NLP for the Greek Language: A Brief Survey

Katerina Papantoniou and Yannis Tzitzikas
kpapantoniou@csd.uoc.gr,tzitzik@ics.forth.gr
Institute of Computer Science, FORTH-ICS, and Computer Science Department, University of Crete
Heraklion, Crete, Greece

ABSTRACT

There is a plethora of methods, tools and resources for processing text in the English language, however this is not the case for other languages, like Greek. Due to the increasing interest in NLP, and since there is a noteworthy number of works related to the processing of the Greek language, in this paper we survey the work related to the processing of Greek language. In particular, we list and briefly discuss related works, resources and tools, categorized according to various processing layers and contexts. This survey can be useful for researchers and students interested in NLP tasks, Information Retrieval and Knowledge Management for the Greek language.

KEYWORDS

natural language processing, Greek language, Survey

ACM Reference Format:
Katerina Papantoniou and Yannis Tzitzikas. 2020. NLP for the Greek Language: A Brief Survey. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3411408.3411410

1 INTRODUCTION

There is a wide variety of methods, tools and resources for processing text in the English language. However this is not the case for the Greek language even though it has the longest documented history of any living Indo-European language, spanning at least 3,500 years of written records, 28 centuries (Archaic period - new) of written text with alphabet[1].

To aid those that are interested in developing or advancing the techniques for Greek processing, in this paper we survey related work. To this end we list related works and resources organized in categories. We hope this collection and categorization of works to be useful for students and researchers interested in NLP tasks, Information Retrieval and Knowledge Management for the Greek language.

[1] https://www.britannica.com/topic/Greek-language

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

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411410

The paper is organized as follows. Section 2 describes previous surveys on Greek NLP, and the methodology that we followed. Section 3 surveys the works on Greek NLP. Section 4 describes some noteworthy resources and evaluation collections that are available, and finally, Section 5 concludes the paper by providing some general observations,

2 RELATED WORK (OTHER SURVEYS) AND METHODOLOGY

There are not many papers that attempt to survey this area. We have found the 2010 paper [93], the white paper [28], as well as surveys about particular topics like [92] that focuses on tools for sentiment analysis in Greek.

2.1 Methodology for the Current Survey

For finding the related works we used Google Scholar in the period: April 2020 - May 2020, without any restriction on the publication date. We restricted to those papers that contain the word "Greek" in the title and we used various extra keywords for finding the related papers. We did not focus on papers that deal with multilingualism (and could potentially contain Greek as well). Consequently, the survey is by no means complete, however we tried to find the more relevant and important works and resources for giving a concise overview. Moreover we have to note that the reader must be aware and keep an eye also on transfer learning approaches since it can be of great benefit on NLP tasks in low-resource languages. In the sections that follow when we refer to Ancient Greek we mean any premodern form of Greek (as specified also in ISO 639-3 code set[2]). The usual cutoff for this distinction is the date of 1453 (the Fall of Constantinople).

3 SURVEY OF WORKS ON GREEK NLP

Below we attempt to group the available works according to various processing layers and contexts, namely: object character recognition (§3.1), morphology (§3.2), phonetics (§3.3), syntax (§3.4), embeddings (§3.5), semantics (§3.6), pragmatics (3.7), sentiment analysis (§3.8), question answering (§3.9), natural language generation (§3.10), machine translation (§3.11). An overview of the topics that are addressed (up to some degree) is given in Figure 1. However we also include a section that refers to works that capture several layers, as well as sections on related topics (§3.12). Resources and evaluation collections are described afterwards, in §4.

[2] https://iso639-3.sil.org/sites/iso639-3/files/downloads/iso-639-3_Name_Index_Latin1.tab

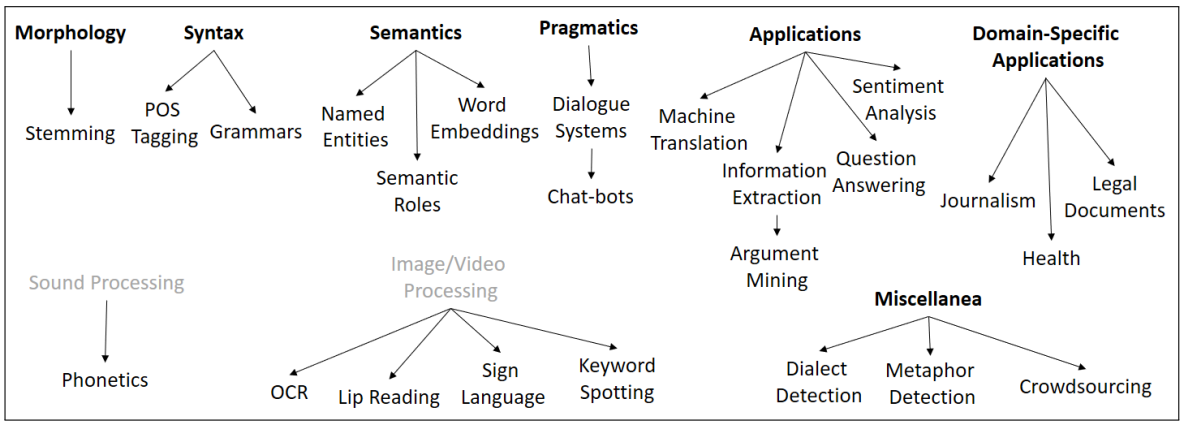

Figure 1: The categorization of NLP tasks that we use in this survey

We are aware that a work may fit to more than one category, however we have decided to avoid duplicates. We have also decided to avoid the inclusion of quantitative results because such numbers would not be comparable since in most cases they are based on experiments over different collections (the interested user can refer to the mentioned papers for quantitative results). The works in each section are described in chronological order.

To aid understanding, Figure 2 shows a small example, i.e. a small Greek phrase, specifically 'Ο Αριστοτέλης είδε την κοπέλα με το τηλεσκόπιο' (Aristotle saw the girl with the telescope), and how the results of some basic NLP tasks might be. However we have to be aware, that the general objective is NLU (Natural Language Understanding), which is broad and challenging issue, due to the various ambiguities. For instance, the slightly modified phrase of Figure 2, 'Ο Αριστοτέλης είδε την κοπέλα με το κινητό' (Aristotle saw the girl with the mobile), could have several meanings, e.g. "Aristotle" could refer either to the ancient philosopher, or to Aristotle Onassis (Greek shipping magnate), or to a modern person, while the "mobile" could refer to a mobile object or to a cell phone, finally the phrase "with the mobile" may refer to "saw" (e.g. Aristotle saw the girl *through his* mobile) or to "girl" (e.g. Aristotle saw the girl *carrying her* mobile). We can understand the intended meaning(s) only by considering (and properly analyzing) the context of this phrase.

