / conventionalization, covering all three *stages* from creation to establishment (see above). For a linguist, who investigates the phenomenon of neology (morphology, semantics, language change etc.) it should be crucial to record all new words which show at least a kind of stabilization of form and meaning (*stage of consolidation*). Many new formations, although ephemeral, i.e. not ad hoc but not (yet) dictionarizable, reveal at least the same amount of linguistic information as the successful, thus dictionarizable neologisms, whilst their life-cycle is rather pragmatically and socio-linguistically conditioned (for discussion see Kerremans 2015: 41-43 and Fischer 1998: 178ff.). Thus, they equally provide the linguist with important information on morphological trends (within word families), on semantic evolution, on the text-morphology interface and partially on language change. In this sense the ephemeral, but not (yet) dictionarizable, neologisms constitute a tank of linguistic information, which should be documented for multiple investigation and further monitoring.

Nevertheless, one of the most important contributions of such an approach to the phenomenon of neology per se would be the following: the systematic monitoring, tracking and analysis of the evolution and life-cycle of the majority of new formations, within a specific period in the recent past – regardless of their possible disappearance or success – could provide us with possible estimations for the behaviour of new words in the future, and consequently with suggestions for their dictionary inclusion or not.

In the following sections there will be an attempt, based on data from the corpus component of ΝΕΟΔΗΜΙΑ project (section 2), to show the contribution of the tracking of new lexical units' evolution, according to specific measurements, to the identification of neologisms, either a. for inclusion in a broader dynamic electronic lexicon of neologisms and/or b. for inclusion in a general-purpose dictionary.

5 Corpus-based Development

5.1 Method

5.1.1 Data

In order to monitor the behaviour and the evolution of new lexical units for the purposes described above we decided to track all new formations recorded in the database of ΝΕΟΔΗΜΙΑ, belonging to the Greek debt crisis (2010-2019) vocabulary. Due to our heuristic procedure the only limitation has been a minimum of 100 occurrences in the search engine Google, in order to ensure repeatability and a form-meaning stabilization (checking the context of use for the web occurrences). In addition, we collected all neological lemmas from *The Vocabulary of Crisis* (Varoufakis 2011). This procedure provided us with a list of 32 new lexical units (derivatives, one- and two-word[20] compounds) concerning the Greek debt crisis.

To study the spread and life-cycle of the 32 Greek debt crisis formations we used a sub-part of the *Monitor Corpus of Neologisms* (see section 2). Instead of applying random-sampling techniques to the entirety of newspaper sources crawled, we selected for the purposes of this study the newspaper feed which produced the largest amount of data per year (*Proto Thema*) and made our searches within all collected written content. Since our focus is not on capturing the overall diachronic trend and the fate of those words in Greek society in general (and between newspapers), but our aim is to

[17] Concerning both b. and c.: cf. the Research Programme of the Aristotle University of Thessaloniki, conducted by Prof. Anastassiadis-Symeonidis, presented in Anastassiadis-Symeonidis, Alexiadou & Nikolaou (2009).
[18] The third *perspective* of entrenchment concerns a level of a psycholinguistic approach addressing mostly the perception level.
[19] Such word formation processes also apply to (not ad hoc) neologisms, albeit much rarer.
[20] See Christofidou et al. (2020) on qualitative and quantitative criteria for compoundhood of (new) multiword expressions.

"freeze time" somehow, zoom into their frequency profiles and study how they developed during a specific time span, by prioritizing continuous coverage[21] to source variation, we maximized our chances of providing a representative picture of those novel formations' unique trajectories, of course with a limitation of our observations to the specific newspaper.[22] Moreover, in that way we avoided the thorny issue of having to disentangle from our results topic-related newspaper bias and newspaper-specific coverage, which unavoidably characterises mixed corpora, i.e. of many newspaper sources, and is discussed by Gabrielatos et al. (2012: 162-164) in their study on the peaks and troughs of corpus-based contextual analysis in the UK Press. They witness great differences between the newspapers they study, to the degree of questioning "the utility of examining the development in the number of articles in the corpus as a whole – thus effectively treating British national newspapers as a homogeneous group" (p. 162) and conclude:

> In light of the above, studies of groups of newspapers, taken as a whole, may miss important individual differences. Conversely, studies of individual newspapers can safely generalise only about the particular newspaper. Therefore, if the corpus comprises distinct sub-corpora (in our case, different newspapers), then frequency developments should be examined in those individually as well as in the corpus as a whole. (Gabrielatos et al. 2012: 163-164)

For the purposes of this study, we divided the corpus into monthly sub-corpora, in order to monitor frequency developments over time to a higher level of granularity. As a *terminus post quem* we decided on September 2015, when Greek national elections took place amidst the economic debt crisis, which was then profoundly consolidated in all aspects of life in the country, following the Greek Bailout Referendum of the summer of 2015, when the bailout conditions of the European Union, the IMF and the European Central Bank were rejected. As a *terminus ante quem* we opted for the end of 2020, a year that the spread of the Covid-19 pandemic crisis took over, still reigns supreme – and in any case overshadowed the Greek debt crisis, which has been at the time already softened (the final bailout came to a formal end in the end of 2019). Therefore, a 5.4-year perspective was adopted with a total of 64 successive sampling points (64 months i.e. 64 corpus sections/large XML files) in order to gain a wider scope from the seamless comparison of frequency patterns across time. So the corpus used in this study comprises 552,975 press articles, of various lengths, from the online version of the newspaper *Proto Thema* covering the period from the 1st of September 2015 to the 31st of December 2020, of the total size of about 160 million running words (tokens).

