

Kyriaki Salveridou/Zoe Gavriilidou

COMPILATION OF AN ANCIENT GREEK – MODERN GREEK ONLINE THESAURUS FOR TEACHING PURPOSES: MICROSTRUCTURE AND MACROSTRUCTURE

Abstract To effectively design online tools and develop sophisticated programs, for the teaching of Ancient Greek language, there is a clear need for lexical resources that provide semantic links with Modern Greek. This paper proposes a microstructure for an online Ancient Greek to Modern Greek thesaurus (AMGthes) that serves educational purposes. The terms of this bilingual thesaurus have been selected from reference Ancient Greek texts, taught and studied during lower and upper secondary education in Greece. The main objective here is to build a semantic map that helps students find relevant and semantically related terms (synonyms and antonyms) in Ancient Greek, and then provide a rich set of suitable translations and definitions in Modern Greek. Designed to be an online resource, the thesaurus is being developed using web technologies, and thus will be available to every school and university student that pursues a degree in digital humanities.

Keywords Online thesaurus; bilingual thesaurus; Ancient Greek; pedagogical lexicography

1. Introduction

The conceptual approach constitutes the core of what we call conceptually organized lexica, onomasiological dictionaries or thesauri, all of which are based on ontologies. In such ontologies the world is structured in hierarchically organized thematic fields. Every field represents a node in the hierarchy that groups together related linguistic signs. These signs are essentially terms linked to each other with semantic relations such as hyponymy, hypernymy, synonymy etc. The goal of lexica that host such structures is to support users in understanding the semantic value of these relations and enable them to access a rich collection of terms. This in turn will help them identify the words that exactly capture their thoughts, upgrading their vocabulary in the process. Furthermore, there is a clear need for tools that serve as conceptual dictionaries and thesauri. The availability of such tools varies significantly according to the language, scope and targeted function (i. e. educational purposes). Specifically, Ancient Greek lacks essential linguistic digital resources. The question arises: How can we use modern technologies to enrich conceptually organized dictionaries?

This paper discusses the design features, the structure and the implementation of an online bilingual thesaurus for Ancient – Modern Greek (AMGthes). The goal is to deliver a web resource that is oriented towards supporting educational functions. In greater detail, the proposed thesaurus aims to aid the better understanding, accurate interpretation and in-depth analysis of Ancient Greek texts, by both high school and university students (humanities).

The paper is organized in the following sections: Section 2 offers the motivation for the creation of the AMGthes along with capturing relevant research. Section 3 provides the basic concepts and the context of the theoretical background. Section 4 describes the methodology that was followed and the process of selecting and filtering the corpus. This section

also analyzes the macrostructure and the microstructure of the thesaurus, explaining the deciding features and components respectively. Section 5 discusses the details for the design and development of the online resource and finally section 6 presents our conclusions and the next steps planned.

2. Motivation for the creation of the AMGthes

The value of dictionaries and the impact of their use during learning, is well established by relevant research (Atkins 1998; Dziemianko 2012a; Gavriilidou/Konstantinidou 2021).

Bilingual thesauri, on the other hand, are a valuable tool in the hands of students and the whole educational context can greatly benefit from such lexical resources, especially when their continuous improvement both in terms of accuracy and content is consistently worked upon. Modern thesauri and dictionaries are powered by advanced implementations and can now be accessed as online platforms that effectively support the learning context of any scale or complexity. Making academic versions of such tools more available to scholars and better adapted to their needs, will help address fragmentation within the pedagogical discipline and fuel the research of new solutions, thus allowing a methodological shift (Bizzoni et al. 2014; Berti et al. 2016).

