Investigating the linguistic representativeness of Early Modern Greek Corpora

Eleni Karantzola
University of the Aegean
karantzola@rhodes.aegean.gr

Yannis Kostopoulos
University of the Aegean
g.kostopoulos@aegean.gr

Konstantinos Sampanis
University of the Aegean
k.sampanis@rhodes.aegean.gr

Abstract

Following a poorly documented period in the history of vernacular Greek (6th-12th c.), the late 15th century sets the beginning of a linguistic era characterized by a quantitatively and qualitatively incomparable production of prose texts written in "common" language. It is at this point that classicizing Greek stops dominating in writing, and a new linguistic variety – albeit a very diverse and fluid one – Early Modern Greek (EMG) starts growing rapidly as a literacy language. The development of this new variety is manifested in its widespread use as *literary language* (in texts with aesthetic function), as well as in its use as a simple *scripta*, namely a written vernacular for legal, administrative, commercial, and other functions. Despite its significance in the history of Greek, this period remains to a large extent unexplored and underrepresented in Greek language corpora. On this view, our understanding of EMG depends crucially on the representativeness of the few available corpora.

The aim of this paper is to investigate the linguistic representativeness of EMG corpora, and to explore possible associations between observed linguistic patterns and corpora design. Focusing on the distribution of contrastive and reformulation markers, our study reveals that the linguistic data illustrated in the available EMG corpora are divergent and largely dependent on the representation of variables, such as text form (poetry/prose), period, geographical region, and genre.

Keywords: Early Modern Greek, corpora representativeness, contrastive markers, reformulation markers.

1 Introduction

The history of Modern Greek language is marked by the co-existence of at least two competitive registers: a "high", learned one, and a "low", non-learned register, often referred to as "vulgar" or "vernacular" Greek (see Hinterberger 2006, Holton and Manolessou 2010). The former, heavily influenced by Attic and Koine Greek, is the register that monopolized the language of ecclesiastical literature and administration, and dominated literary production, from the 4th century until the 11th century, at least. The latter, generally thought of as the register of oral, everyday communication, remains to a large extent unknown. Intrinsically connected to spoken discourse, the exact form of the vernacular register is lost with its speakers. A significant number of texts from the late 15th century and onwards provides evidence for the spread of a linguistic variety whose basic characteristic is that it no longer adheres to the standards of archaic language. However, this variety, usually called Early Modern Greek, is far from homogenous: apart from the amount of ongoing linguistic developments, the available sources on Early Modern Greek reveal significant variations related to local dialects, language contact, and authors' personal style. In view of this variability, any attempt to provide a general description of Early Modern Greek inevitably calls for reliable, quantitative data from representative corpora.

Aiming at contributing to a reliable description of Early Modern Greek, in this paper, we investigate the representativeness of four relevant corpora, namely the Thesaurus Linguae Graecae corpus (TLG), the corpus of Vernacular Greek created by the Centre for the Greek Language (CGL), the Anthology of vernacular prose texts (Kakoulidou-Panou et al. in press), and the collection of

Autograph manuscripts of the 16th and 17th century (Papaioannou 2016). Our study focuses on linguistic representativeness (Biber 1993; Sinclair 1996), and in particular, on the representation of contrastive and reformulation markers in the available EMG corpora. Through a number of empirical tests, we explore comparatively the use and frequency of contrastive and reformulation markers in EMG corpora, and we examine possible associations between our findings and corpora design. On this view, our analysis departs from the dominant approach that treats representativeness as an issue pertaining to corpus design alone, and adopts a perspective which evaluates corpora representativeness on the basis of the empirical investigation of the illustrated linguistic data (see also Gray, Egbert and Biber 2017).

In the following section, we try to provide a working definition for the term *Early Modern Greek*. Section 3 illustrates EMG corpora characteristics, while Section 4 presents the methodology that we used in our study. The findings of our empirical investigation are given in Section 4 and are further discussed in Section 5. In the concluding Section, we summarize our findings with suggestions for improving EMG corpora representativeness.

2 Early Modern Greek: Periodization and characteristics

Any attempt to slice the history of a linguistic variety into separate pieces with a given start and a definite ending inevitably involves theoretical abstractions and approximations that rarely go undebated. The periodization of Early Modern Greek is no exception. For some authors, the era examined in this paper does not suggest some distinct linguistic period but falls into some larger part of the history of Greek, either Medieval or Modern Greek[1]. In a nutshell, the argument against the assumption of a distinct Early Modern Greek era is that most linguistic developments observed in this period had started to appear in previous phases of Greek (with some of them going back to the Koine), while some of the characteristic changes of the era continued to exist even after the standardization of the Modern Greek language (19th century). Other researchers hold that the amount and the frequency of significant linguistic changes observed in the temporal span between 1500 and 1700 justify the view that Early Modern Greek is a distinct linguistic era (Holton and Manolessou 2010, Holton et al. 2019). An elaboration of the debate, or some contribution to the periodization of Greek, goes beyond any ambition of the present study. In our analysis, we use the term "Early Modern Greek" in order to refer to a relatively distinct, but non-unified, linguistic form of Greek, which seems to have thrived in the second half of the 15th century and until the emergence of the Modern Greek state (1830)[2].

The annexation of the last remaining Byzantine territories by the Ottoman Empire introduced a long period in which Greek-speaking communities were separated in different states – mainly under Ottoman or Venetian rule – and different linguistic contexts. The lack of a central Greek-speaking authority, or any other institution of standardization, favored the rise of decentralized, local vernaculars, whose use was generalized and spread beyond poetry and oral communication. This process seems to have followed similar developments that took place at the same time in Western Europe, in the context of the first "ecolinguistic revolution" (see Baggioni 1997: 74). The construction of national states (Spain, France, England), the humanist movement and the extension of the literate public, the rise of national literatures, the religious factor, i.e., Reform and Counter-Reform (1517-1580) are some of the socio-cultural factors that contributed to the emergence of vernaculars as literacy languages and, consequently, to the radical change of both the Greek and the Western European ecology of communication in the 16th and 17th centuries. At the same time, vernaculars were stabilized and codified through grammars,

[1] For an overview of the debate on the periodization of Early Modern Greek see Kakoulidou-Panou et al. in press: Introduction.
[2] Burke (2004) argues that the period between 1450 and 1789 should be regarded as a discrete period for the languages spoken or written in Europe – at least from a sociolinguistic point of view. In what concerns Greek, apart from the invention of the printing press by Gutenberg, the fall of the Byzantine Empire (1453) suggests another significant "external" (historical and cultural) criterion for the linguistic periodization and an additional indication of a turning point to the modern era.

dictionaries, spelling guides etc. (see Auroux 1994). The massive production of vernacular texts, which has been fundamental for the elaboration of Western European vernacular languages, was equally significant for Early Modern Greek. An impressive amount of text production in thematic areas, such as legal arrangements, science, geography, history, and philology, give evidence of the impact of the wider ecolinguistic and techno-linguistic revolutions on the growth of Early Modern Greek. As it was the case for other European vernaculars, the rise of the movement for national independence, and the consequent introduction of the ideal of a unified, national language (late 18th - early 19th century), lead to the subsidence of Early Modern Greek.

On the level of linguistic description, Early Modern Greek is distinguishable by several phonological, syntactic, morphological, and lexical phenomena, including the development of palatalization and the raising of vowels in Northern dialects, the change in the placement of clitics, which tend to be established in preverbal position, the total loss of the infinitive, the generalization of the periphrasis *θα* + subjunctive for the expression of the future tense, the loss of the aorist gerund, and the prevalence of structural and lexical borrowings from Italian and Turkish (see Holton and Manolessou 2010). Apart from the developments that first appear in the 16th century, EMG is also characterized by the generalization of phenomena that appeared in previous phases of Greek, such the leveling of nominal paradigms, the unification of past tense endings, the use of active gerunds in *-οντας*, and the restriction of infinitives (for a comprehensive overview, see Holton et al. 2019).