5.1.2 Procedure

Specific (and time-consuming) pre-processing steps were performed semi-automatically to prepare the 160 million tokens corpus for analysis (for instance, article deduplication and removal of repeated content noise, whitespace and non-whitespace character normalization, cleaning of residual CDATA or stripping non-parsable XML entities, conversion to a custom-TEI schema), since the articles were collected (except for the last four months of 2020) using a previous, less automated version of NEOΔHMIA's crawler. Then 32 queries were compiled for each lexical unit, covering all inflectional or spelling variants (a total of 1,055 case-insensitive searches were written, using simple regular expression notation) to be performed with two software packages (Voyant Tools and WordSmith Tools).
The corpus was imported to the Voyant text analysis environment (server edition) using the XML teiCorpus ingest module integrated in Voyant Tools (VTs), with the tokenization parameter set on "Simple Word Boundaries". WordSmith Tools 8.0 (WSTs) produced concordances and enhanced the frequency profiling of the selected neologisms by the computation of dispersion metrics.
Due to time limitations we did not perform manual lemmatization for the lexical units under examination (there were 14,502 occurrences of the search terms in total – see Table 1) and, instead, made use of VTs' search syntax to match items separated with pipes as a single term. Where needed, some spelling variants with hyphens or parentheses, such as *meta-mnimoniakos*, *(meta)mnimoniakos*, *neo-troikanos* were normalized using the TEI element <*reg*> to eliminate noise in the recall of single terms (*mnimoniakos, troikanos*) and make sure that all instances of the 32 formations were correctly retrieved. Finally, the 6 multiword units were annotated and enclosed within the element <*mwu*>, in order to be processed as single items with the WSTs' WordList procedure. For every lexical unit queried with VTs we used the Trends and Terms tools (with the *Relative Frequency* per million, *Peakedness* and *Skew* columns activated besides the default ones – *Count* a.k.a. absolute frequency and *Trends*). The degree of neologism consolidation within a community of discourse receivers (newspaper readers) and producers (journalists, audience commentators) should be captured with the computation of frequency profiles throughout the corpus, as well as the use of time-lined dispersion statistics.

5.2 Corpus Exploration and Analysis

Table 1 presents the quantitative results of the procedure discussed in the previous section. From left to right, *Lemmas* correspond to cumulative searches for each selected new formation. The *Lexical Frequency* values (*Absolute and Relative*) range from one occurrence (two hapaxes*, chreofreno*, *dimokratoria*) to 3,554 hits (22.5 words per million for *mnimoniakos*) and are rather low if we take the size of the corpus into consideration. The *Peakedness* statistic measures

[21] Due to technical reasons that interrupted data collection, 75 days are missing, but the gaps are spread across 1,874 days in total.
[22] According to Alexa's site rankings for Greece (https://www.alexa.com/topsites/countries/GR), the online version of *Proto Thema* (https://www.protothema.gr) was, in 2015, and still is (as of March 2021) first among all other Greek online newspapers ranked by the specific service for overall traffic calculated by the combination of daily visitors and pageviews (thus indirectly measuring degree of readability). Moreover, the newspaper addresses a wider audience, publishes on a diverse range of topics besides the central, typical news categories (great emphasis is given on popular topics such as lifestyle and celebrities, psychology, entertainment, cooking etc.) and produces an overall multi-genre inventory of resources (besides multimedia content, there is a constantly updated blog section with point of view articles, advertorial columns, verticals, recipes, connection with magazines etc.).

how much the relative frequencies of the lemmas are bunched up into peaks, whereas a peak is defined as a region with high values, where the rest are lower.[23] The peaks are formed when there are extreme differences between documents (i.e. corpus periods) and represent significant outliers, that is discourse fluctuations and instability due to topicality. Large spikes denote uneven patterns of sudden rises/falls in usage. In essence, the values in Table 1 provide an overall kurtosis estimation.[24] *Skew* is a statistical measure of the symmetry (skewness) of the relative frequencies. A positive skew is formed when the mass of the distribution is concentrated on the left and the right tail is longer, suggesting that the overall frequency profiles follow a declining path, irrespective of periods of regression. A skew value approaching zero corresponds to usage consistency, whilst negative values would mean that frequencies started low and increased. The *Peakedness* / *Skew* metrics, therefore, holistically evaluate the *shape* of the frequency curve, thus usage intensity.