While significant steps have been made, the availability of digital resources capable of supporting the teaching of Ancient Greek are still limited and not actively maintained. Modern versions (web-based, online) for established tools and platforms are not being developed, since most are treated as legacy systems with rare or no updates. More importantly, their content is not often curated with scheduled contributions from the community, thus reducing the time window of their validity. In order to gain momentum and increase their user base, digital lexical tools for the Ancient Greek language have to be systematically enriched with new knowledge and upgraded to utilize modern technologies and access patterns (i. e. web or mobile). Attempting to address this need, our research delivers an online bilingual thesaurus that hosts the vocabulary wealth from the corpora that currently comprises the syllabus for teaching the Ancient Greek language at Greek schools. AMGthes adopts a widely known semantic classification scheme, to power its taxonomy, and employs no pivoting language, to facilitate the translation of the selected terms and making it an integral part of their learning process.

3. Theoretical background

The design of the proposed bilingual thesaurus was based on two well-established and widely used thesauri with significant value and impact on many modern linguistic resources: the Roget's Thesaurus (1911) for the English Language and the Antilexicon (1962) by Vostantzoglou for the Greek Language.

In 1852 Peter Mark Roget's Thesaurus, known as the Roget's Thesaurus of English words and phrases, occupied a prominent position in English Lexicography. In this thesaurus, words and terms are not presented alphabetically, but instead grouped together according to the semantic relations that bind them (Hüllen 2004). Roget's Thesaurus comprises one of the largest collections of semantically organized terms, and Roget's name has become a synonym for thesaurus (Jarmasz 2012).

Adopting Roget's classification theme for organizing terms and concepts, the Antilexicon by Vostantzoglou (1862) is a reference Greek lexical resource and a typical example of a similarly structured thesaurus. The Antilexicon is considered a versatile linguistic tool that facilitates effective learning of Greek language. A key feature of this thesaurus is that Vostantzoglou extensively analyzes each term, managing to populate a rich profile with many correlations and semantic links (Trapalis et al. 2005). At its core, linguistic material is grouped into categories of semantically related words, distinguished by individual properties. The Antilexicon comprises a total of 1.500 such categories and interpretations are provided along with both grammar and syntax, while Roget's thesaurus offers no such type of information.

In order to compile the AMGthes, we also focused on tools that capture and analyze the concepts and semantics of words, such as WordNet. This lexical database included English nouns, verbs, adjectives and pronouns, all semantically organized in sets of synonyms, named synsets (Miller/Fellbaum 2007). From relevant previous research, the Ancient Greek WordNet (AGWN) is the first attempt to create the WordNet version for the Ancient Greek language. This version followed the main paradigm of Princeton WordNet (PWN), while adopting features from other WordNets such as the Italian and the Latin versions. This approach resulted in a thesaurus where users can access the synsets' equivalents for each different language (Bizzoni et al. 2014).

AGWN started with evaluating and analyzing bilingual dictionaries to bootstrap Greek – English pairs. As a first step, Greek words were linked to PWN synsets that included the respective English translations. An issue that emerged from using this approach was the invalid overpopulation of synsets and semantic relations, due to the multifaceted nature of English homonymy and polysemy. Furthermore, AGWN is not tailored to effectively serve the learning requirements and modern needs of Greek students. This motivated us to study and create a new online thesaurus, whose design and development are discussed and analyzed in the following sections.

4. Methodology

According to Gouws/Prinsloo (2010) the lexicographic protocol or dictionary plan is a workflow of well-defined tasks, which includes two main components: the organization plan and the dictionary conceptualization plan. Wiegand (1998, p. 151) describes the dictionary conceptualization plan as a process of five distinct phases: the general preparation phase, the material acquisition phase, the material preparation phase, the material processing phase and the publishing preparation phase. Adapting this workflow, Klosa (2013) offers a computer-lexicographical process for the development of online dictionaries. Her research examines a plan that includes: the phase of preparation, the phase of data acquisition, the phase of computerization, the phase of data processing, the phase of data analysis and the phase of preparation for online release.

Considering the above protocols, the compilation of AMGthes follows Klosa's process phases, while also adopting the planning for continuous maintenance and improvement of both the dictionary and its online implementation. All the decisions relevant to the design and features of AMGthes macrostructure and microstructure were made in accordance with this protocol and its methodological principles. Benefiting from modern lexicographical practices and methods, AMGthes is based on a workflow that determines how best to plan, develop, control and deliver its content in an online format. Currently, its progress reached the data

preparation phase for online release, with an alpha version already available and pilot tested. Apart from finalizing and testing its implementation, our work now focuses on conducting multiple sprints of data processing and analysis for the expansion of its vocabulary with new terms.