The debate on the existence of a distinct Early Modern Greek era, as well as the main characteristics of the Greek language in the period between 16th-17th century – which are essentially developments extending backwards or forwards in time –, indicate that the variety that concerns us shapes a very *dynamic* synchrony marked by numerous changes and transitions. Moreover, beyond the diachronic dimension, the linguistic situation in which we are interested in this paper is largely defined by the emergence and spread of local vernaculars that do not necessarily follow the same path or pace in their development. These facts suggest that EMG is a linguistic period of extended variation, whose description requires representative corpora. At the time of writing this paper, we are aware of only four corpora that cover the EMG era. In the next section, we present the characteristics of these corpora before proceeding to the investigation of their representativeness.

3 Early Modern Greek corpora

Unlike Ancient or Koine Greek, which are very well documented in digital text collections, more recent phases of the history of the Greek language remain highly underrepresented in the available corpora. The Hellenic National Corpus (HNC), a collection of various texts amounting to 47,000,000 words, and the Corpus of Greek Texts (CGT), a collection of both written and oral discourse made up of 3,000,000 words, reflect the use of contemporary Modern Greek, but do not cover any linguistic period before 1976[3]. As for Modern Greek in its earlier phases – and especially in the period between 16th and 18th centuries – corpus representation is rather scarce and restricted to a small amount of collections, such as the Thesaurus Linguae Graecae (TLG), the digital collection of texts created by the Centre for the Greek Language (CGL), the Anthology of vernacular prose texts of the 16th century (Kakoulidou-Panou et al. in press), and the collection of autographs of the 16th and 17th centuries (Papaioannou 2016), which is digitally available at the repository of the University of the Aegean (Laboratory of Linguistics of SE Mediterranean, Rhodes).

Thesaurus Linguae Graecae is undoubtedly the largest and most prestigious corpus of Greek language, made up of 110,000,000 words that cover the history of Greek from Homer to Byzantine times. In the course of its recent expansions, TLG integrated 75 full texts (37 authors) from Early

[3] On the characteristics of the HNC and CGT corpora see Hatzigeorgiou et al. 2001 and Goutsos 2010, respectively.

Modern Greek which amount to 2,636,469 words. Reflecting contemporary editions of both prose and poetry, and having an average length of 35,153 words (min. 192 words - max. 262,455 words), the EMG texts included in the TLG corpus offer a considerable sample of the Greek vernacular literature of the 16th and 17th centuries. As regards the genres included in the corpus, TLG offers statistics only for six colloquial texts[4].

The digital collection of texts created by the Centre for the Greek Language (CGL) is an anthology of vernacular literacy comprised of 239,814 words. With a text length average of 3,114 words, the CGL collection includes only fragments of texts and is oriented towards the representation of poetic production. Out of a sum of 77 fragments, 49 are poetic works and 28 prose. In terms of periodization, the corpus covers the period between 12th and early 17th century, extending, thus, beyond the temporal delimitation of Early Modern Greek that we assumed in Section 2. The texts included in the CGL corpus are taken from manuscripts produced in the EMG period, as well as from manuscripts that illustrate works of the EMG era but were created in later periods. This fact raises some doubts about the authenticity of the linguistic data offered by the CGL.

The Anthology of vernacular prose texts (hereafter "Anthology"), edited by Kakoulidou-Panou, Karantzola and Tiktopoulou (in press), is a collection of 250 prose text excerpts from manuscripts or original printed editions of the 16th century. The average length of these excerpts is 630 words (min. length 63 words - max. length 1,712 words), while the total sum of the corpus amounts to 155,717 words. The relatively small number of words comprising the Anthology, and the short length of the excerpts used, suggest a possible challenge for the representativeness of the data. On the other hand, the corpus created by these excerpts is structured into 11 thematic categories according to the content of texts (*Forewords*, *Theology*, *Sermons*, *Lives of Saints*, *Philological texts*, *History-Chronicles*, *Geography-Travel Literature*, *Sciences*, *Legal texts*, *Notary books*, *Correspondence*), a feature that adds an aspect of stratification to the designed collection.

	TLG	**CGL**	**ANTHOLOGY**	**AUTOGRAPHS**
words	2,636,469	239,814	155,717	44,026
number of texts	75	77	250	101
length	35,153 [192 - 262,455]	3,114 [1,274 - 10,672]	630 [63 - 1,712]	435 [111 - 980]
period	16th -17th	12th-17th	16th	16th-17th
form (poetry/prose)	poetry-prose	poetry- prose	prose	prose
genres/thematic categories	ONLY FOR COLLOQUIAL TEXTS	NO	YES	NO

Table 1: Early Modern Greek corpora characteristics

The collection of autographs (hereafter "Autographs"), edited by Papaioannou (2016), is a selection of 101 excerpts from prose texts of the 16th and 17th century. The fragments included in the Autographs have an average length of 435 words (min. length 111 words – max. length 980 words) and make up a total of 44,026 words. Including only autograph manuscripts of the era, the corpus of Autographs illustrates authentic cases of 16th and 17th century written discourse.

[4] According to the information provided by the TLG corpus, these colloquial texts include 2 Chronographies, 1 Hagiography, 1 Comic text, 1 Historiography, and 1 Fabula. Despite any apparent similarities, there is no direct correspondence between the genre characterization used in TLG and the characterization applied by Kakoulidou-Panou, Karantzola and Tiktopoulou in the Anthology.

As Table 1 shows, the existing corpora on EMG vary significantly, not only in their length, but most importantly in aspects related to the period, the form of discourse (poetry/prose), and the genres covered. The question is whether these differences are reflected in the data offered, and consequently, in the linguistic representativeness of EMG corpora. In the following sections, we try to address the issue through a number of empirical tests.

4 Methodology

The question of whether a given corpus is representative of the population or the variation it targets is always a major concern in corpus-based studies. In the case of EMG, this question becomes even more crucial, since the available corpora are few and relatively small. Moreover, EMG corpora are based on written texts alone, which means that they are more or less adapted to the restrictions imposed by writing and genre conventions, departing, thus, from the ordinary, spoken language of the period. In any case, if we are to use the corpora under discussion in the description of EMG, we first have to provide an answer about their representativeness.

There is widespread assumption – usually tacit, but sometimes explicit too (see, for instance, Hanks 2012) – that the bigger a corpus is, the higher its representativeness. If this is the case, then all four corpora examined in this paper are very unlikely to be representative of EMG. According to several scholars though, size is not a solid criterion in evaluating the representativeness of a corpus (see, for instance, Raineri and Debras 2019). For Biber (1993:244), "representativeness refers to the extent to which a sample includes the full range of variability in a population". Population is considered to be the sum of texts for a given variety – in our case, the sum of EMG texts. Variability, on the other hand, can refer to two different things: a) situational variability, i.e., extralinguistic parameters, such as author, addressee, gender, topic, etc., and b) linguistic variability, that is, parameters concerning the use and frequency of linguistic elements (Biber 1993; McEnery et al. 2006). Representativeness is further supported by the sampling of a corpus, which is entirely defined on non-linguistic considerations (e.g., social and demographic parameters), and the balance of a corpus, that is, the proportionality illustrated in a corpus with respect to frequencies (linguistic and situational) observed in the population (i.e., the sum of texts).

Considering the characteristics of EMG corpora discussed in the previous section, we expect that TLG, CGL, Anthology, and Autographs show different degrees of representativeness. Anthology and Autographs, for instance, do not include any poetic texts and, therefore, they are not expected to cover the full extent of EMG variability, at least in what concerns the situational parameters involved. Similarly, TLG and CGL lack a number of genres, such as notary and legal texts (the latter appearing only scarcely in TLG), which make up an important part of EMG literary production. These omissions also affect the sampling of EMG corpora, which seems problematic in addressing the external parameters involved in EMG literary production. In view of these flaws, balance does not even come into question.