Lemma	Absolute Frequency	Relative Frequency (pmw)	Peakedness	Skew	Sparkline	Dispersion	Kendall's τ coef/ent	Trend [**/* Correlation is significant at the 0.01/0.05 level (2-tailed)]
ftochopiisi	558	3.534884	-0.04406605	0.6568374		0.83	-0.481	Slow decline** ↘
metamnimoniakos	1324	8.387431	0.92924464	1.4513253		0.59	0.209	Slight increase* ↗
pragmatiki_ikonomia	1199	7.595566	1.0510801	1.0915446		0.92	-0.255	Slow decline** ↘
eyroieratio	121	0.766525	2.196963	1.7333602		0.67	-0.434	Slow decline** ↘
ftochopio	286	1.811786	2.415639	1.3975376		0.75	-0.201	Slow decline* ↘
troikanos	114	0.722181	3.4694684	1.809957		0.73	-0.338	Slow decline** ↘
merkelistis	30	0.190048	3.81772	1.9048449		0.79	-0.135	Stable usage (decreasing) →
esoteriki_ypotimisi	90	0.570143	7.9706416	2.4776852		0.67	-0.384	Slow decline** ↘
titlopiisi	422	2.673335	8.054505	2.2520626		0.7	0.503	Moderate increase** ↗
mnimoniakos	3554	22.5143	8.213318	1.9867238		0.78	-0.581	Moderate decline** ↘
ypertamio	1570	9.945821	9.340388	2.8904257		0.77	-0.246	Slow decline** ↘
domimeno_omologo	55	0.34842	9.52648	2.9386773		0.65	-0.316	Slow decline** ↘
trapezokratia	13	0.082354	9.68198	3.1034594		0.58	-0.018	Stable usage (decreasing) →
posotiki_chalarosi	1696	10.74402	9.995694	2.9633954		0.61	-0.352	Slow decline** ↘
ithikos_kindynos	44	0.278736	10.819942	2.7555737		0.71	0.060	Stable usage (increasing) →
eyrofovikos	82	0.519463	11.275036	3.0954626		0.67	-0.119	Stable usage (decreasing) →
apikia_chreoys	43	0.272401	16.617157	3.6415455		0.65	-0.268	Slow decline** ↘
chreokratia	5	0.031675	19.061302	4.3090296		0.47	-0.227	Slow decline* ↘
anakefaleopiisi	1952	12.36576	20.783533	4.4174266		0.78	-0.706	Sharp decline** ↓
ypermnimonio	3	0.019005	24.487204	4.9032397		0	-0.232	Slow decline* ↘
menoymeyropeos	31	0.196382	25.19234	4.7880397		0.67	-0.028	Stable usage (decreasing) →
stasimochreokopia	50	0.316746	25.698536	4.445233		0.7	-0.249	Slow decline** ↘
eyrokratia	2	0.01267	29.37236	5.518538		0.35	0.193	Stable usage (increasing) →
eyroarnitismos	2	0.01267	29.39645	5.5200977		0	-0.201	Slow decline ↘
antimnimonio	102	0.646162	30.239964	4.8551636		0.69	-0.259	Slow decline** ↘
eyroomologo	389	2.464283	37.208927	5.864297		0.78	-0.101	Stable usage (decreasing) →
antimnimoniakos	752	4.763858	40.01884	5.821955		0.73	-0.589	Moderate decline** ↘
merkelismos	3	0.019005	42.524155	6.3934593		0	-0.098	Stable usage (decreasing) →
antimerkelistis	6	0.03801	62.433693	7.8648567		0.17	-0.090	Stable usage (decreasing) →
chreofreno	1	0.006335	64	8		0	0.177	Stable usage (increasing) →
dimokratoria	1	0.006335	64	8		0	-0.003	Stable usage (decreasing) →
germanopio	2	0.01267	64	8		0	-0.076	Stable usage (decreasing) →

Table 1: Frequency distribution / development of the search terms in the corpus during the 64-month period (sorted on *Peakedness*).

Sparkline graphs are generated for each query, namely the concordance hits are visually represented as lines that show the distribution of relative frequencies across the chronologically ordered corpus documents, followed by the *Dispersion* statistic, which is the Juilland's *D* implementation of WSTs and is computed with the WordList tool. Due to the absence of lemmatization, for every lemma, only the highest dispersion value amongst all variants is displayed in Table 1 to represent the degree to which frequencies are evenly spread (maximum value=1, suggesting uniform dispersion | minimum value=0, suggesting burstiness).[25] The last two columns introduce a further abstraction: the detection of trends in the data, by correlating the observed relative frequencies with the sequence of the different temporal stages. Following Hilpert & Gries (2009: 388-390), Kendall's τ correlation coefficients and their p-values are produced for each lemma. Values close to 0 indicate the absence of a trend, values approaching -1/+1 indicate sharp decrease/increase.