4.1 Selecting the entries

The first stage of compiling the AMGthes vocabulary included tasks that focused on extracting, collecting and capturing the initial set of terms, from the Ancient Greek texts that comprise the teaching curriculum of Greek middle and high school courses. These tasks span across the phases of data preparation, acquisition, and computerization.

As part of the preparation phase, a set of school courses were selected, forming the official syllabus for learning the Ancient Greek language. Each course includes the analysis and interpretation of texts, from the work of famous Ancient Greek philosophers, poets, orators and historians. In terms of data acquisition, all the appropriate texts were accessed and retrieved from Photodentro – the Greek National Aggregator of Educational Content.¹ A corpus was created from the texts of the appropriate courses: i) For all three grades of middle school this includes excerpts from various Ancient Greek authors. ii) For the 1st grade of high school, excerpts from Xenophon’s *Hellenica* and Thucydides’ *History of the Peloponnesian War*. iii) For the 2nd grade of high school, the *Speech of Lysias for Manditheos*. iv) For the 3rd grade of high school excerpts from the book *Philosophical Reason*, including among others Aristotle’s *Politics* and *Nicomachean Ethics*, Protagoras and Plato’s *Republic*. *Odyssey*, *Iliad*, *Antigone* and other subjects, while being part of the curriculum, they are taught as translated texts and thus were not included.

Completing the computerization phase, the above corpora was processed, organized and stored into multiple text files, properly formatted to act as the input of AntConc,² a text analysis software tool. A starting set of 5.566 unique words was retrieved from the text analysis, acting also as the first task of the data analysis phase. These words were filtered and reviewed for further refinement, detecting misspelled duplicates and other minor mistakes. The resulting list was then studied, identifying the lexical category of each word and bringing it to its basic or citation form. In open lexical categories, i) the citation form of a noun is the singular nominative, ii) the citation form of a verb is the first person singular, present tense in active voice, and iii) the citation form of an adjective is masculine singular nominative. Any category other than verbs, nouns, adjectives and pronouns were filtered out, creating a final set of 2.052 candidate words for the AMGthes vocabulary. These steps can be classified as tasks for both data processing and data preparing.

As an example of how the final set of terms is evaluated and studied, the term “εὐνομία” / “favor”, found in one of the Ancient Greek texts, is coupled with the term “νόμος” / “law” as its hypernym, and then labeled as a member of “ἦθος” / “ethos” semantic class. Following the same process, the term “αὐτάρκεια” / “self-sufficiency” is coupled with its hypernym “ἐπάρκεια” / “sufficiency” and labeled as a member of “ποσότητα” / “quantity” semantic class. Processing terms this way, gradually builds the semantic context for a growing num-

¹ Photodentro Greek National Aggregator of Educational Content (<http://photodentro.edu.gr/>).

² AntConc Text Analysis Software (<https://www.laurenceanthony.net/software/antconc/>).

ber of concepts, grouping together relevant terms and phrases as part of their semantic grid. Following such relations of hyponymy and hypernymy allows the above process to step up from the semantic level of the initial terms and focus on concepts that consistently participate as thematic pillars of the Ancient Greek texts.

The semantic analysis of terms, for an Ancient Greek to Modern Greek bilingual thesauri, requires the cross referencing and extensive studying of relevant thesauri and dictionaries. Currently, there is a sufficient number of curated Ancient to Modern Greek dictionaries, able to facilitate the accurate translation and interpretation of terms. On the other hand, only a small number of Ancient Greek thesauri are available, all quite old and most having received few to no updates. As a result, conducting a semantic analysis for a large volume of terms, for an Ancient to Modern Greek bilingual thesauri, is an ambitious feat that requires time, dedication and making the most out of the above resources. This work's goal is to study the extracted set of unique terms and produce a piloting set of concepts that will be consistently expanded and maintained. These are concepts that hold significant semantic and educational value, supporting the teaching of the respective Ancient Greek texts.