Nevertheless, the criteria on corpora representativeness briefly discussed above are mainly concerned with corpora design, not the phenomena illustrated in a given corpus. Parameters such as the representation of situational variability, sampling, and balance tell us how a corpus is made with respect to certain variables, and consequently, whether findings from that corpus can be legitimately generalized to the targeted population. On this view, estimations on corpora representativeness grounded on design parameters are probabilistic rather than empirical (see also Gray et al. 2017). In some cases, we want to know whether linguistic patterns observed in a corpus is representative of a certain linguistic variety, regardless of whether the design of the overall corpus is representative of the targeted population. And, certainly, in some cases we must have an empirical answer on the representativeness of corpora with flawed design, simply because we do not have any alternatives for studying the language of a given period. Considering that the corpora discussed in this paper are our only sources for the quantitative

study of EMG, we are compelled to come up with an empirical evaluation of their linguistic representativeness.

The methodology that we followed in investigating the representativeness of EMG corpora is based on a neat assumption: if the language illustrated in TLG, CGL, Anthology, and Autographs is representative of EMG, then the linguistic uses and frequencies observed in these corpora should not show significant divergences. If, on the other hand, we observe that the corpora under investigation exhibit divergencies in linguistic patterns, then we should assume that these corpora have varying degrees of representativeness, which reflect respective differentiations in the external (or situational) variables that they include. Following this assumption, in our investigation we tried first to explore how EMG corpora behave with respect to certain linguistic phenomena, and second, to identify associations between possible linguistic divergencies and divergencies in the design of the discussed corpora. In order to restrict our investigation in a small set of phenomena, we explored the behavior of EMG corpora with respect to certain discourse markers, and in particular, contrastive and reformulation markers.

Discourse markers (also, discourse connectives, discourse particles or discourse operators) are a non-unified class of elements usually including connectors (e.g., but, nevertheless), adverbials (e.g., now, anyway) propositional phrases (e.g., on the contrary, after all). According to most scholars, discourse markers have a non-propositional contribution to the utterance that contains them, and perform a connective function, providing instructions on the way discourse segments or communicated propositions in general (explicit or implicit) should be related in the interpretation of an utterance (see Schiffrin 1987; Blakemore 1987, 2002; Fraser 1996, 1999; Schourup 1999). Optionality, weak clause association, and the tendency to appear in sentence-initial position have also been proposed as defining properties of discourse markers, however they do not seem to apply for all elements usually included in the class (Schourup 1999). In some works, discourse markers are considered to be associated to oral discourse, but written texts also exhibit a considerable use of non-propositional, connective expressions, although not necessarily the same as the ones appearing in oral texts (see Brinton 1996). For some authors, elements usually labeled "discourse markers" differ from lexical, conceptual expressions in that they are unavailable to the speakers' conscious knowledge (Blakemore 2002), and remarkably difficult to translate (Furkó 2014). These properties suggest that discourse markers are direct links to a speaker's/author's linguistic intuitions and are less susceptible to adaptations – interlinguistic and intralinguistic. The association between discourse markers and the speakers' authentic style has been effectively integrated in stylometric studies, which use discourse markers and other functional words as tools for authorship attribution (Stamatatos 2009). Drawing on these considerations, in our study, we used discourse markers as relatively solid indications of authors' linguistic intuitions and authentic style.

Our investigation is focused on two categories of discourse markers: contrastive markers and reformulation markers. Contrastive markers are elements with non-truth conditional meaning indicating the existence of some sort of contrast between two discourse segments or between a discourse segment and an assumption previously communicated in the discourse. Drawing on the literature on contrastive conjunctions, we take contrast to include three main relations: a) semantic opposition, b) denial of expectations, and c) correction/substitution (Lakoff 1971; Anscombre and Ducrot 1977; Blakemore 1989; Mauri 2008). In EMG, these relations are covered by a number of elements, including alla, ami, ma, omos, and pouri (αλλά, αμή, μα, όμως, πούρι, see Karantzola and Kalokerinos 2005). In our study, we examined the distribution of three of these markers, namely, alla, ami, ma. Examples (1)-(3) below illustrate some characteristic uses of these markers:

(1) *Kai eseis katharoi eisten,* ***alla*** *ochi oloi.*
"And you are clean, ***but*** *not you all."*
(Kartanos, ed. Venice 1536)[5]

[5] The fragments appearing in examples (1)-(6) are taken from Kakoulidou-Panou et al. in press.

(2) *Kai etouton den eginiken mia i dyo,* ***ma*** *polles kai polles fores…*
"And this did not happen once or twice, ***but*** *many, many times…"*
(Morezinos, monastery of Xiropotamos, ms 202, 1602)

(3) *Dioti oi epistimes mathainontai ochi monon me tin ellinikin glossan,* ***ami*** *kai me pasan allin glossan…*
"Because sciences are learned not only through Greek language, ***but*** *also through any other language…"*
(Sofianos, ms Par. Gr. 2592 (autograph), 16th c.)

Reformulation markers, on the other hand, are expressions indicating that the following discourse segment has a reformulative function with respect to a preceding chunk of discourse. Despite the bulk of relevant literature, the definition of reformulation remains under debate. Most authors would agree that reformulation is a process of reinterpretation, which involves some sort of identity (semantic, pragmatic or enunciative) between two discourse segments X and Y (Gulich and Kotschi 1983; Rossari 1994; Steuckardt 2009). This identity can lie between repetition and complete correction (Conceiçao 2005: 82). The reformulating segment, usually introduced by a marker of reformulation, can be either equally, more, or less specific than the segment X that it targets (see Meyer 1992). Blakemore (1997: 9) notes that reformulations are often used with a pedagogical and even patronizing purpose. Culpeper (1994) and Blakemore (1994) observe that reformulations often communicate ideological, educational, and social distance between speakers/authors and their audience. EMG has a considerable number of reformulation markers including *igoun, itoi, dilonoti, diladi, toutestin* (*ήγουν*, *ήτοι*, *δηλονότι*, *δηλαδή*, *τουτέστιν*) which seem to compete for the same functional space. In our investigation, we focused on three reformulation markers, i.e., *igoun, dilonoti, diladi*. Some characteristic uses of these markers are given in (4)-(6) below:

(4) *Kai tote, en ekeini tin imera,* ***igoun*** *eis tin Deuteran Parousian…*
"And then, on that day, ***that is****, on the Second Coming…"*
(Resinos, National Library of Greece, ms 639 (autograph), 16th c.)

(5) *Kai ekaman tin Ierousalim ta ethni touta,* ***dilonoti*** *tin Konstantinoupolin, i opoia eklithi nea Ierousalim, os oporofylakeion…*
"And these nations made Jerusalem, ***that is*** *Constantinople – which has been called new Jerusalem – a place to keep fruit…"*
(anonymous, Patriarchal Library of Alexandria, ms 97, 16th c.)