3.1 Optical Character Recognition (OCR)

Optical Character Recognition (OCR) refers to the process of converting images of typed, handwritten or printed text into machine-readable text.

3.1.1 Ancient Greek. [60] focuses on the problem of recognizing old Greek manuscripts and proposes a recognition technique tested in a large number of historical manuscript collections which are written in lowercase letters and originate from St. Catherine's Mount Sinai Monastery.

[101] describes a process of training the open source Tesseract OCR engine[3] to support Ancient Greek. One of the challenges in this work is the recognition of the two types of diacritical marks of Ancient Greek breathing marks and accents.

[77] focuses on large-scale optical character recognition of ancient (or polytonic) Greek and it reports results from processing 1,200 volumes comprising 329,002,271 Greek words.

3.1.2 Modern Greek. [99] and [98] describe methods for Greek handwritten character recognition evaluated over handwritten character and digit databases and the emphasis is given to the use of novel features extraction techniques.

[84] presents an approach for the recognition of Greek polytonic scripts by training and testing a Long Short-Term Memory (LSTM)-based recognizer using the OCRopus framework[4]. In this context the publicly available Polyton-DB [27] has beed used that is a database containing 15,689 textlines of synthetic and authentic printed Greek Polytonic script.

3.2 Morphology

Morphological parsing is the analysis of word structure and it is usually a prerequisite to many kinds of computational processing of text. Morphology is a rich component of the Greek language since both Ancient, Modern and Greek dialects have a relatively rich inflectional morphology especially in the noun and verb systems. To this end, some of the first efforts in processing Greek text were towards the creation of tools for this level of linguistic analysis.

3.2.1 Ancient Greek. [10] describes GLEM, a publicly available lemmatizer[5] for Ancient Greek that uses part-of-speech information for disambiguation that also assigns output to unseen words, while [5][6] focuses on the recovery of missing characters from a damaged text input using deep neural networks.

[3] https://github.com/tesseract-ocr/tesseract
[4] https://github.com/tmbarchive/ocropy
[5] https://github.com/GreekPerspective/glem
[6] https://github.com/sommerschield/ancient-text-restoration

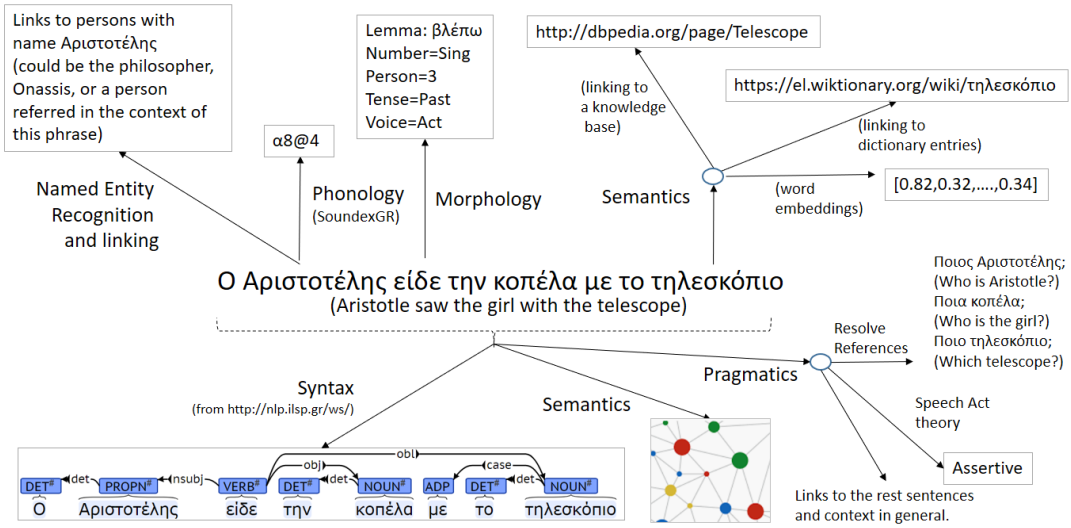

Figure 2: Example of some basic NLP tasks over a Greek phrase

3.2.2 Modern Greek. [83] presents the implementation of a morphological processor for the Modern Greek language that is based on the two-level morphology model introduced by Koskenniemi [46].

[71] describes a large-scale Greek morphological lexicon, specifically its architecture, the procedure followed to develop it, as well as the provided functionalities to update it. The described lexicon was used to develop a lemmatiser and a morphological analyser.

[59][7] describes a stemmer for the Greek language that follows the known Porter algorithm for the English language and it is developed according to the grammatical rules of the Greek language. Along the same line, [61] is a stemmer for the Greek language build in 2008 that follows again an affix removal approach (as the Porter's Algorithm) adapted for the Greek language.

[82] introduces another stemmer[8] for the Greek language and compares it with previous implementations, while [1] provides a stemming and tagging procedure for the Greek language that can be easily adapted to existing systems.

3.3 Phonetics

Phonetics is a branch of linguistics that studies the sounds of human speech. Phonetic algorithms fall in this category.

[20] presents an algorithm for the generation of intonation (F0 contours) for the Greek Text-To-Speech system developed at Wire Communications Laboratory[9].

[24] analyzes the acoustic characteristics of the Greek vowels (duration, fundamental frequency, amplitude, and others).

[39] describes a machine learning (ML) approach that learns rules of how to transcribe Greek words into the International Phonetics Association's (IPA's) phonetic alphabet. It mines rules of phonologic transcription from training data (phonologically transcribed Greek words) that are available on the internet and in other resources.

[89] describes classification approaches based on deep neural networks for distinguishing two Greek dialects, namely Athenian Greek, the prototypical form of Standard Modern Greek and Cypriot Greek.

Finally, [42] is an adaptation and extension of the phonetic algorithm Soundex for the needs of the Greek language.

3.4 Syntax

Syntactic analysis aims at revealing the syntactic role of the words.

3.4.1 Ancient Greek. [9] describes the first release of the Ancient Greek Dependency Treebank[10] (AGDT), a 190,903-word syntactically annotated corpus of literary texts including the works of Hesiod, Homer and Aeschylus.

[14] reports the results from five part-of-speech (POS) taggers tested on data of the Ancient Greek Dependency Treebank.

[31] describes a dataset comprising a collection of dependency syntax trees of representative texts from ancient Greek prose authors (Aeschines, Antiphon, Appian, Athenaeus, Demosthenes, Dionysius of Halicarnassus, Herodotus, Josephus, Lysias, Plutarch, Polybius, Thucydides, and Xenophon), totaling to date 550,000+ tokens.