4.2 Macrostructure

Defining the semantic structure of the AMGthes was a result of studying other well established bilingual thesauri. The purpose of this task was to understand and adopt semantic relations from both old and modern lexical resources with similar goals and targeted functions. More specifically, the classification scheme of three thesauri was studied in detail: i) the Roget's Thesaurus, ii) Antilexicon by Vostantzoglou, and iii) the Cambridge French – English Thesaurus by Marie Noelle Lamy (1997). The decided macrostructure is based on a taxonomy that is essentially a combination of the above, primarily adopting the semantic classes of Vostantzoglou. The reason for this alignment is that Vostantzoglou is an older lexical resource that analyzes terms which are conceptually more relevant to the semantic context of the Ancient Greek texts. In contrast to more modern thesauri, Vostantzoglou's vocabulary is more compatible with Ancient Greek concepts and as such his taxonomy offers classification with improved semantic accuracy.

In terms of typological features, AMGthes can be categorized as an online, bilingual, special, conceptual dictionary. This classification denotes that the provided information is organized, based on semantic criteria, and helps users find language signs for specific concepts (Atkins/Rundell 2008 p. 4). Aligned with Antilexicon, AMGthes too functions as a hierarchical hypernymic taxonomy of concepts that sorts its terms based on their semantic class. At its top level, the semantic taxonomy of AMGthes is divided into General and Special concepts. General concepts are then subdivided in Abstract and Specific, while Special into Tangible and Intangible. Each subclass is further analyzed as follows (Fig. 1):

- [General][Abstract]: Existence, Relation, Unity, Order, Quantity, Number, Time
- [General][Specific]: Space, Dimension, Shape, Energy, Motion
- [Special][Tangible]: Inorganic Matter, Organic Matter
- [Special][Intangible]: Mind, Will, Action, Values, Emotion, Ethos, Religion

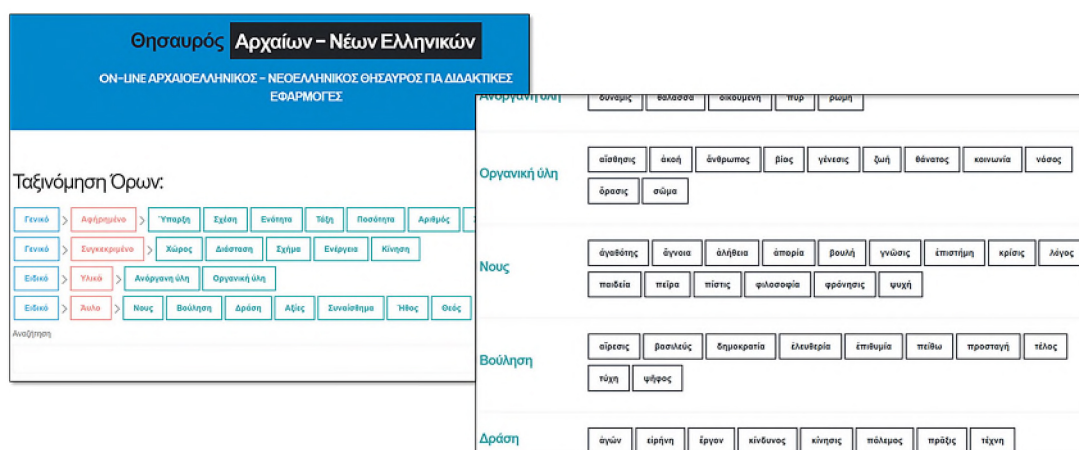


Fig. 1: Macrostructure Semantic Classes – Term Classification Catalog

Being an online dynamic resource, AMGthes can also render an alphabetically ordered list of its terms. Its users can navigate and view the terms included in each of the above semantic classes. Furthermore, both of its currently available catalogs include a search field, allowing the dynamic filtering of their content (Fig. 1). It is worth mentioning that AMGthes data representation and data persistence (the way information is stored and handled) is future proofing its ability to expand and use new semantic classes. The above taxonomy and macrostructure are stored and structured in a human-readable data format (JSON), and as such they can be easily updated and extended to support new semantics and mechanisms for the optimal organization of the constantly enriched AMGthes vocabulary.