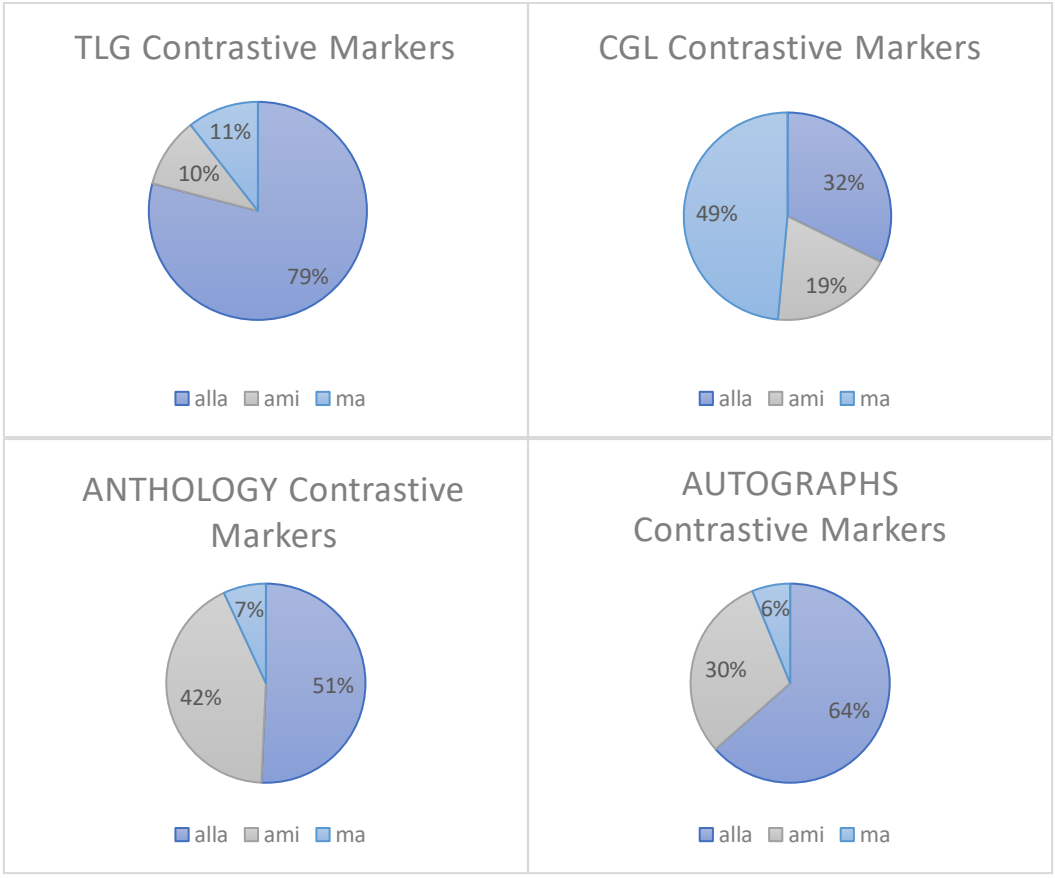
(6) *kai mas ermineuoun tas technas tautas kai epistimas oi presviteroi,* ***diladi*** *oi apostoloi kai oi mathites tou Christou…*
"and these arts and sciences are explained to us by the presbyters, ***that is****, the apostles and the disciples of Christ…"*
(Rartouros, ed. Venice 1560)

5 Results

5.1 Contrastive markers

The first question that we investigated is the distribution of contrastive markers in EMG corpora. Our results show that lexical choices and frequency of use in the examined contrastive markers is remarkably

varied in the four corpora. A striking finding is the dominance of *alla* and the low representation of *ami* in the TLG corpus, amounting to 80% and 10% respectively. Although very frequent in all four corpora, *alla* does not appear as dominant in any other corpus (CGL: 32%, Anthology: 51%, Autographs: 64%). Similarly, in none of the other corpora is *ami* as underrepresented (CGL: 19%, Anthology: 42%, Autographs 30%). The CGL corpus also exhibits striking peculiarities in the use of contrastive markers. It is the only corpus in which *ma*, and not *alla* is the most frequent contrastive marker (49%) and it is also the only corpus in which *alla* covers less than one third of the overall frequency of contrastive markers (32%). Figures 1-4 illustrate these points in detail.



Figures 1-4: The distribution of contrastive markers in EMG corpora

Another interesting finding of our investigation is the similarity observed between Anthology and Autographs. In both corpora, *ma* shows an equally low frequency (Anthology 7%, Autographs 6%), and both corpora exhibit the same hierarchy in the use of the examined markers contrastive markers (*alla* > *ami* > *ma*). Nevertheless, while in Anthology the frequencies of *ami* and *alla* are somewhat balanced (51% and 42% respectively), in Autographs the use of *alla* is twice more frequent than the use of *ami* (64% and 30% respectively). These similarities become more apparent when we compare the contrastive markers' absolute frequencies in the examined corpora. Figure 5 illustrates this comparison.

