Corpora and language variation in Greek

Fatima Eloeva[1], Maxim Kisilier[2], Olga Nikolaenkova[3]
[1]Department of Classical Philology, Vilnius University, Lithuania
[2]Hellenic institute, Saint Petersburg State University, Russia
[3]Department of General Linguistics, Saint Petersburg State University, Russia
https://doi.org/10.36505/ExLing-2019/10/0018/000380

Abstract

It seems that there are certain linguistic situations ideally adapted for the usage-based approach and the Greek case is one of them. This is a corpus-based study, which is based on the analysis and processing of a large variety of Greek texts in everyday spoken interactions. As our starting point, we argue the panchronic character of Greek lexicon, its extreme conservatism and the man-made character of the formation of the Greek literary standard. Some practical issues such as the choice between monotonic and polytonic orthography, lexemes tagging to obtain more data are addressed here.

Key words: language corpus, diglossia, modern Greek, language variation.

Introduction

It seems that there are certain linguistic situations ideally adapted for the usage-based approach and the Greek case is one of them. Anyhow up to the present moment, Modern Greek does not posess its complete **electronic corpora** like most European languages.

There is no doubt that the Greek language question should not be neglected while creating linguistic corpora of Modern Greek. At the moment there are at least three corpora of Modern Greek but none of them currently reflects the existence of Greek diglossia.

The Hellenic National Corpus (http://hnc.ilsp.gr/) was created in the Institute for Language and Speech Processing in Athens. It consists of more than 47 million words in various types of texts. However, the texts with any peculiarities (for example dialectisms) are not included as well as the texts written before 1990. It means that the Hellenic corpus does not take into account Modern Greek literature.

The Corpus of Spoken Greek is a part of the Greek talk-in-interaction and Conversation Analysis research project, directed by prof. Pavlidou in the Institute of Modern Greek Studies at the Aristotle University of Thessaloniki (http://ins.web.auth.gr/index.php?option=com_content&view=article&id=626&Itemid=251&lang=en). It was originally designed for the qualitative analysis of language and linguistic communication, especially from the perspective of Conversation Analysis. The material has been drawn from naturally-occurring circumstances of communication and comprises everyday conversations among

friends and relatives, classroom conversations, telephone calls and various types of TV broadcasts.

Material

In the present paper, we introduce Corpus of Modern Greek (http://web-corpora.net/GreekCorpus/search/?interface_language=en). It consists of approx. 35.7 million tokens — but it is not restricted to any specific type of texts. The majority of texts incorporated in Corpus of Modern Greek come from contemporary Greek newspapers (Η Καθημερινή, Μακεδονία, Το Βήμα, Ελευθεροτυπία). However, there are also fiction, poetic, official, scientific, and religious texts, both original and translated, that were created in the 19th or in the 20th centuries. Thus, both the authors often regarded as supporters of Katherevusa, like Papadiamandis or Viziinos and demoticists (Nikos Kazandzakis) are present (Arkhangelskiy & Kisilier, 2018).

All texts have been morphologically annotated. It means that each word is provided with a lemma (dictionary form) and a set of morphosyntactic tags (such as case values, number values, etc.) which can be used in a search query. Morphological annotation was carried out with the help of a digital grammatical dictionary and a morphological analyzer (UniParser). Unfortunately, corpus still has no disambiguation, i. e. each word was annotated with all possible morphological analyses in all contexts (cf. Kuzmenko & Mustakimova, 2015). Another peculiarity of this corpus is a dictionary module that enables translation from English into Greek.

It is impossible to avoid the Greek language issue in the corpus since a variety of literary and non-literary texts is involved. The easiest solution was to tag "Katharevusa" or "Demotic Greek" authors or texts. Thus, for example, Alexandros Papadiamandis and his oeuvres were tagged as "Katharevusa" while Nikos Kazandzakis should be tagged as "Demotic Greek". Recent corpus-oriented studies in Katharevusa Greek (Yakovleva, 2017) clearly demonstrated that this approach is inapplicable to non-Demotic material. We are going to illustrate it using the example of the renown Greek writer Alexander Papadiamandis (1851–1911) whose language is pretty often described as Katharevusa. This statement is mostly based on multiple archaic features in morphology or word formation:

τὴν διήγησιν 'story' (accusative) vs. την διήγηση

ἐσηκώθη '[he] rose' vs. σηκώθηκε

ἦσαν '[they] were' vs. ήταν

and vocabulary:

ὕδωρ ‘water’ vs. νερό

οὖς ‘snake’ vs. φίδι

ωρίζω

A thorough analysis of his texts shows that there are a lot of lexical borrowings from

Turkish — ταμάμ ‘okay’ (< Turk. *tamam*)

Albanian — τσούπα ‘girl’ (< Alb. *çupë*)

Slavic — βάλτος ‘swamp’ (< Slav. *блато*)

along with dialectal and vernacular features:

πλιά ‘more’ vs. *πιά*

γρουνίζω ‘[I] know’ vs. γνωρίζω

Evidently, it would be a mistake to treat these words and morphological forms as Katharevusa, and if Papadiamandis and his texts are tagged as “Katharevusa”, all forms and words included in these texts will be compulsory ascribed to Katharevusa and may lead the user to various false conclusions concerning the language. We are sure that corpus may not decide for the user but it should provide enough information to let the user make his own decision.

Corpus of Modern Greek is now being moved to the new platform able to provide whole morphological paradigm with different statistic data, it will have a possibility for a dictionary mode and simultaneous representation of textual and audio material, later there shall be a module for translating from Russian into Greek, etc.

Conclusions

Working on a new approach to the Greek language issue implementation includes two rather different options:

First of all, it is important to have a choice between monotonic and polytonic orthography. This choice formally exists even today, but it is supported only by few texts in polytonic. Introduction of new polytonic texts will certainly face difficulties in their recognition, so special software is required.

The second option has to do with a new way of tagging. We suggest that instead of texts or authors one should tag lexemes, paradigms or even separate forms. The opposition should not be binary ("Demotic" vs. "Katharevusa") but a triple one: "Demotic" vs. "Archaic" and "General" for the units that do not differ in Katharevusa and Demotic Greek. This kind of tagging will give the user new advanced search tools like choice of flexions or inflexion types from the point of view of their stylistic properties.

As a result, the user will have quantitative data for the texts he intends to analyze and he will be able to decide himself to which language or stylistic variant this text should be attributed.

Acknowledgements

Corpus of Modern Greek compiled by Timofey Arkhangelskiy and Maxim Kisilier within the program "Corpus Linguistics" of the Russian Academy of Sciences. New platform design has been performed by Timofey Arkhangelskiy with Aleksander Rusakov and Maxim Kisilier financial support.

References

Arkhangelskiy T.A., Kisilier M.L. 2018. Greek corpora: achievements, goals and prospects. (Корпуса греческого языка: достижения, цели и задачи) Indoeuropean linguistics and classical philology (Индоевропейское языкознание и классическая филология) XXII. 1, 50–59.

Kuzmenko E., Mustakimova E. 2015. Automatic disambiguation in the corpora of Greek and Yiddish. Annual International Conference "Dialogue" (2015, 27–30.05.2015), Moscow

Yakovleva A.V. 2017. Ablative and allative marking of spatial position in katharevousa: corpus study (Аблативное и аллативное маркирование положения в пространстве в кафаревусе: корпусное исследование). Thesis presented for Master degree, Language theory and computational linguistics department, National Research University Higher School of Economics.