4.3 Microstructure

As mentioned before each term is assigned to a semantic class defined as part of the macrostructure taxonomy. Combining information from the decided reference lexical resources, each concept is extensively studied and researched, to identify and map the semantic relations of synonymy, antonymy, hypernymy and frozen expressions. These are the core components for building the semantic profile of each term and thus creating the grid of semantic relations of AMGthes.

The AMGthes microstructure is organized into modules, each profiling a general concept (a thesaurus lemma) by providing synonyms and related words in Ancient and Modern Greek. The large majority of concepts are nouns. The relation of antonymy is also addressed and studied as a separate concept. Antonym concepts are coupled together and presented side by side as linked modules. Inheriting its formalizations from the three reference thesauri, AMGthes renders its modules as online content, using the following template and formatting rules (Fig. 2):

- As a title at the top of each module the reference number (code) and the lemma (the concept) are displayed in bold text. Next to them, the semantic class of the concept is displayed between brackets. Every module is displayed in one column of the page. In case an antonym concept has been studied and profiled, its module will be displayed at the opposing side as a second column.
- A module contains different sections for each lexical category of available synonyms and related words. This means that all the terms (Ancient or Modern Greek) inside a section

share the same lexical category. Each section starts with an abbreviation of the lexical category: “Ουσ.” for nouns, “Ρ.” for verbs, “Επιθ.” for adjectives and “Επιρ.” for pronouns. Each section lists every Ancient Greek synonym or related word of the main concept in a separate line. The first synonym or related word is displayed right next to the section’s abbreviation for its lexical category. A module may also feature an extra section for phrases, using the abbreviation “Φρ.”. Each phrase is formatted in bold text, followed by an optional reference source and relevant Modern Greek phrases.

- Every Ancient Greek synonym or related word is formatted in bold text. It is followed by an optional definition in Modern Greek between square brackets. This optional definition is followed by a set of Modern Greek synonyms or related words, separated by commas. This part of each section functions as a bridge between Ancient and Modern Greek, providing the semantic links that enable AMGthes to serve as a bilingual thesaurus.

To instantiate the above microstructure for every selected concept and produce its content, both online and printed versions of widely approved dictionaries were studied. This extended task represents the data processing and data analysis phase of AMGthes. The digital version of Lidell Scott dictionary³ served as the primary resource for Ancient Greek, along with the Stamatakou dictionary.⁴ For Modern Greek, the Antilexikon by Vostantzoglou provided the main resource of material and the overall point of reference. The dictionaries by Dimitrakou⁵ and Triantafyllidis,⁶ along with Babiniotis⁷ dictionary of synonyms and antonyms, also supported the data processing in Modern Greek. Proper fusion of semantics and definitions from all the above resources was critical for producing balanced modules, suitable for learning purposes and consumption by students.