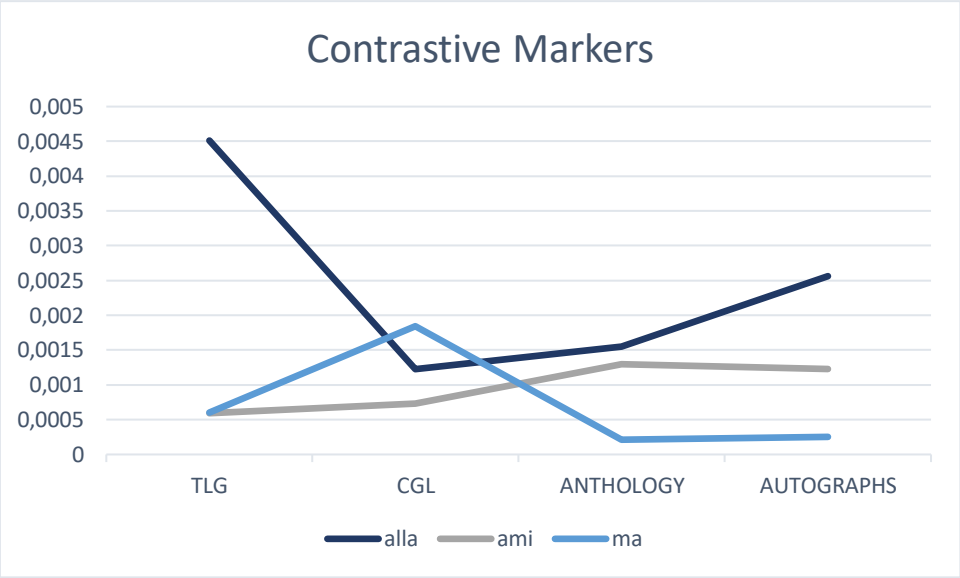


Figure 5: Absolute frequencies of contrastive markers in EMG corpora

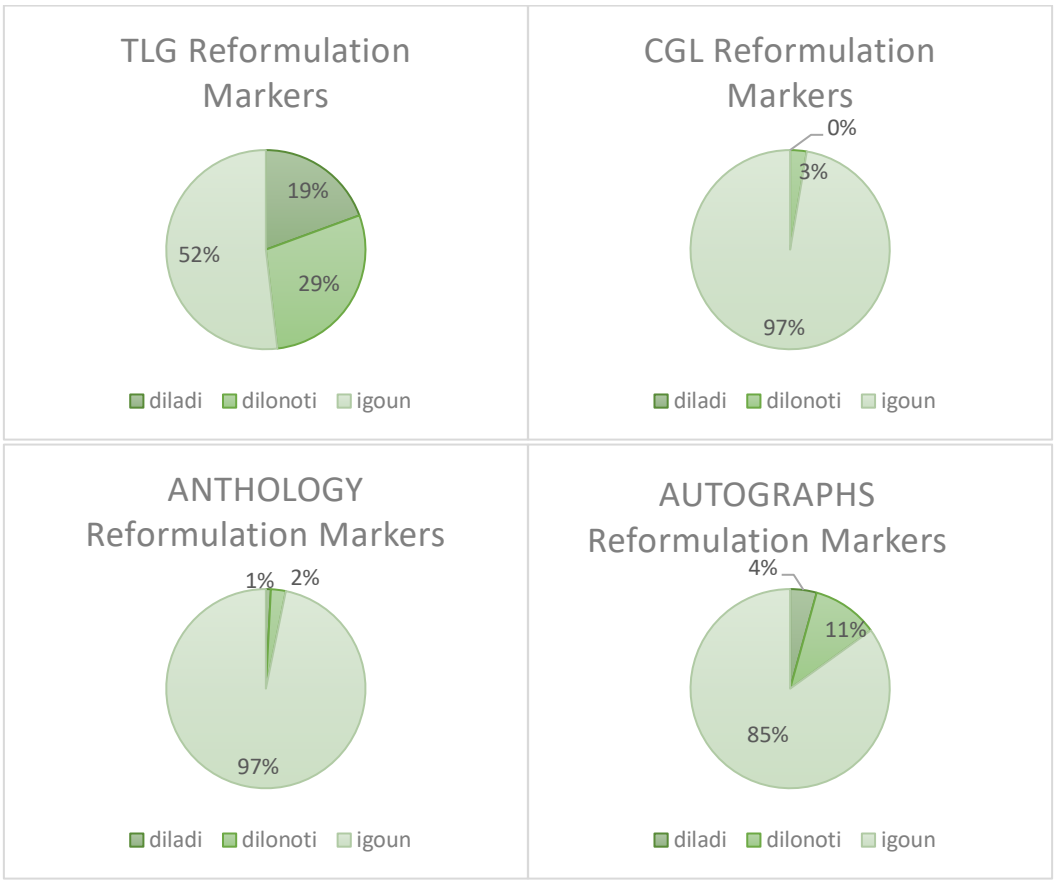
5.2 Reformulation markers

The second question that we investigated is the distribution of reformulation markers in the corpora under discussion. All EMG corpora revealed a common pattern concerning the order of preference in the use of reformulation markers. As concerns relative frequencies, CGL, Anthology, and Autographs are very much alike, while TLG illustrates a unique pattern. Our search in TLG, CGL, Anthology, and Autographs showed that *igoun* is the predominant reformulation marker in Early Modern Greek. In both the CGL and the Anthology corpora *igoun* is used in 97% of linguistically marked reformulations, while in the Autographs the respective proportion amounts to 85%. This preference is confirmed in the TLG corpus too, although *igoun* in TLG is used only in 52% of marked reformulations. In all examined corpora, the second most frequent reformulation marker is *dilonoti*, which covers though only a small proportion of reformulations (3% in CGL, 2% in the Anthology and 11% in the Autographs). TLG represents a different situation, with *dilonoti* covering as much as 29% of marked reformulations. Probably the most surprising finding of our investigation of reformulations is the frequency of *diladi* – the predominant, if not the only, reformulation marker in contemporary Modern Greek – which covers only 1% of marked reformulations in the Anthology, 4% in the Autographs, and is unattested in CGL. Again, findings from TLG differ significantly, with *diladi* appearing in 19% of reformulations. Figures 6-9 below illustrate in detail the patterns observed in the examined corpora.

As in the case of contrastive markers, the comparative examination of absolute frequencies shows that there are two main tendencies in the use of reformulation markers. On one hand, TLG and CGL show a very low frequency of use for all reformulation markers, while on the other, Anthology and Autographs show a low frequency of use for *diladi* and *dilonoti*, but a much higher frequency for *igoun* (see Figure 10).

The cross-examination of our findings reveals that the corpora under investigation share some basic patterns concerning the appearance of contrastive and reformulation markers, but they also show some significant divergences. In our searches, TLG exhibits tendencies in lexical preferences that are not found in the other three corpora. CGL also shows unique patterns, especially in what concerns the expression of contrast. On the other hand, Anthology and Autographs appear to be more alike, both in the absolute frequencies that they exhibit, as well as in the expression of contrast and reformulation. The similarities between Anthology and Autographs could be assigned to the design of these corpora. They both include only prose texts, they cover shorter temporal spans, and they extend to a wider range of genres compared to TLG and CGL. In order to scrutinize the association between the design of EMG

corpora and the patterns that they exhibit, in the next section, we examine the impact of certain corpora characteristics on the frequency of contrastive and reformulation markers.



Figures 6-9: The distribution or reformulation markers in EMG corpora

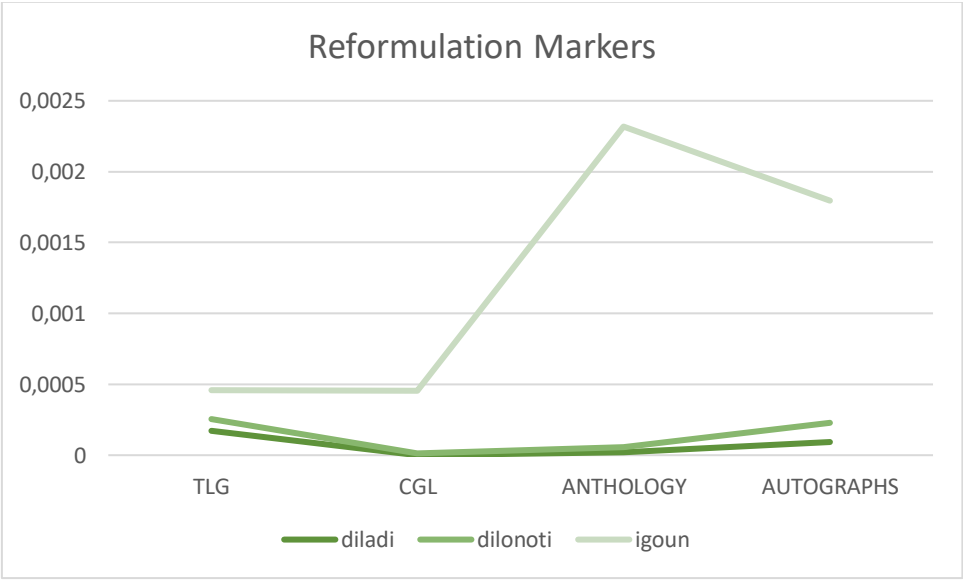


Figure 10: Absolute frequencies of reformulation markers in EMG corpora

6 Discussion

An essential difference between Anthology and Autographs on one hand, and TLG and CGL on the other, is that only the latter contain poetic works, and in fact, a significant number thereof. Considering that the inclusion of poetic texts in these corpora might suggest a possible cause for their differentiation compared to Anthology and Autographs, we tried to investigate the impact of the poetry/prose variable on the distribution of discourse markers. Focusing on the CGL corpus, we investigated the appearance of reformulation markers in poetic and prose works. Our findings (illustrated below) suggest that reformulation markers appear almost exclusively in prose texts and rarely in poetry. Given that poetic works amounts to 65% of the CGL corpus, we can assume that the low frequency of reformulation markers in CGL is due to the relatively low representation of prose in the overall corpus.