[43] presents a "roadmap" on how different already available treebanks in Ancient Greek, namely Perseus Ancient Greek Dependency Treebanks, Aphthonius, Pedalion, Gorman, Sematia and PROIEL, can be combined to enhance the progress

[7] An implementation is available at https://github.com/skroutz/greek_stemmer.
[8] http://saroukos.com/stemmer
[9] http://www.wcl.ece.upatras.gr/en/ai/resources/demo-greek-text-to-speech
[10] https://perseusdl.github.io/treebank_data/

in the automated parsing of classical and post-classical Greek texts.

3.4.2 Modern Greek. [72] investigated the use of Transformation Based Error-Driven learning for resolving part-of-speech ambiguity in the Greek language.

[62] describes a machine learning approach for POS disambiguation and unknown word guessing. Both problems are cast as classification tasks carried out by decision trees.

[65] proposes a unified POS tagging architecture that uses a feature-based multi-tiered approach to cover the requirements of tagging highly inflective languages such as Greek.

[74] describes the construction of the Greek Dependency Treebank while [75] describes the work towards the harmonization of the Greek Dependency Treebank[11] with the Universal Dependencies v2 standard, and the extension of the treebank with enhanced dependencies.

[25] presents a tool for analyzing morphologically and syntactically sentences, phrases, and words in order to correct syntactic, grammatical, and stylistic errors.

[68] describes a ML approach to POS tagging and named entity recognition for Greek, focusing on the extraction of morphological features and classification of tokens into a small set of classes for named entities.

[48] presents a method for morphological disambiguation in the absence of any gold standard corpora using several statistical models, such as the Brill algorithm [12]. The methods were tested both on Greek and Yiddish.

3.4.3 Endangered Dialects. [2] evaluates POS tagging techniques on an actual endangered language, Griko (a dialect spoken in an area of Southern Italy called Grecìa Salentina). The combination of a semi-supervised method with cross-lingual transfer seems to be the more appropriate for this task given the small size of the used corpus.

3.5 Word Embeddings

A Word Embedding method generally tries to map words, phrases or documents, to vectors of real numbers so that semantic, syntactic and world knowledge can be captured and reflected in similar vectors.

[64] presents word embeddings[12] and other linguistic resources trained on the largest to date (2018) digital Greek language corpus.

[50] presents an ensemble method, Continuous Bag-of-Skip-grams (CBOS), for producing word representations for the Greek language.

[15][13] presents a Greek edition of the BERT [18] pre-trained on large Greek corpora.

[63] describes a Greek version of WordSim353 test collection [22] for a basic evaluation of word similarities, and the testing of seven word vector models.

[11] http://gdt.ilsp.gr and https://universaldependencies.org/treebanks/el_gdt/index.html.

[12] http://archive.aueb.gr:7000/

[13] https://github.com/nlpaueb/greek-bert

3.6 Semantics

The semantic processing usually considers a representation of sentence structure (syntactic tree), a representation of the possible meaning(s) of each word, and aims at producing a representation of the meaning of the sentence. This process includes steps like word sense disambiguation, semantic role tagging, named entity recognition and linking, and other tasks.

3.6.1 Ancient Greek. [13] presents an annotation scheme for semantic role annotation for the Ancient Greek Dependency Treebank.

[70] describes a dynamic semantic change model (Bayesian model) to study the evolution of word senses in ancient text that leverages categorical metadata about the texts' genre to boost inference and uncover the evolution of meanings in Ancient Greek corpora.

[78] describes a vector space model where every word is represented by a vector which encodes information about its linguistic context(s).

3.6.2 Modern Greek. [40] describes a prototype NER (Named Entity Recognition) system for Greek texts developed based on a NER system for English. Both systems are evaluated on corpora of the same domain and of similar size.

[17] describes work in progress for a named entity recognizer for Greek based on pattern matching and non-recursive regular expressions.

[51] describes a freely available named-entity recognizer[14] for Greek texts that identifies temporal expressions, person, and organization names. An ensemble of support vector machines (SVM) and active learning was used for person and organization names while for the temporal expression semi-automatically produced patterns.

[4] describes a method for named entity recognition in Greek legislation using deep neural network architectures. The recognized entities are used to enrich the Greek legislation knowledge graph with more detailed information about persons, organizations, geopolitical entities, legislation references, geographical landmarks and public document references.

3.7 Pragmatics (Dialogue Systems)

Pragmatic analysis attempts to put each sentence into its general situational context, taking into account the contexts in which it is said. This process includes various (and challenging) tasks including: resolution of references (e.g. pronouns), resolution of semantic ambiguities, inferring the missing objects and actions, drawing inferences based on world knowledge and common scenarios and scripts, linking sentences, identifying speech acts, understanding discussions and dialogues.

3.7.1 Ancient Greek. [79] uses linguistic dependency treebanks and digitized texts created by the Perseus Digital Library, to create social networks for a collection of Greek

[14] http://nlp.cs.aueb.gr/software.html

tragedies that allow users to visualize the interactions between characters in the plays.

3.7.2 Modern Greek. The thesis [38] describes a conversational chatbot system based on public services for a greek web portal called diadikasies.gr to help citizens find easily the desired service.

3.8 Sentiment Analysis

Sentiment analysis is the process of computationally identifying opinions, emotions and attitudes expressed in a piece of text towards a particular topic, product, etc.

[55] implements a sentiment classifier applying SVM on hotel reviews written in Modern Greek. The corpus of hotel reviews was collected from the Greek version of TripAdvisor and consists of 1,800 reviews (900 positive and 900 negative). A manual annotated sentiment lexicon also created in the context of this work that comprises a total of 27,388 types of positive words and 41,410 types of negative words.

[7] analyzes tweets related to schizophrenia in Greek language, and examines schizophrenia tweets in comparison with other illness (diabetes).

[69] describes methodologies for both fully automatic and semi-automatic moderation in user comments. They propose the use of a gating mechanism (GRU) in recurrent neural networks (RNN) on word embeddings. The results has been evaluated both on English and Greek datasets.

[86] presents an approach to recognize opinions in Greek language and examines the impact of feature selection on the analysis of opinions and the performance of the classifiers (text-based and part-of-speech based features from textual data are extracted).

[6] focuses on sentiment analysis applied in Greek texts, specifically it investigates a process where gradient boosting, a technique for dealing with high-dimensional data, is applied on Greek texts.

[21] examines the hashtag #greferendum, focusing on both social and semantic networks (#Grexit, #oxi campaign), and it analyses Twitter data, which were collected using NodeXL, on three significant days: the announcement of the referendum, the day of the bailout expiration and the actual date of the referendum.