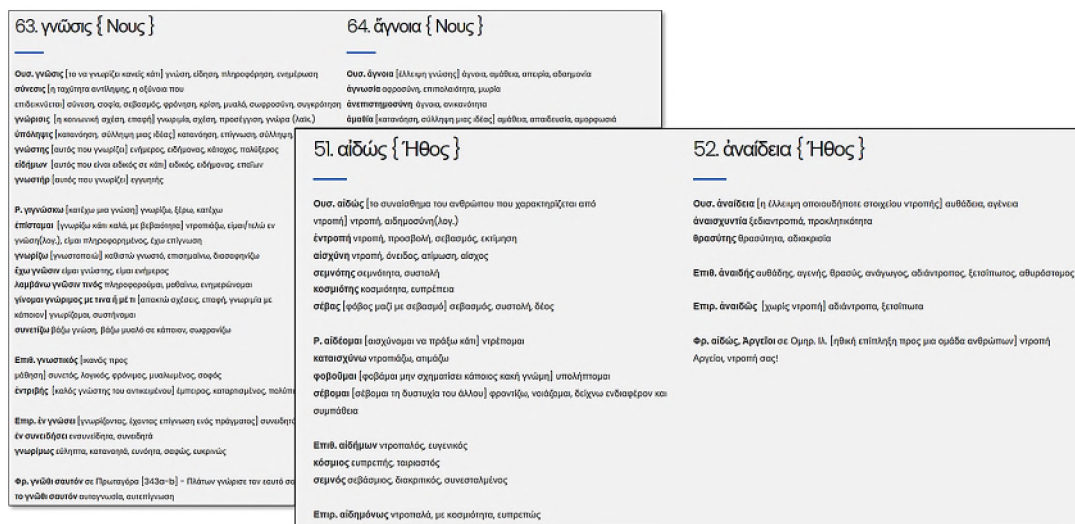


Fig. 2: Microstructure for Concepts “αἰδώς”, “γνώσις” and their antonyms

3 Liddell Scott online dictionary (https://www.greek-language.gr/digitalResources/ancient_greek/tools/liddell-scott/).

4 Stamatakos, I. (1972): Dictionary of the Ancient Greek language. Athens.

5 Dimitrakos D. (1953): Great lexicon of the Greek language. Domi Publications.

6 Triantafyllidis online dictionary (https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/).

7 Babiniotis, G. (2011): Dictionary of synonyms and antonyms of Modern Greek. Lexicology Center.

5. Implementation of the online thesaurus

Following the description of the thesaurus' core semantic components and the reference term's analysis, this section offers details about the online features and web-based implementation. The online thesaurus is being developed as a dynamic website that utilizes modern technologies for its front-end interfaces, back-end services and data instantiation methods.

More specifically, the online thesaurus offers two views, for the presentation of its terms: i) a sorted list of terms in alphabetical order and ii) a semantic taxonomy of terms based on the decided macrostructure (Vostantzoglou classification scheme). Both interfaces were developed with HTML, CSS and JavaScript technologies, allowing for page designs that are aesthetically appealing and easy to navigate. Each listing provides internal links for quick access into the alphabetical order and the taxonomy classes, along with a search bar that facilitates the filtering of terms.

In terms of back-end implementation, instead of offering a simple static web page, the proposed thesaurus functions as a dynamic resource. Its views/interfaces are powered by web services, developed in JavaScript using the Node.JS platform. These services access and process the files, where the structured data for each term's semantic profile is stored. To properly capture the information for a populated microstructure, our data instantiation employs the widely used JSON format, currently driving most internet data and metadata. These data modeling decisions provide the means for the versatile referencing inside each term's profile, giving direct access to specific semantic relationships and their translations.

The online thesaurus is built with modern technologies that ensure its responsiveness and fast rendering. Both its mechanisms (services) and information architecture, offer the means for future extensions and enriched content. The design theme and navigation mechanisms, ensure efficient browsing and an overall pleasant experience while exploring this online lexical resource.

6. Conclusions

Discussing the design decisions for the AMGthes, this paper highlights its development in terms of content and semantic features. The role of a lexicographic process is studied, describing the phases and compilation tasks of the adopted workflow. Starting with the data acquisition of unique terms, from Ancient Greek texts of the Greek school curriculum, our work gradually prepares, processes, and models the data and semantics of AMGthes. Its macrostructure and microstructure are explained, profiling their alignment with methods and techniques from the employed reference dictionaries and thesauri. Finally, the paper provides information regarding the online implementation and web access pattern. The role of modern technologies is described, enabling the development of a dynamic web portal with data handling mechanisms. By design, its online content will support scheduled updates, improvements and optimizations, offering a consistently enriched resource for students and academics.