[23] See VTs Help. WSTs implement a *Peakiness* sorting function of time-lined concordances to graphically display outliers in frequency development within lemmas, where "Peakiness uses the standard deviation of the proportion of hits to word count in each period of a time-line", but the scores per search word are not displayed for a between-lemmas comparison.
[24] High positive values correspond to leptokurtic distributions (extreme fluctuations) lower to mesokurtic and negative to platykurtic.
[25] Aggregated – thus more accurate – dispersion values were also computed with WSTs (generation of time-lined dispersion plots for every lemma but only for the text files in which the search terms appeared, see WSTs Help). After examination of the results, the overall trend was the same, so we opted for the first method of calculation, i.e. using all the files of the corpus.

As the frequency profiles show, overall, there is no considerable pattern of growth for the selected Greek debt crisis new terms at the end of 2020 (especially after August 2019, on close inspection of all the time-graphs), as compared to the beginning of the period of observation (last quarter of 2015). All distributions are positively skewed (with varying degrees of skewness), which means that in principle the words are already in use in 2015, thus seem to undergo a stabilization process (notably *anakefaleopiisi, mnimoniakos, antimnimoniakos, pragmatiki ikonomia, ftochopiisi, troikanos*) but then follow a declining path, most of them fade away with sudden rises and regressions. Some neologisms are visibly attested at the end of the period (*eyroomologo, titlopiisi, chreofreno*) displaying the prototypical exponential curve of neologism diffusion (Cabré & Nazar 2012) but their frequency development, within the window of our observation does not seem to stabilize on a steady trajectory.

Furthermore, when lemmas are sorted on *Peakedness* (see Table 1), we can observe the following pattern: the least peaky neologisms, irrespective of overall frequency, dispersion and direction of change (upwards, downwards) are those with the fuller and more "resilient" life-cycles, whether presently still evolving (thus with the best chance of "survival", such as *ftochopiisi, pragmatiki ikonomia*) or at a time solidly evolved (such as *metamnimoniakos*).[26] On the contrary, as we move up the scale, neologisms with higher *Peakedness* values, thus greater fluctuations and temporal instability, irrespective of overall frequency, dispersion and direction of change (upwards, downwards) are the most transient and their use is purely topical. Thus, in developing practical heuristics regarding the most important quantitative features of successful neologisms undergoing lexicalization (stabilization of form and meaning), the dynamic notion of *Peakedness,* as a predictor that affects "lexical sustainability", seems promising to explore.

Figure 1 plots the Kendall's τ correlation coefficient values of Table 1 in the horizontal axis and the relative frequencies for every lemma a. at the beginning and b. at the end of the period of observation in the vertical axis (see Grieve, Nini & Guo 2016 for a similar methodology to detect emerging word forms in English).

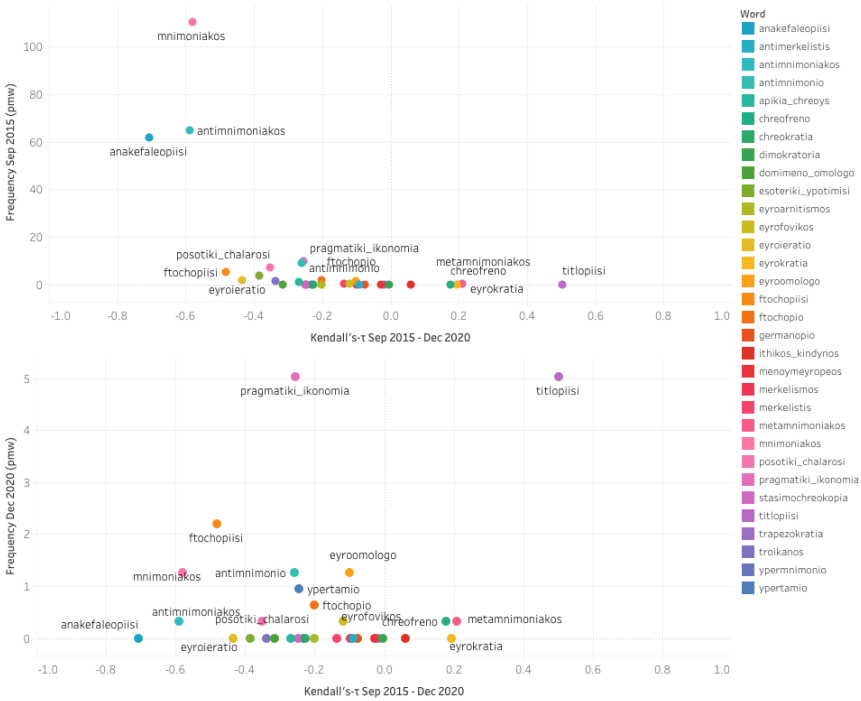

Figure 1: Kendall's τ coefficient *vs.* September 2015 and December 2020 relative frequencies.