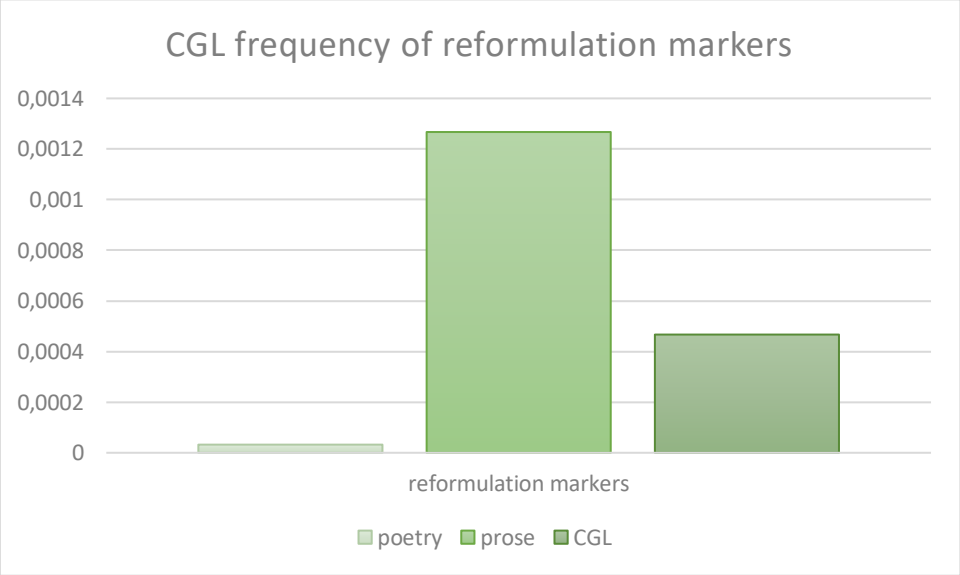


Figure 11: Reformulation markers in poetic and prose texts in CGL

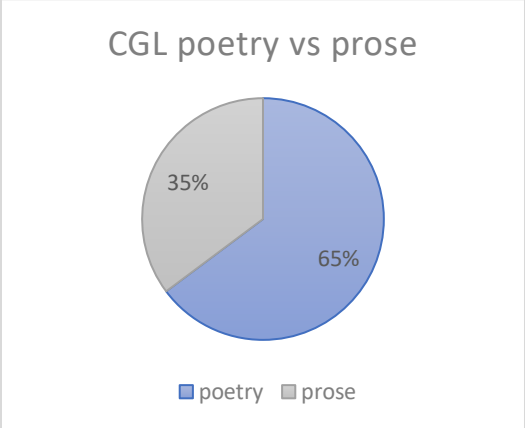


Figure 12: sampling of the poetry/prose variable in the CGL corpus

Another phenomenon with which we tried to investigate the impact of the poetry/prose variable is the use of *ma*. As we have seen in Section 5, *ma* appears rarely in the Anthology and the Autographs, more frequently in TLG, and is the predominant contrastive marker in CGL. Considering the characteristics of the four corpora, we hypothesized a possible association between the frequency of *ma*, and the poetry/prose variable. In order to investigate this association in conditions that neutralize the effects of other variables (such as period and geographical variation), we examined the appearance of *ma* only in the Cretan texts of TLG. Again, our results revealed a strong association of the phenomenon

with the poetry/prose variable, as *ma* in Cretan poetic texts is 5 times more frequent than it is in Cretan prose. This association, illustrated in Figure 13, suggests a possible explanation for the low frequency of *ma* in the Anthology and the Autographs, and its dominance in a largely poetic corpus, such as CGL.

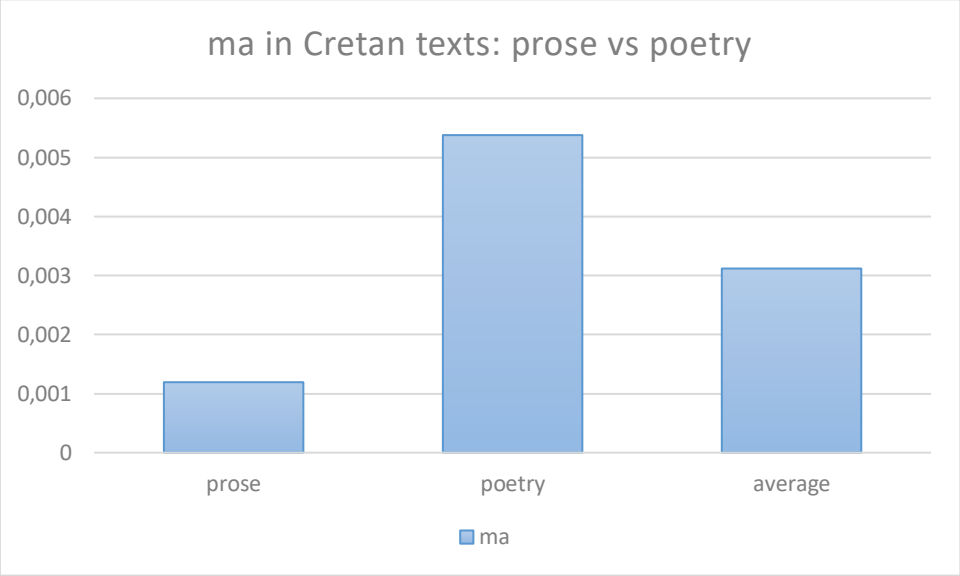
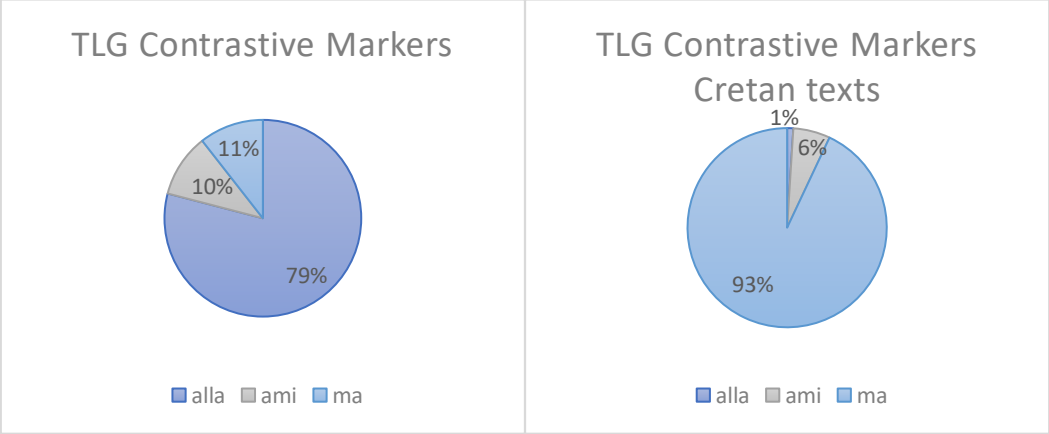


Figure 13: *ma* in prose and poetic Cretan texts in TLG

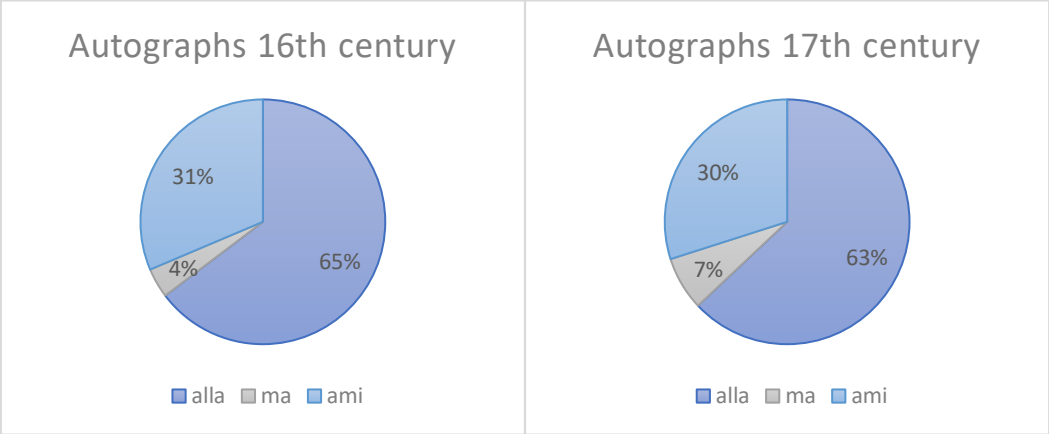
Our investigation of *ma* brought to our attention a possible association between this marker and Cretan texts. Trying to explore the connection of *ma* with the variable of geographical region, we compared the frequencies of contrastive markers in the Cretan texts of TLG with the frequencies of contrastive markers in the sum of TLG. This comparison revealed that *ma* in TLG Cretan texts covers 93% of the expressions of contrast, while the respective proportion for the overall TLG corpus is only 11%. We take this result as an indication for the impact of the geographical variable on the use of contrastive markers, and especially on the frequency of *ma*.