[91] focuses on the 2015 Greek bailout referendum, aiming to nowcast on a daily basis the voting intention of 2,197 Twitter users. It proposes a semi-supervised multiple convolution kernel learning approach, leveraging temporally sensitive text and network information.

[11] explores whether and how the Greek media and social media discourses on the refugee crisis and the political decisions regarding it contribute to the discursive reconstruction of the Greek nation-state.

[92] presents a survey for sentiment analysis tools and resources in Greek, and make publicly available a rich set of such resources[15], ranging from a manually annotated lexicon, to semi-supervised word embedding vectors and annotated datasets for different tasks.

[87] presents an approach to analyze Greek text and extract indicative info towards users' opinions and attitudes. It describes a supervised approach adopted that analyzes and classifies comments and reviews into the appropriate polarity category and discretization techniques are also applied to improve the performance and the accuracy of classification procedures.

[44] describes the creation of a corpus of over 1,300 real use tweets and it analyzes these exchanges based on the analytical tools of the Lexical Constructional Model [76] (a framework that integrates the pragmatics and discourse dimensions of language use).

[73] presents the first Greek annotated dataset for offensive language identification: the Offensive Greek Tweet Dataset (OGTD). OGTD is a manually annotated dataset containing 4,779 posts from Twitter annotated as offensive and not offensive. They have also trained a number of systems on this dataset and report that the best results have been obtained from a system using LSTMs and GRU with attention.

3.9 Question Answering (QA)

QA refers to the process of retrieving precise answers to questions expressed in natural language; it is a long standing goal, first introduced in late 60s and early 70s.

[54] presents APANTISIS, a modular QA system implemented for the Greek language for plugging it to databases or knowledge bases.

3.10 Natural Language Generation (NLG)

Natural Language Generation refers to a process that takes as input data and text, and generates text in some human language.

[19] describes a computational grammar of Greek for generating automatically descriptions of museum exhibits from a database.

[52] presents an algorithm based on grammatical rules and semantic information dedicated for Greek language for the extraction of document content summary. The proposed method has been tested on the news articles domain.

[3] presents the NaturalOWL an open-source natural language generation system that produces texts describing individuals or classes of OWL ontologies (in both English and Greek).

3.11 Machine Translation (MT)

MT aims at translating a text from a source language to its counterpart in a target language.

[8] addresses the problem of translating from morphologically poor (English is this case) to morphologically rich languages (Greek and Czech) by adding per-word linguistic information to the source language.

[96] presents an annotation tool that allows the manual annotation of certain linguistic items in the source text and

[15] https://github.com/MKLab-ITI/greek-sentiment-lexicon

their translation equivalent in the target text, by entering useful information about these items based on their context.

[58] applies constraint logic programming techniques to deal with polysemy issue in machine translation.

3.12 Miscelanea

3.12.1 Dialect Detection. The thesis [80] describes the collection of a bidialectal corpus of Greek[16] (Cypriot Greek and Modern Greek) and the construction of a classifier based on character and word n-gram features to distinguish between the dialects.

3.12.2 Lip Reading. [41] focuses on the problem of Lip Reading with Greek words in an unconstrained driving scenario by using only visual information. It provides a dataset with image sequences from Greek (10 persons spoke 50 words while they were either driving or simply sitting in the passenger's seat of a car) and a recognition pipeline that consists of a Convolutional Neural Network followed by a Long-Short Term Memory Network with a plain attention mechanism.

3.12.3 Greek Sign Language. [85] presents a method for sign language recognition using a Leap Motion controller and apply it to the recognition of the Greek sign language (GSL) alphabet. The method utilizes 3D positional data provided by the device along with SVM classification.

[30] describes a system that can recognize GSL vocabulary in translation mode using Kinect technology (the sensor captures 3D hands movement trajectory and then a set of features in the form of body joints are fed to a classifier to recognize the input sign).

[47] proposes a Rule Based Machine Translation (RBMT) system for the creation of large and quality written GSL glossed corpora from Greek text. The proposed RBMT system assists the professional GSL translator in speeding up the production of different kinds of GSL glossed corpora.

3.12.4 Keyword Spotting. [45] focuses on keyword spotting, a methodology for document indexing based on spotting words directly on images without the use of a character recognition system, and describes all steps of this process (preprocessing for image binarization, enhancement and segmentation, feature representation, matching and word retrieval).

3.12.5 Argument Mining. [32] presents a two-step classification approach for argument extraction from social media texts. During the first step, argumentative sentences are detected while in the second step, the premises in the argumentative sentences are identified. An analogous procedure is presented in [81] to identify boundaries for claims and premises in argumentative text segments with the addition of word embeddings.

3.12.6 Computational Journalism. [66] presents a platform for automated data processing in the context of Computational Journalism along with a general methodology for event extraction from different data sources (news articles and tweets). Another platform that can support journalists and advertisers is the Social Web Observatory [94] that through a pipeline offers text collection/crawling, entity recognition, clustering of texts into events related to entities, entity-centric sentiment analysis, and also text analytics and visualization functionalities.

3.12.7 Health Domain. [97] presents a suite of linguistic resources and tools for use in high level NLP applications in the domain of biomedicine. In this context a Greek morphological lexicon of about 100,000 words, a lemmatiser and a morphosyntactic tagger, a specialized corpus of biomedical texts and an ontology of medical terminology has been developed.

[100] describes GREECAD corpus that is annotated Greek corpus of aphasic spoken discourse.

[90] uses a set of language dependent and language independent linguistic features to represent the poems of 13 Greek poets of the 20th century, and applies multiple machine learning algorithms to predict a writer's likelihood of committing suicide.

3.12.8 NLP and Crowd-sourcing. [88] describes data collection tasks for parallel translation implemented using a crowdsourcing platform.

3.12.9 Metaphor Detection. [23] focuses on the automatic differentiation between literal and metaphorical meaning in authentic non-annotated phrases from the Corpus of Greek Texts[17] [33] by means of computational methods of machine learning.

3.12.10 Irony Detection. [16] describes a classification schema for irony detection in Greek political tweets relying on limited labeled training data, thus a semi-supervised approach is followed, where collective-learning algorithms take both labeled and unlabeled data into consideration.

3.12.11 Genre Identification. In [29] almost all classical Greek literature was identified as prose or verse with supervised learning based on a stylometric feature set for Ancient Greek language.

3.12.12 Fake News Detection. The thesis [56] examines the methods and tools Greek users implement in order to spot fake news on social media and counter its spread. The data was collected from the members of the Ellinika Hoaxes Facebook group.