The vast majority of the selected formations share a declining trend (see also Table 1). In September 2015, *mnimoniakos, antimnimoniakos, anakefalaiopiisi* were frequently used and kept spreading, but in December 2020 a statistically significant decline in their usage is observed, implying that they did not eventually stabilize in newspaper popular discussions. It is only the least peaky neologisms, *pragmatiki ikonomia* and *ftochopiisi* that reign supreme, remain stable and can be safely considered best candidates "to have marched the long way" towards an *establishment stage*. On the other hand, a few Kendall's τ values are positive, displaying a mild, almost stable upward pattern, but with high *Peakedness* and low *Dispersion* scores (see Table 1). Only *titlopiisi*, as we observe in Figure 1, at the end of 2020 is characterized by rapid growth, namely the formation followed a clear emerging – under way of stabilization – trajectory, as it is also suggested by its middle-range *Peakedness* score. Conversely, *metamnimoniakos* seems from Figure 1 alone, to be "frozen" into an ever-emerging state. Its *Peakedness* values, however, predict otherwise. In fact, *metamnimoniakos* displays the second less peaky frequency development across the period under observation and the mere shape of its

[26] *Eyroierateio,* seems to be a successful, consistent, albeit newspaper-specific preference, since *I Kathimerini,* the second largest newspaper in the corpus disfavors the use of this emotionally-vivid formation. Similar searches were performed for the rest of the sources collected during the specific period (*Ta Nea, To Vima, Ethnos, I Avgi*) and confirm that analysis. A cumulative frequency from both newspapers would then distort this stylistic preference of *Proto Thema*, so we can reliably categorize it as an outlier.

frequency distribution reveals that its usage has not only grown consistently *over a narrower time frame* but also formed a rather stable plateau from, roughly, the summer of 2017 until the summer of 2019, with small regressions until the end of February 2020. At the same time, the *Dispersion* values are rather low, showing uneven distribution across the corpus parts. *Metamnimoniakos* thus seems to have completed a full life-cycle with a shorter life-span than the one we set *a priori* in this study. Had we considered a shorter time-frame of examination, low *Peakedness* would suggest stability and consistency. In other words, a *Peakedness* estimation can also be used *retroactively* to predict subsequent "lexical sustainability" and we argue that the measurement reflects strong on-demand communicative needs (see Discussion).
The candlestick graph in Figure 2 mirrors the lower part of Figure 1 and summarizes frequency development, following Brezina's (2018) adaptation of this type of data visualization used in financial reports to corpus linguistics. The boxes visually represent the y axis of Figure 1 (initial point, September 2015 *vs.* final point, December 2020) and the colour shows when the frequencies were higher (at the beginning – red box – or at the end – blue box). The spikes represent the minimum and the maximum frequencies throughout the 5.4 years. The longer the body of a candle is, the greater is the variation in frequency profiles. The spikes denote frequency fluctuations (when projected outside the box) and, the longer they are in relation to the box and themselves (upper and lower wicks), the less smooth the transitions are. The candles for *pragmatiki ikonomia* and *ftochopiisi* are almost symmetrical if we compare them with the rest.[27] For *metamnimoniakos,* the beginning of its path is also the end and the positive spike in between fits in its whole life-cycle.

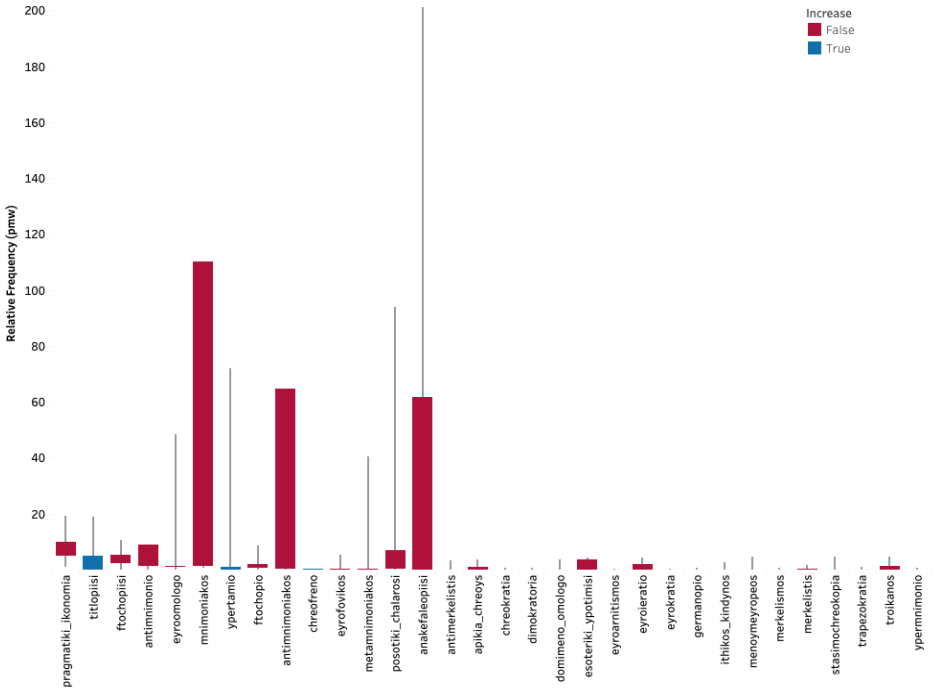

Figure 2: Candlestick graph ordered by declining relative frequencies as of December 2020.