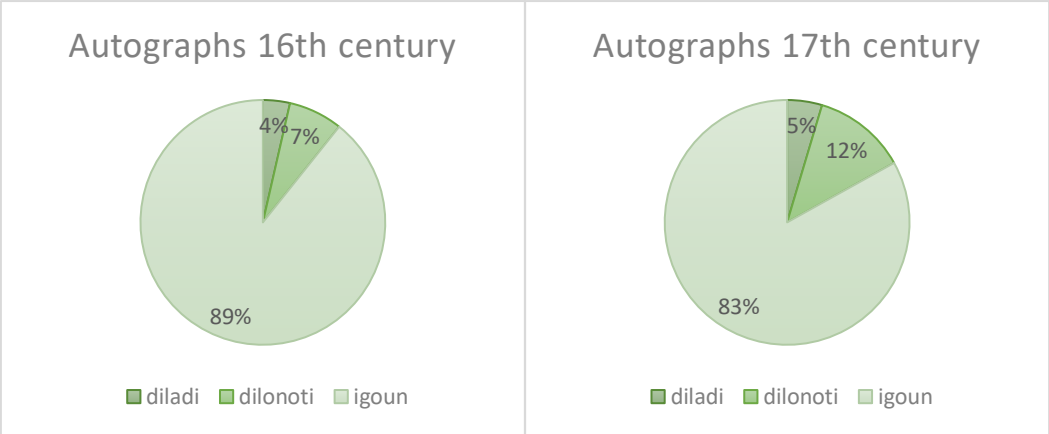
Figures 14-15: Contrastive markers in the sum of TLG and in TLG Cretan texts

As we have seen in Section 3, apart from other divergences, the examined corpora differ also in the periods that they cover. The question, thus, arises whether the period variable plays a role in the appearance of contrastive and reformulation markers in the examined corpora. Following this query, we first investigated the impact of the period variable on contrastive and reformulation markers in the Autographs, a corpus which covers 16th and 17th century and patterns similarly to the Anthology. Concerning the use of contrastive markers, our results from Autographs did not reveal a significant

difference in the expression of contrast or reformulation between 16th and 17th century. In what concerns the use of contrastive markers in the Autographs, the only divergence that we observed between 16th and 17th century texts is an increase in the use of *ma*, from a relative frequency of 4% in the 16th century to 7% in the 17th. Similarly, the expression of reformulation in the Autographs seems not to be affected by the period variable. The increase in the use of *dilonoti* – from 7% in 16th century texts to 12% in 17th century texts – is the only noticeable difference that the temporal variable brings to the expression of reformulation in the Autographs corpus.



Figures 16-17: Contrastive markers in Autographs in 16th and 17th century texts



Figures 18-19: Reformulation markers in Autographs in 16th and 17th century texts

The impact of the period variable becomes more apparent when investigated in larger temporal spans, like those covered by the CGL corpus. Contrary to the situation observed in the Autographs, our research on the distribution of contrastive markers in CGL revealed that lexical choices in the expression of contrast are largely dependent on the period in which a text belongs. Thus, *alla*, which is the dominant contrastive marker in the 12th century CGL texts, is rarely attested in 17th century, while *ma*, which scarcely appears in the 12th century, dominates the expressions of contrast in the 17th century. We consider these results to be indicative of an association between period and the expression of contrast.

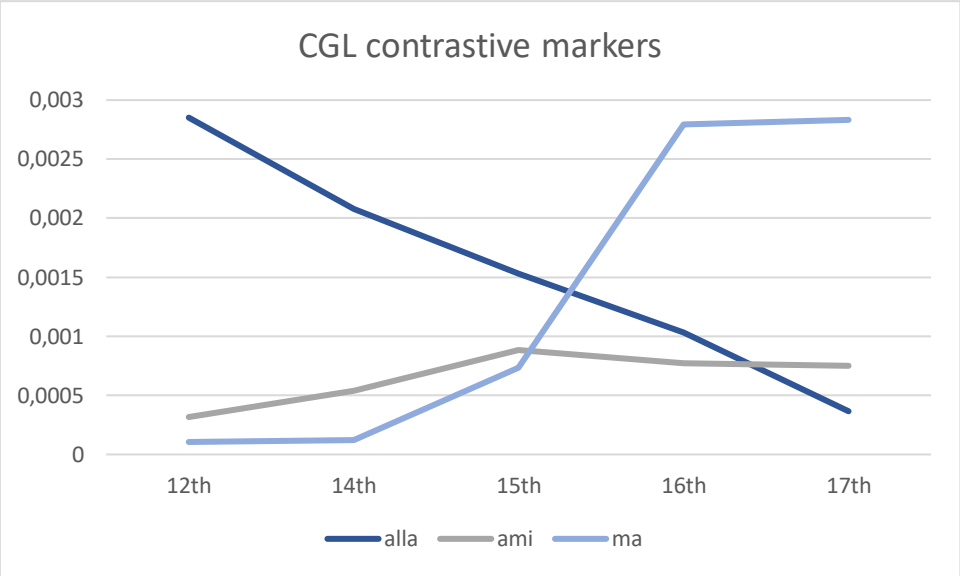


Figure 20: Absolute frequencies of contrastive markers in the periods covered by CGL

The last parameter that we examined in our study is that of the text type or genre. Having already observed that the poetry/prose variable plays in important role in determining the choice and the frequency of discourse markers, we tried to explore how the thematic orientation of a text affects the frequencies of contrastive and reformulation markers. The best candidate for a test of this sort is the Anthology corpus which is structured according to 11 thematic categories: *Forewords*, *Theology*, *Sermons*, *Lives of Saints*, *Philological texts*, *History-Chronicles*, *Geography-Travel Literature*, *Sciences*, *Legal texts*, *Notary books*, and *Correspondence*. Our investigation into the Anthology texts showed that both the choice and the frequency of contrastive markers depends on the thematic orientation of a text. For instance, genres like *Forewords*, *Sermons*, *History*, and *Lives of Saints* (Saints biographies), make extensive use of contrastive markers, while genres like notary documents (*Notary Books*), *Sciences* and *Travel Literature* show low frequencies in the use of contrastive markers. Interestingly, beyond an overall pattern concerning the order of preference of contrastive markers, some text types exhibit unique associations with the use of particular contrastive markers. In *Travel Literature*, for instance, *ma* is the predominant contrastive marker and *ami* is unattested, while in *Correspondence* *alla*, *ami* and *ma* share the same proportion in the expression of contrast (absolute frequency: 0,0009). On the other hand, *Philological* texts exhibit a very strong dominance of *alla*, and *History-Chronicles* are remarkably associated with the use of *ami* (compare Figure 21). These findings suggest that the text type variable plays an important role for the expression of contrast in EMG.

Extending our investigation into the expression of reformulation, we next examined the impact of the text type/genre variable on the distribution of reformulation markers. Contrary to our findings on contrastive markers, in the expression of reformulation lexical choices do not seem to be associated with the text type variable. However, our result showed that the frequency of marked reformulations is largely dependent on the thematic orientation of a text. *Anthology*, *Science*, *Travel Literature*, *Philological texts*, and *Theology* show high frequencies in the use of reformulation markers, while *Correspondence*, *Forewords*, *Lives of Saints*, and *Chronicles* exhibit a rather scarce use of reformulation markers (see Figure 22).