3.12.13 Legal Domain. [95] presents a method for the automated extraction of acronyms in legal Greek texts.

[26] presents a project about the automated analysis and processing of Greek legal documents for their transformation into Legal Open Data.

3.12.14 Authorship. In [53] a long lasting dispute about the authorship of Rhesus play, traditionally attributed to Euripides, is investigated. Frequency of words and character

[16] https://github.com/hb20007/greek-dialect-classifier

[17] http://www.sek.edu.gr/index.php?en

n-grams analysis lead the authors to conclude that the author of Rhesus was probably an actor in whose repertoire Euripides and Aeschylus, but not Sophocles, were included and a great admirer of Aeschylean grandeur and style. For Modern Greek [57] employs a supervised learning approach for authorship attribution and uses as features a mixture of n-grams both in terms of size (bigrams and trigrams) and type (word and character n-grams). [36] presents a comparative assessment of the difficulty of authorship attribution in Greek and in English. Based on a corpus introduced in this work different methods of author attribution were applied and the results show that, overall, performance on English is higher than performance on Greek.

4 RESOURCES AND EVALUATION COLLECTIONS

Here we list a few portals containing various NLP resources for the Greek language. We also include collections that are suitable for evaluating various NLP tasks (over Greek resources) since they are very useful for obtaining comparative results and thus advancing the current techniques. Although roughly every paper reports some evaluation results, and thus it contains or uses an evaluation collection, here we list only some of them, therefore we recommend the reader to refer to the particular papers of interest.

Portals:

- `metashare.ilsp.gr`[18] providing access to 99 resources related to the Greek language.
- `qt21.metashare.ilsp.gr/`[19] with 179 resources and tools related to machine translation.
- `clarin:el`[20] providing access to 514 resources.
- NLP explorer[21] [67] for searching, and visualizing the ACL Anthology dataset[22].

Dictionaries, Lexicons and Thesauri:

- Greek Wordnet[23]
- GreekLex2[24] [49]: freely available lexical database with part-of-speech, syllabic, phonological, and stress information.
- Sentiment Lexicon[15] [92]: provides a manually annotated sentiment lexicon (32,885 lemmas) as well as two automatically generated sentiment lexicons.

Various Corpora:

- The Ancient Greek and Latin Dependency Treebank 10
- Greek Dependency Treebank and Greek UD treebank (UD_Greek-GDT)[11] [9].
- Hellenic National Corpus (HNC)[35]: 47 million words which contains texts published from 1976 to 2007. A web interface is available for searching the corpus[25].
- Corpus of Greek Texts (CGT)[17] [33]: 30 million words which contains texts from two decades (1990 to 2010).
- Corpus of Spoken Greek[26]. The motivation (from a linguistic and pedagogical point of view) and its design is described in [34].

Datasets:

- Gazzetta dataset[27] [69]: contains manually moderated Greek user comments from Gazzetta, a Greek sports news portal.
- Sentiment greek tweets dataset[28] [37]: 683 manually rated tweets for their sentiment intensity
- Greek political tweets[29] [16]: contains 61,427 Greek tweets collected on the week before and the week after the May 2012 parliamentary elections in Greece. This dataset is available for research purposes.

Table 1: Tools

Lemmatizer:	GLEM[5] [10][Ancient].
Stemmer:	Mitos[30], Skroutz[31], NCSR[32].
NER:	AUEB[14] [51].
Embeddings:	Word2Vec[12] [64], GreekBERT[13] [15].
NLG:	NaturalOWL[14] [3].
Suite:	ILSP[33].

5 CONCLUDING REMARKS

In total we have found, categorized and described in brief, around 99 works; we hope this collection and categorization to be useful for students and researchers interested in NLP tasks, Information Retrieval and Knowledge Management in general.

A few questions that we have had at the beginning were: (a) how the number of papers about Ancient Greek compares with those about Modern Greek, and (b) what is the distribution of papers per year? Table 2 shows the distribution in decades of the papers mentioned in this survey. We have split the last decade in two rows (each corresponding to 5 years) to make clear the trend. We have observed: (i) a significant overall increase the last years, (ii) an increase in works based

[18] http://metashare.ilsp.gr:8080/repository/search/?q=greek&page=3
[19] http://qt21.metashare.ilsp.gr/repository/search/?q=greek
[20] https://inventory.clarin.gr/ and https://inventory.clarin.gr/resources/search/?q=greek&page=6
[21] http://lingo.iitgn.ac.in:5001/
[22] https://www.aclweb.org/anthology/
[23] https://okfn.gr/983/
[24] https://github.com/CypressA/GreekLex-2 and https://psychology.nottingham.ac.uk/greeklex/
[25] http://corpus.ilsp.gr/index.php?current_page=help
[26] http://ins.web.auth.gr/index.php?lang=en&Itemid=251
[27] https://archive.org/details/gazzetta-comments-dataset.tar
[28] https://www.researchgate.net/publication/301888705_Sentiment_Ratings_for_Greek_Tweets_Dataset/link/572b181908aef7c7e2c68639/download
[29] https://hilab.di.ionio.gr/wp-content/uploads/2020/02/HILab-Greek_POS_Tagged_Tweets.csv
[30] https://github.com/YannisTzitzikas/GreekMitosStemmer
[31] https://github.com/skroutz/elasticsearch-skroutz-greekstemmer
[32] https://github.com/kpech21/Greek-Stemmer
[33] http://nlp.ilsp.gr/ws

on neural networks but also some subfields of NLP such OCR steadily are based on such approaches, and (ii) a relatively high number of papers on sentiment analysis with specialized solutions for the Greek language. We also have to note that we found more works about Ancient Greek than we expected, 14% of the works that we found concern and Ancient Greek.

Our intention is to extend and enrich this survey in the future.

Period	Modern Greek	Ancient Greek	Dialects	Total
[1990-2000)	7	-	-	7
[2000-2010)	15	1	-	16
[2010-2015)	9	2	-	11
[2015-2020)	50	11	4	61
TOTALS	**79**	**15**	**4**	**99**

Table 2: Distribution of papers per year

REFERENCES

[1] Giorgos Adam, Konstantinos Asimakis, Christos Bouras, and Vassilis Poulopoulos. 2010. An efficient mechanism for stemming and tagging: the case of Greek language. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 389–397.

[2] Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2529–2539.

[3] I. Androutsopoulos, G. Lampouras, and D. Galanis. 2013. Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System. *Journal of Artificial Intelligence Research* 48 (Nov 2013), 671–715.

[4] Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation.. In *JURIX*. 1–10.