AMGthes is built using a pipeline of compilation tasks that effectively fuse knowledge to produce a much-needed lexical resource. Creating and maintaining AMGthes is a challenge and a requirement for bringing the teaching of Ancient Greek language up to speed with modern learning technologies and teaching strategies. The presented work delivers the pilot version of a bilingual thesaurus that trains the students' ability to traverse semantic rela-

tionships and empowers them to explore, translate and accurately interpret the context of Ancient Greek texts. Upon the completion of AMGthes development and lexical coverage expansion, it will be hosted at the webpage of +MorPhoSe lab of Democritus University of Thrace, free to be accessed by students and any prospect visitors.

References

- Atkins, S. B. T. (ed.) (1998): *Using dictionaries. Studies of dictionary use by language. Learners and translators.* Tübingen.
- Atkins, B. S./Rundell, M. (2008): *The Oxford guide to practical lexicography.* Oxford.
- Babiniotis, G. (2011): *Dictionary of synonyms and antonyms of Modern Greek.* Lexicology Center.
- Berti, M./Bizzoni, Y./Boschetti, F./Crane, G./Del Gratta, R./Yousef, T. (2016): Ancient Greek WordNet meets the dynamic lexicon: the example of the fragments of the Greek historians. In: *Proceedings of the 8th Global WordNet Conference (GWC)*, pp. 34–38
- Bizzoni, Y./Boschetti, F./Diakoff, H./Del Gratta, R./Monachini, M./Crane, G. (2014): The making of ancient greek wordnet. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1140–1147.
- Dimitrakos, D. (1953): *Great lexicon of the Greek language.* Domi Publications.
- Dziemianko, A. (2012a): On the use (fulness) of paper and electronic dictionaries'. In: Granger, S./Paquot, M. (eds.): *Electronic lexicography.* Oxford.
- Gavriilidou, Z./Konstantinidou, E. (2021): The design of an explicit and integrated intervention program for pupils aged 10–12 with the aim to promote dictionary culture and strategies. In: Gavriilidou, Z./Mitis, L./Kiosses, S. (eds.): *XIX Euralex Proceedings. Vol. 2*, pp. 735–745.
- Gouws, R. H./Prinsloo, D. J. (2010): *Principles and practice of South African lexicography.* Stellenbosch.
- Hüllen, W. (2004): *A history of Roget's thesaurus: origins, development, and design.* Oxford.
- Jarmasz, M. (2012): *Roget's thesaurus as a lexical resource for natural language processing.* arXiv preprint. arXiv:1204.0140.
- Klosa, A. (2013): The lexicographical process (with special focus on online dictionaries). In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie. Supplement Volume: Recent developments with focus on electronic and computational lexicography.* (= *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.4*). Berlin/Boston, pp. 517–524.
- Lamy, M. N. (1997): *The Cambridge French-English thesaurus.* Cambridge.
- Liddell Scott Online Dictionary.
https://www.greek-language.gr/digitalResources/ancient_greek/tools/liddell-scott/.
- Miller, G. A./Fellbaum, C. (2007): WordNet then and now. *Language resources and evaluation.* In: *Language Resources and Evaluation 41 (2)*, pp. 209–214.
- Roget, P. M. (1911): *Roget's thesaurus of English words and phrases.* New York.
- Stamatakos, I. (1972): *Dictionary of the Ancient Greek language.* Athens.
- Trapalis, G./Markantonatou, S./Alexopoulou, A./Fotopoulou, A./Maistros, Y. (2005): The Antilexicon by Vostantzoglou and the development of a small scale semantically organized lexical database of Modern Greek. In: *7th International Conference on Greek Linguistics, York, UK, September 8–10.*

Triantafyllidis Online Dictionary.

https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/.

Vostantzoglou, Th. (1962): Antilexikon or onomasticon of Modern Greek language. Athens.

Wiegand, H. E. (1998): Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. Berlin/New York.

Contact information

Kyriaki Salveridou

Democritus University of Thrace

ksalveri@helit.duth.gr

Zoe Gavriilidou

Democritus University of Thrace

zgabriil@helit.duth.gr