6 Discussion

In the previous section we tracked the frequency development of selected new formations coined during the Greek debt crisis to observe their behavior within a time span of 5.4 years (the end of the crisis) and gain a more fine-grained understanding of their life-cycles within that period. The mere shape of their distribution in the corpus was shown to form a continuum of cases and permitted a glance into the dynamics of the spread process. Although overall, their appearance in journalistic discourse is decaying,[28] the less peaky distributions traced longer paths into the future and characterized the more resilient neologisms. Conversely, high *Peakedness* was an indicator of instability and transient, topical, thus ad hoc uses. These patterns, of course, derive from the different weight of communicative needs that triggered word usage towards the end of the Greek debt crisis and the beginning of the pandemic in 2020. Some new terms were only produced on-demand and never left the *consolidation* (or even the creation) *stage*; high peaks seem to negatively affect the spread of new formations. Others, although slowly diminishing in use, have endured and entered the *establishment stage*; low peaks seem to positively affect the spread of new formations. Kurtosis measures, such as the *Peakedness* values can be used as a top-down, dynamic quantitative filter to monitor "lexical sustainability", together with similar metrics, such as *Dispersion.* As distributional evidence showed (see Table 1), extremely peaky lemmas have very low *Dispersion*, since they not uniformly spread in the corpus. Not all less peaky lemmas, however, are more uniformly spread (see Table 1, *metamnimoniakos*). Moreover, there are peaky lemmas which are indeed uniformly spread (see Table 1, *anakefaleopiisi)*. Therefore, *Dispersion* measures partially correlate with *Peakedness* scores in an inverse relationship (r_s = -0.575, $p_{2\text{-tailed}}$

[27] They almost resemble the "Spinning Tops" candlestick pattern in financial jargon, representing little movement in the market.
[28] At least for the specific newspaper we selected on theoretical grounds (see section 5.1).

= 0.001). In fact, since *Dispersion* measurements are obviously affected by the duration of the time span (on a horizontal view of the data), they can only be used complementary to *Peakedness* evaluations (an essentially vertical view of the data), for instance as an initial cut-off threshold of under-dispersed, thus ephemeral new formations (for Table 1, see *Dispersion* scores ≤ 0.60).

We nonetheless emphasize the potential of such tangible criteria that corpus-linguistic methods and tools offer and their diagnostic (as if prognostic, for *Peakedness*) value in assessing the "success stories" of different new formations in their way to establishment in a community of language producers / receivers. Once fine-tuned empirically they can contribute to the development of solid prediction models (cf. Jiang et al. 2021) or simply serve as practical heuristics complemented by corpus-based, bottom-up lexicographical assessment.

7 Concluding Remarks and Future Research

In this paper there has been an attempt to conjointly illustrate the importance of quantitative explorations and measurements, like *Peakedness, Dispersion* etc., applied on a newspaper corpus for a selected list of new formations – designating aspects of the Greek debt crisis and covering certain criteria – in order to co-estimate their behaviour and evolution retroactively in time. Specific tools and statistical procedures were used, and it was shown that *Peakedness* was an important indicator for the sustainability of emerging formations. In addition to this, we assume that text type and media diversity should be a second crucial factor for the success of the new lexical units. Thus, the same approach, i.e. a retroactive analysis to lists of new formations of certain periods should be further applied to the entire corpus, including the rest of the newspapers, in order to also detect and evaluate their diffusion to the speech community, thus the beginning of institutionalization, the key notion for dictionary inclusion.

The results of the corpus exploration and analysis can prove both linguistically and lexicographically very profitable. It seems to be important for linguistic research to identify and register new formations – exhibiting a certain repetitive use and a form-meaning stabilization (thus entering the *consolidation stage*) – in a dynamic electronic lexicon of neologisms in order to monitor their behaviour and evolution for certain selected periods. Although many of these registered new formations may not yet be at the *stage of establishment* within the *perspective of institutionalization* (Kerremans 2015: 40, see Schmid 2008), which according to our presentation (see section 4) would signal the step for inclusion in general-purpose dictionaries, their monitoring seems a. to build a valuable linguistic information tank and b. to further facilitate the answer to the desideratum of the inclusion (or not) decision for neologisms.