[5] Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. *arXiv preprint arXiv:1910.06262* (2019).

[6] Vasileios Athanasiou and Manolis Maragoudakis. 2017. A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern Greek. *Algorithms* 10, 1 (2017), 34.

[7] Christina Athanasopoulou and Evanthia Sakellari. 2016. 'Schizophrenia'on Twitter: Content Analysis of Greek Language Tweets.. In *ICIMTH*. 271–274.

[8] Eleftherios Avramidis and Philipp Koehn. 2008. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, 763–770.

[9] David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the eighth international workshop on treebanks and linguistic theories (TLT 8)*. 5–15.

[10] Corien Bary, Peter Berck, and Iris Hendrickx. 2017. A memory-based lemmatizer for Ancient Greek. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. 91–95.

[11] Salomi Boukala and Dimitra Dimitrakopoulou. 2018. Absurdity and the "blame game" within the Schengen area: Analyzing Greek (social) media discourses on the refugee crisis. *Journal of Immigrant & Refugee Studies* 16, 1-2 (2018), 179–197.

[12] Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics* 21 (1995), 543–565.

[13] Giuseppe GA Celano and Gregory Crane. 2015. Semantic role annotation in the ancient greek dependency treebank. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*. 26.

[14] Giuseppe GA Celano, Gregory Crane, and Saeed Majidi. 2016. Part of speech tagging for ancient greek. *Open Linguistics* 1, open-issue (2016).

[15] Ilias Chalkidis. 2020. Greek BERT. https://github.com/nlpaueb/greek-bert

[16] Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence* 51 (2016), 50–57.

[17] Iason Demiros, Sotiris Boutsis, Voula Giouli, Maria Liakata, Harris Papageorgiou, and Stelios Piperidis. 2000. Named Entity Recognition in Greek Texts.. In *LREC*.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[19] Aggeliki Dimitromanolaki, Ion Androutsopoulos, and Vangelis Karkaletsis. 2001. A large-scale systemic functional grammar of Greek. In *5th International Conference on Greek Linguistics, Sorbonne, France*.

[20] George Epitropakis, Nikos Yiourgalis, and George Kokkinakis. 1993. High quality intonation algorithm for the Greek TTS-system. In *ESCA Workshop on Prosody*.

[21] Ioanna Ferra and Athina Karatzogianni. 2017. A tale of cyber-conflict in Greece: Polarization and mobilization for the Greek referendum on Twitter. *Transformations of Protest in Greece. Atenas: Papazisis Publishers & EKKE. Online available at: https://works. bepress. com/athina_karatzogianni/28* (2017).

[22] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.* 20, 1 (2002), 116–131.

[23] Eirini Florou, Konstantinos Perifanos, and Dionysis Goutsos. 2018. Neural embeddings for metaphor detection in a corpus of Greek texts. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 1–4.

[24] Marios Fourakis, Antonis Botinis, and Maria Katsaiti. 1999. Acoustic characteristics of Greek vowels. *Phonetica* 56, 1-2 (1999), 28–43.

[25] Panagiotis Gakis, Christos Panagiotakopoulos, Kyriakos Sgarbas, Christos Tsalidis, and Vassilios Verykios. 2017. Design and construction of the Greek grammar checker. *Digital Scholarship in the Humanities* 32, 3 (2017), 554–576.

[26] John Garofalakis, Konstantinos Plessas, Athanasios Plessas, and Panoraia Spiliopoulou. 2018. A project for the transformation of Greek legal documents into legal open data. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*. 144–149.

[27] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, V. Papavassiliou, F. Sunistira, and V. Katsouros. 2015. GRPOLY-DB: An old Greek polytonic document image database. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 646–650.

[28] Maria Gavrilidou, Maria Koutsombogera, Anastasios Patrikakos, and Stelios Piperidis. 2012. *The Greek language in the digital age*. White Paper. Editors: Georg Rehm and Hans Uszkoreit. http://www.meta-net.eu/whitepapers/e-book/greek.pdf Springer.

[29] Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri, and Joseph Dexter. 2019. Stylometric Classification of Ancient Greek Literary Texts by Genre. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, Minneapolis, USA, 52–60.

[30] Nikolaos Gkigkelos and Christos Goumopoulos. 2017. Greek sign language vocabulary recognition using Kinect. In *Proceedings of the 21st Pan-Hellenic Conference on Informatics*. 1–6.

[31] Vanessa B Gorman. 2020. Dependency Treebanks of Ancient Greek Prose. *Journal of Open Humanities Data* 6, 1 (2020).

[32] Theodosios Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument Extraction from News, Blogs, and Social Media. In *Artificial Intelligence: Methods and*

Applications - 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings (Lecture Notes in Computer Science), Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles (Eds.), Vol. 8445. Springer, 287–299.

[33] Dionysis Goutsos. 2010. The Corpus of Greek Texts: a reference corpus for Modern Greek. *Corpora* 5 (05 2010), 29–44. https://doi.org/10.3366/cor.2010.0002

[34] Dionysis Goutsos, Ourania Hatzidaki, and Philip King. 1994. Towards a corpus of spoken Modern Greek. *Literary and linguistic computing* 9, 3 (1994), 215–223.

[35] Nick Hatzigeorgiu, Maria Gavrilidou, Stelios Piperidis, George Carayannis, Anastasia Papakostopoulou, Athanassia Spiliotopoulou, Anna Vacalopoulou, Penny Labropoulou, Elena Mantzari, Harris Papageorgiou, et al. 2000. Design and Implementation of the Online ILSP Greek Corpus.. In *LREC*.

[36] Patrick Juola, George Mikros, and Sean Vinsick. 2019. A comparative assessment of the difficulty of authorship attribution in Greek and in English. *Journal of the Association for Information Science & Technology* 70, 1 (January 2019), 61–70.

[37] Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, and Avi Arampatzis. 2015. Sentiment analysis of greek tweets and hashtags using a sentiment lexicon. In *Proceedings of the 19th panhellenic conference on informatics*. 63–68.

[38] Ioannis Karamitsos. 2019. Chatbots for Greek/EU Public Services. (2019).

[39] Nikitas N Karanikolas. 2019. Machine learning of phonetic transcription rules for Greek. In *AIP Conference Proceedings*, Vol. 2116. AIP Publishing LLC, 350005.

[40] Vangelis Karkaletsis, Georgios Paliouras, Georgios Petasis, Natasa Manousopoulou, and Constantine D Spyropoulos. 1999. Named-entity recognition from Greek and English texts. *Journal of Intelligent and Robotic Systems* 26, 2 (1999), 123–135.

[41] Dimitris Kastaniotis, Dimitrios Tsourounis, Aristotelis Koureleas, Bojidar Peev, Christos Theoharatos, and Spiros Fotopoulos. 2019. Lip Reading in Greek words at unconstrained driving scenario. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 1–6.

[42] Antrei Kavros and Yannis Tzitzikas. 2020. A Soundex-like Algorithm for the Greek Language. https://github.com/YannisTzitzikas/SoundexGR. *GitHub repository*.

[43] Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, Enriching and Valorizing Treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*. Association for Computational Linguistics, Paris, France, 109–117.

[44] Sophia Kefalidou and Angeliki Athanasiadou. 2019. APO X, Y: A discourse topicalization construction within Greek Twitter. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association* 17, 1 (2019), 187–218.

[45] Anastasios L Kesidis and Basilis Gatos. 2020. Providing Access to Old Greek Documents Using Keyword Spotting Techniques. *Visual Computing for Cultural Heritage* (2020), 85–102.

[46] Kimmo Koskenniemi. 1983. Two-level model for morphological analysis.. In *IJCAI*, Vol. 83. 683–685.

[47] Dimitrios Kouremenos, Klimis Ntalianis, and Stefanos Kollias. 2018. A novel rule based machine translation scheme from Greek to Greek sign language: Production of different types of large corpora and language models evaluation. *Computer Speech & Language* 51 (2018), 110–135.

[48] Mustakimova E. Kuzmenko E. 2015. *Automatic Disambiguation in the Corpora of Modern Greek and Yiddish", journal = "Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015)*. Vol. 1. Russian State University for the Humanities, Chapter 21, 388–398. Issue 14.

[49] Antonios Kyparissiadis, Walter J. B. van Heuven, Nicola J. Pitchford, and Timothy Ledgeway. 2017. GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PLOS ONE* 12, 2 (02 2017), 1–20. https://doi.org/10.1371/journal.pone.0172493

[50] Michalis Lioudakis, Stamatis Outsios, and Michalis Vazirgiannis. 2019. An Ensemble Method for Producing Word Representations for the Greek Language. *arXiv preprint arXiv:1912.04965* (2019).

[51] Giorgio Lucarelli, Xenofon Vasilakos, and Ion Androutsopoulos. 2007. Named entity recognition in greek texts with an ensemble of svms and active learning. *International Journal on Artificial Intelligence Tools* 16, 06 (2007), 1015–1045.

[52] G. Mamakis, A. G. Malamos, A. Axaridou, Y. Kaliakatsos, and A. Ware. 2005. An Algorithm for Automatic Content Summarization in Modern Greek Language. In *2005 International Conference on Information and Communication Technology*. 579–591.

[53] Nikos Manousakis and Efstathios Stamatatos. 2017. Devising Rhesus: A strange 'collaboration' between Aeschylus and Euripides. *Digital Scholarship in the Humanities* 33, 2 (04 2017), 347–361.

[54] Emmanouil Marakakis, Haridimos Kondylakis, and Papakonstantinou Aris. 2017. APANTISIS: A Greek Question-Answering System for Knowledge-Base Exploration. In *Strategic Innovative Marketing*. Springer, 501–510.

[55] G Markopoulos, G Mikros, A Iliadi, and M Liontos. 2015. Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features. In *Cultural Tourism in a Digital Era*, V Katsoni (Ed.). Springer International Publishing, Cham, 373–383.

[56] George Mavridis. 2018. Fake news and social media: How Greek users identify and curb misinformation online. (2018).

[57] George Mikros and Kostas Perifanos. 2015. Gender Identification in Modern Greek Tweets. In *Recent Contributions to Quantitative Linguistics*, Arjuna Tuzzi, Martina Benesová, and Ján Macutek (Eds.). Quantitative Linguistics, Vol. 70. DeGruyter, 75–88.

[58] O. N. Nikolaenkova. 2019. Applying CLP to machine translation: a Greek case study. *Journal of applied linguistics and lexicography* 1 (2019), 69–79. Nikolaenkova, O. N. (2019) Applying CLP to machine translation: A Greek case study. Journal of Applied Linguistics and Lexicography, 1 (1), 69–78. DOI: 10.33910/2687-0215-2019-1-1-69-78.

[59] Georgios Ntais. 2006. Development of a Stemmer for the Greek Language. *Department of Computer and Systems SciencesMaster Thesis at Stockholm University/Royal Institute of Technology, pp1-40* (2006).

[60] Kostas Ntzios, Basilios Gatos, Ioannis Pratikakis, Thomas Konidaris, and Stavros J Perantonis. 2007. An old greek handwritten OCR system based on an efficient segmentation-free approach. *International Journal of Document Analysis and Recognition (IJDAR)* 9, 2-4 (2007), 179–192.

[61] Students of the CSD-UOC. 2006. Mitos Greek Stemmer. https://github.com/YannisTzitzikas/GreekMitosStemmer.

[62] Giorgos Orphanos, Dimitris Kalles, A Papagelis, and Dimitris Christodoulakis. 1999. Decision trees and NLP: A case study in POS tagging. In *Proceedings of annual conference on artificial intelligence (ACAI)*.

[63] Stamatis Outsios, Christos Karatsalos, Konstantinos Skianis, and Michalis Vazirgiannis. 2019. Evaluation of Greek Word Embeddings. *arXiv preprint arXiv:1904.04032* (2019).

[64] Stamatis Outsios, Konstantinos Skianis, Polykarpos Meladianos, Christos Xypolopoulos, and Michalis Vazirgiannis. 2018. Word Embeddings from Large-Scale Greek Web Content. *arXiv preprint arXiv:1810.06694* (2018).

[65] Harris Papageorgiou, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. European Language Resources Association (ELRA), Athens, Greece.

[66] Konstantina Papanikolaou, Haris Papageorgiou, Nikos Papasarantopoulos, Theoni Stathopoulou, and George Papastefanatos. 2016. "Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter. In *Social Media in the Newsroom, Papers from the 2016 ICWSM Workshop, Cologne, Germany, May 17, 2016 (AAAI Workshops)*, Bahareh Rahmanzadeh Heravi and Arkaitz Zubiaga (Eds.), Vol. WS-16-19. AAAI Press.

[67] Monarch Parmar, Naman Jain, Pranjali Jain, P Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2019. NLPExplorer: Exploring the Universe of NLP Papers. arXiv:cs.IR/1910.07351

[68] Eleni Partalidou, Eleftherios Spyromitros-Xioufis, Stavros Doropoulos, Stavros Vologiannidis, and Konstantinos I Diamantaras. 2019. Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 337–341.