Our proposal for a dynamic electronic lexicon of neologisms is being supported by the evolution and the enormous possibilities of corpus linguistics and electronic lexicography. Both fields contribute to the investigation, monitoring and recording of a huge amount of data. Taking advantage of these new possibilities the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* is planning to expand its research area from neologisms for inclusion in general-purpose dictionaries to the construction of a broader dynamic lexicon of (possible) neologisms.

8 References

[Afentoulidou & Christofidou] Αφεντουλίδου, Β. & Χριστοφίδου, Α. (2017 [2018]). Σώμα Εποπτείας Νεολογισμών της Νέας Ελληνικής: Σχεδιασμός και κειμενική ταξινόμηση. Στο Α. Χριστοφίδου (επιμ.) *Όψεις της σωματοκειμενικής γλωσσολογίας: Αρχές, εφαρμογές, προκλήσεις. Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 14, σσ.129-180. Αθήνα: Ακαδημία Αθηνών.

[Anastassiadis-Symeonidis] Αναστασιάδη-Συμεωνίδη, Α. (2003). *Αντίστροφο λεξικό της Νέας Ελληνικής*. Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

[Anastassiadis-Symeonidis, Alexiadou & Nikolaou] Αναστασιάδη-Συμεωνίδη Α., Αλεξιάδου, Χ. & Νικολάου, Γ. (2009). Ηλεκτρονική βάση νεολογισμών της Νέας Ελληνικής. Στο Α. Χριστοφίδου (επιμ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 9-10, σσ. 419-439. Αθήνα: Ακαδημία Αθηνών.

[Christofidou] Χριστοφίδου, Α. (2015). Εισαγωγή. Σχεδιασμός και παρουσίαση της έρευνας. Στο Α. Χριστοφίδου (επιμ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 13, σσ.11-24. Αθήνα: Ακαδημία Αθηνών.

[Christofidou, Afentoulidou, Karasimos & Dimitropoulou] Χριστοφίδου, Α., Αφεντουλίδου, Β., Καρασίμος, Θ. & Δημητροπούλου, Ε. (2013). Ηλεκτρονικό πρόγραμμα Νεοδημία. Προκλήσεις και δικτυο-λύσεις. Στο Α. Χριστοφίδου (επιμ.) *Δημιουργία και μορφή στη γλώσσα, Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 12, σσ. 198-243. Αθήνα: Ακαδημία Αθηνών.

[Christofidou, Karasimos & Afentoulidou] Χριστοφίδου, Α., Καρασίμος, Θ. & Αφεντουλίδου, Β. (2014). Έλεγχος, παρακολούθηση και ταξινόμηση νεολογισμών με το ηλεκτρονικό πρόγραμμα *Νεοδημία*: Η προσέγγιση των νέων δανείων. Στο G. Kotzoglou, K. Nikolou, E. Karantzola, K. Frantzi, I. Galantomos, M. Georgalidou, V. Kourti-Kazoullis, C. Papadopoulou, E. Vlachou (επιμ.) *Selected Papers of the 11th International Conference on Greek Linguistics,* σσ. 1850-1868. Rhodes: University of the Aegean.

[Nikolaou & Anastassiadis-Symeonidis] Νικολάου, Γ. & Αναστασιάδη-Συμεωνίδη, Α. (2017). Ο ρόλος του παγκόσμιου ιστού στη μελέτη της νεολογίας και της μορφολογικής ανάλυσης: Η περίπτωση των νεολογικών επιθέτων της Κοινής Νεοελληνικής. Στο Α. Χριστοφίδου (επιμ.) *Όψεις της σωματοκειμενικής γλωσσολογίας: Αρχές, εφαρμογές, προκλήσεις. Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 14, σσ. 271-285. Αθήνα: Ακαδημία Αθηνών.

[Varoufakis] Βαρουφάκης, Γ. (2011) *Κρίσης λεξιλόγιο. Οι οικονομικοί όροι που μας καταδυναστεύουν*. Αθήνα: Ποταμός

Andersen, G. & Hofland, K. (2012). Building a large corpus based on newspapers from the web. In G. Andersen (ed.) *Exploring newspaper language.* Amsterdam/Philadelphia: John Benjamins, pp. 1-28.

Aubry, S, Cartier, E. & Stirling, P. (2018). Néonaute: Mining web archives for linguistic analysis. Presentation at the *International Internet Preservation Consortium Web Archiving Conference, 12-15 November 2018*. Wellington, NZ.

Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.
Cabré, T. & Nazar, R. (2012). Towards a new approach to the study of neology. In *Neologica,* 6, pp. 63-80.
Cartier, E. (2016). Néoveille, système de repérage et de suivi des néologismes en sept langues. In *Neologica,* 10, pp.101-131.
Cartier, E. (2019). Néoveille, plateforme de détection, de repérage et de suivi des néologismes en corpus dynamique. In *Neologica,* 13, pp. 23-54.
Christofidou, A., Afentoulidou, V., Karasimos, A. & Vassiliadou, R. (2020). Compoundhood: Defining, extracting and monitoring multiword A+N compounds in a database of Greek neologisms. In S. Markantonatou, A. Christofidou (eds.) *Multiword expressions: Drawing on data from Modern Greek and other languages. Bulletin of Scientific Terminology and Neologisms,* 15. Athens: Academy of Athens, pp. 133-192.
Connor Martin, K. (2019). New Words Prioritization Engine: A system for evaluating multiple data inputs to prioritize neologisms for inclusion in dictionary projects. Presentation at *Globalex 2019: Workshop on Lexicography and Neologism, 8 May 2019*. Bloomington Indiana.
De Schryver, G.-M. (2002). Web for/as corpus: A perspective for the African languages. In *Nordic Journal of African Studies,* 11(2), pp. 266-282.
Fischer, R. (1998). *Lexical change in present-day English. A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms*. Tübingen: Gunter Narr Verlag.
Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionarization of new words in Spanish. In *Dictionaries,* 41(1), pp. 131-153.
Gabrielatos, C., McEnery, T., Diggle, P., Baker, P. & ESRC (Funder). (2012). The peaks and troughs of corpus-based contextual analysis. In *International Journal of Corpus Linguistics*, 17(2), pp. 151-175.
Gérard, C., Falk, I. & Bernhard, D. (2014). Traitement automatisé de la néologie : Pourquoi et comment intégrer l'analyse thématique ? In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschäfer, S. Prévost (eds.) *Actes du 4e CMLF, Berlin, 19-23 July 2014*, SHS Web of Conferences 8. France: EDP Sciences, pp. 2627-2646.
Grieve, J., Nini, A. & Guo, D. (2016). Analyzing lexical emergence in Modern American English online. In *English Language and Linguistics,* 21(1), pp. 99–127.
Guerra, A. R. (2016). Dictionaries of Neologisms: A review and proposals for its improvement. In *Open Linguistics*, 2, pp. 528-556. Accessed at doi.org/10.1515/opli-2016-0028 [28/02/2021].
Hagen, T. M. (2012). Automatic topic classification of a large newspaper corpus. In G. Andersen (ed.) *Exploring newspaper language.* Amsterdam/Philadelphia: John Benjamins, pp. 111-130.
Herberg, D., Kinne, M. & Steffens, D. (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen.* Berlin: de Gruyter.
Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. In *Literary and Linguistic Computing*, 24(4), pp. 385-401.
Hohenhaus, P. (2005). Lexicalization and institutionalization. In P. Štekauer, R. Lieber (eds.) *Handbook of word-formation*. Dordrecht: Springer, pp. 353-373.
Jiang, M., Shen, X. Y., Ahrens, K. & Huang, C.-R. (2021). Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008-2016. In *PLoS ONE,* 16(2). Accessed at doi:10.1371/journal.pone.0245984 [28/02/2021].
Kerremans, D. (2015). *A web of new words. A corpus-based study of the conventionalization process of English neologisms*. Frankfurt: Peter Lang Edition.
Kerremans, D., Prokić, J., Würschinger, Q. & Schmid, H.-J. (2018). Using data-mining to identify and study patterns in lexical innovation on the web: The Neo Crawler. In *Pragmatics and Cognition*, 25(1), pp. 174-200.
Klosa-Kückelhaus, A. & Wolfer, S. (2019). Considerations on the acceptance of German neologisms from the 1990s. In *International Journal of Lexicography*, 33(2), pp. 1-18. Accessed at doi:10.1093/ijl/ecz033 [28/02/2021].
Lüdeling, A., Evert, S. & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf, C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam/New York: Rodopi, pp. 7-24.
Renouf, A. (2007a). Corpus development 25 years on: From super-corpus to cyber-corpus. In R. Facchinetti (ed.) *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, pp. 27-49.
Renouf, A. (2007b). Tracing lexical productivity and creativity in the British Media: 'the Chavs and the Chav-Nots'. In J. Munat (ed.) *Lexical creativity, texts and contexts*. Amsterdam/Philadelphia: John Benjamins, pp. 61–89.
Renouf, A. (2013). A finer definition of neology in English: The life-cycle of a word. In H. Hasselgård, J. Ebeling, S. O. Ebeling (eds.) *Corpus perspectives on patterns of lexis*. Amsterdam/Philadelphia: John Benjamins, pp. 177-208.
Renouf, A. (2016). Big data and its consequences for neology. In *Neologica,* 10, pp. 15-37.
Schmid, H.-J. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. In *Anglia. Zeitschrift für Englische Philologie,* 126(1), pp. 1-36.
Teubert, W. (1998). Korpus und Neologie. In W. Teubert (ed.) *Neologie und Korpus*. Tübingen: Gunter Narr Verlag, pp. 129-170.
Thelwall, M. (2018). *Big data and social web research methods*. University of Wolverhampton. Accessed at http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf [28/02/2021].

Acknowledgements
This research is funded by the Research Committee of the Academy of Athens.