[69] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep Learning for User Comment Moderation. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 25–35.

[70] Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. *arXiv preprint arXiv:1903.05587* (2019).

[71] George Petasis, Vangelis Karkaletsis, Dimitra Farmakiotou, George Samaritakis, Ion Androutsopoulos, and Constantine Spyropoulos. 2001. A Greek Morphological Lexicon and its Exploitation by a Greek Controlled Language Checker. In *Proceedings of the 8th Panhellenic Conference on Informatics*. 8–10.

[72] Georgios Petasis, Georgios Paliouras, Vangelis Karkaletsis, Constantine D Spyropoulos, and Ion Androutsopoulos. 1999. Resolving part-of-speech ambiguity in the Greek language using learning techniques. *arXiv preprint cs/9906019* (1999).

[73] Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. *arXiv preprint arXiv:2003.07459* (2020).

[74] Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, and Stelios Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT*. Citeseer.

[75] Prokopis Prokopidis and Harris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. 102–106.

[76] Lorena Pérez-Hernández and Francisco Ruiz de Mendoza. 2011. A Lexical-Constructional Model Account of Illocution. *Vigo International Journal of Applied Linguistics* 8 (01 2011), 98–137.

[77] Bruce Robertson and Federico Boschetti. 2017. Large-Scale Optical Character Recognition of Ancient Greek. *Mouseion* 14, 3 (2017), 341–359.

[78] Martina Astrid Rodda, Philomen Probert, and Barbara McGillivray. 2019. Vector space models of Ancient Greek word meaning, and a case study on Homer. *Traitement Automatique des Langues (TAL)* 60 (2019).

[79] Jeff Rydberg-Cox. 2011. Social networks and the language of greek tragedy. (2011).

[80] Hanna Sababa and Athena Stassopoulou. 2018. A classifier to distinguish between cypriot greek and standard modern greek. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 251–255.

[81] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, 56–66.

[82] Spyridon Saroukos. 2009. *Enhancing a greek language stemmer-efficiency and accuracy improvements*. Master's thesis.

[83] Kyriakos Sgarbas, Nikos Fakotakis, and George Kokkinakis. 1995. A PC-KIMMO-based morphological description of Modern Greek. *Literary and Linguistic Computing* 10, 3 (1995), 189–201.

[84] Fotini Simistira, Adnan Ul-Hassan, Vassilis Papavassiliou, Basilis Gatos, Vassilis Katsouros, and Marcus Liwicki. 2015. Recognition of historical Greek polytonic scripts using LSTM networks. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 766–770.

[85] Merkourios Simos and Nikolaos Nikolaidis. 2016. Greek sign language alphabet recognition using the leap motion device. In *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*. 1–4.

[86] Nikolaos Spatiotis, Michael Paraskevas, Isidoros Perikos, and Iosif Mporas. 2017. Examining the Impact of Feature Selection on Sentiment Analysis for the Greek Language. In *International Conference on Speech and Computer*. Springer, 353–361.

[87] Nikolaos Spatiotis, Isidoros Perikos, Iosif Mporas, and Michael Paraskevas. 2019. Examining the Impact of Discretization Technique on Sentiment Analysis for the Greek Language. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 1–6.

[88] E Takoulidou, V Sosoni, K Kermanidis, and M van Zaanen. 2016. Social media and NLP tasks: Challenges in crowdsourcing linguistic information. In *2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. IEEE, 53–58.

[89] Charalambos Themistocleous. 2019. Dialect Classification From a Single Sonorant Sound Using Deep Neural Networks. *Frontiers in Communication* 4 (2019), 64.

[90] A Toulakis, A Papamichail, D Triantafylloy, T Tasoulas, and K Kermanidis. 2019. Language Processing for Predicting Suicidal Tendencies: A Case Study in Greek Poetry. In *Artificial Intelligence Applications and Innovations: AIAI 2019 IFIP WG 12.5 International Workshops: MHDW and 5G-PINE 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings*, Vol. 560. Springer, 173.

[91] Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: The case of the Greek Referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 367–376.

[92] Adam Tsakalidis, Symeon Papadopoulos, Rania Voskaki, Kyriaki Ioannidou, Christina Boididou, Alexandra Cristea, Maria Liakata, and Yiannis Kompatsiaris. 2018. Building and evaluating resources for sentiment analysis in the Greek language. *Language Resources and Evaluation* 52, 4 (2018), 1021–1044.

[93] Christos Tsalidis, Mavina Pantazara, Panagiotis Minos, and Elena Mantzari. 2010. NLP tools for lexicographic applications in Modern Greek. *eLexicography in the 21st Century: New Challenges, New Applications (Proceedings of eLex2009), S. Granger and M. Paquot* (2010), 457–462.

[94] Leonidas Tsekouras, Georgios Petasis, George Giannakopoulos, and Aris Kosmopoulos. 2020. Social Web Observatory: A Platform and Method for Gathering Knowledge on Entities from Different Textual Sources. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2000–2008.

[95] Charalampos Tsimpouris, Kyriakos Sgarbas, and Sofia Panagiotopoulou. 2014. Acronym identification in Greek legal texts. *Digital Scholarship in the Humanities* 30, 3 (03 2014), 440–451.

[96] Mara Tsoumari and Georgios Petasis. 2011. Coreference Annotator-A new annotation tool for aligned bilingual corpora. In *Proceedings of The Second Workshop on Annotation and Exploitation of Parallel Corpora*. 43–52.

[97] Aristides Vagelatos, Elena Mantzari, Mavina Pantazara, Christos Tsalidis, and Chryssoula Kalamara. 2011. Developing tools and resources for the biomedical domain of the Greek language. *Health Informatics Journal* 17, 2 (2011), 127–139.

[98] G Vamvakas, B Gatos, and S Perantonis. 2010. Handwritten character recognition through two-stage foreground sub-sampling. *Pattern Recognition* 43, 8 (2010), 2807–2816.

[99] G Vamvakas, N Stamatopoulos, B Gatos, I Pratikakis, and S Perantonis. 2007. Greek Handwritten character recognition. In *Proceedings of the 11th Panhellenic Conference in Informatics*. 18–20.

[100] S. Varlokosta, S. Stamouli, A. Karasimos, G. Markopoulos, M. Kakavoulia, M. Nerantzini, A. Pantoula, V. Fyndanis, A. Economou, and A. Protopapas. 2016. A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC-2016), Workshop "Resources and Processing of linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID-2016)*, Dimitrios Kokkinakis (Ed.). Slovenia, 14–21.

[101] Nick White. 2012. Training Tesseract for Ancient Greek OCR. *Eiiruzov* 28–29 (2012).