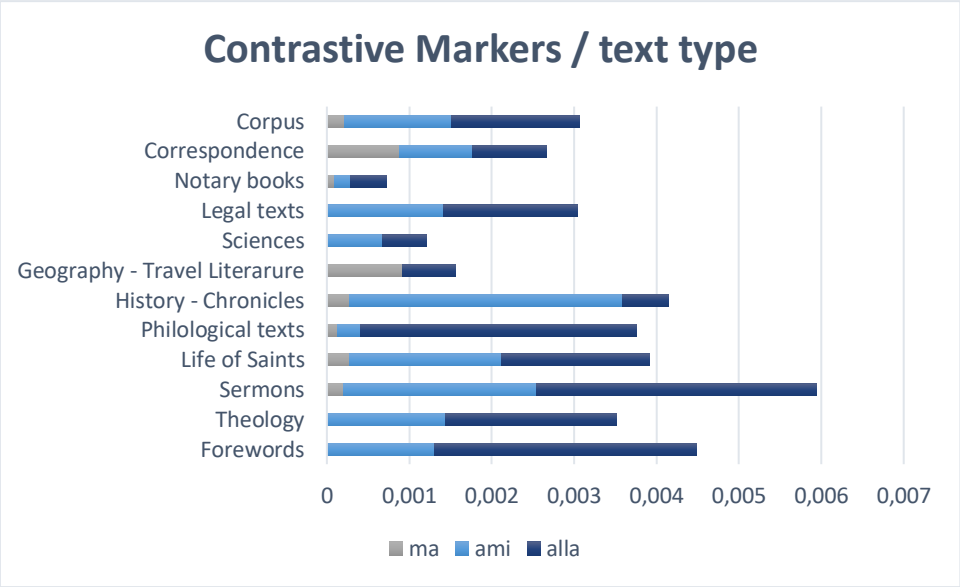


Figure 21: Contrastive markers in the Anthology according to the text type variable

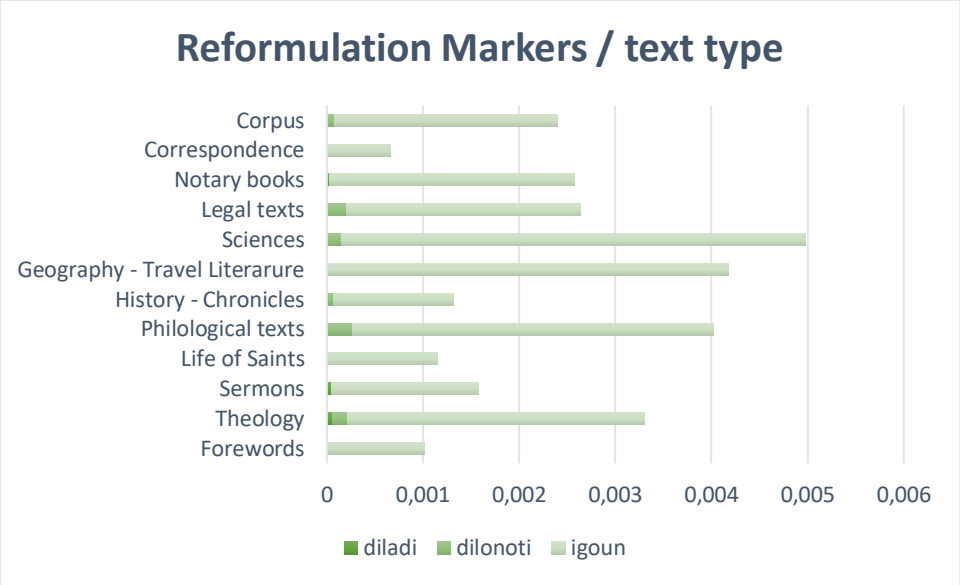


Figure 22: Reformulation markers in the Anthology according to the text type variable

Beyond the level of empirical evidence, the association of certain categories of discourse markers with certain text types has theoretical motivation, too. In principle, argumentative and narrative texts are very likely to include discourse relations, such as contrast or denial of expectations. Similarly, texts introducing terminology or exposing new information are likely to include rephrasings, definitions, or metalinguistic alternatives, whose expression relies heavily on the use of reformulation markers[6]. On this view, it is rather normal to find that science texts in EMG make excessive use of reformulation markers and scarce use of contrastive markers or that reformulation markers in chronicles are far less frequent than contrastive markers. The point is that corpora not taking into consideration the text type variable fail to capture the full spectrum of a language's lexical manifestations.

[6] The relation between contrast and argumentation has been underlined in Anscombre & Ducrot 1977, and Ducrot & Vogt 1979.

7 Conclusions

The extensive variability in the use of contrastive and reformulation markers evidenced in this paper confirms the view that Early Modern Greek is a very dynamic synchrony, in which linguistic alternatives compete with each other and form grammatical paradigms that are rather fluid. This instability of the linguistic system makes the description of EMG risky and calls for the creation of representative corpora of the period. The fact that EMG is a variety to which we have access only through written documents is a further indication of the need for reliable, representative corpora.

In this paper, we investigated the representativeness of four existing corpora on EMG, namely TLG, CGL, Anthology and Autographs. Focusing on linguistic representativeness, and adopting an empirical methodology, we explored how contrastive and reformulation markers are represented in the aforementioned corpora. According to our hypothesis, if EMG corpora are representative of the language which they intend to document, then the illustrated quantitative data should be similar across corpora. On the contrary, if EMG corpora exhibit different linguistic patterns, then their linguistic representativeness should be considered to be restricted and dependent on the characteristics of their design.

The results of our study showed that EMG corpora illustrate different situations concerning the distribution of contrastive and reformulation markers. Following our hypothesis, we consider this variation to be indicative of variability in the degree of representativeness of EMG corpora, which can only be assigned to divergences in EMG corpora design. As we have seen, EMG corpora differ according to a number of aspects, including their length, the period that they cover, the forms that they include (i.e., poetry/prose), the geographical distribution of the illustrated texts, and the genres they represent. Leaving aside the effect of the size parameter, which has not been examined in this paper, the results of our research showed that all investigated variables play an important role in the choice and frequency of contrastive and reformulation markers.

The poetry/prose variable seems to have a considerable impact on the representativeness of EMG corpora. Our investigation of CGL showed that reformulation markers appear rarely in poetry, while our results from TLG revealed that choices and frequencies of certain contrastive markers, such as *ma*, heavily depend on whether a given texts belongs to poetry or prose. Geographical variation also shows a strong effect on the use of discourse markers. In the TGL corpus, the expression of contrast in Cretan texts differs radically from the expression of contrast observed in texts from other Greek speaking regions. The genre variable appears to be the most important one, both for the frequency of the markers used in a text, and for the preference order in which the members of a category appear. Expository or "pedagogical" texts, such as science and travel literature, tend to exhibit higher frequencies of reformulation markers and lower frequencies of contrastive markers. On the contrary, narratives or argumentative texts, such as chronicles and sermons, tend to include more contrastive markers and fewer reformulative ones. Concerning the period variable, its impact on the linguistic representatives seems to depend on the extend of the temporal spans involved. In Autographs, for instance, a corpus covering a temporal span of two centuries (16^{th}-17^{th}), period does not seem to have an impact on the use of either contrastive or reformulation markers. On the other hand, in CGL, a corpus covering six centuries (12^{th}-17^{th}), period has a more significant role in the examined phenomena, especially in terms of the expression of contrast.

An important issue that we have not addressed in our investigation is the effect of size in EMG corpora representativeness. As we have seen, EMG corpora differ in their size, ranging from very small ones (e.g., Autographs) to relatively large ones (e.g., TLG). Although we acknowledge that large corpora are statistically more likely to cover the variability of the targeted population, our results seem to imply that increased corpus size does not necessarily entail increased linguistic representativeness. In our study, CGL – a corpus which is five times larger than the Autographs – illustrates tendencies that are

not confirmed by any other EMG corpus. Considering that CGL is a largely poetic corpus, covering a period that goes beyond EMG, we assume that this peculiarity is an indication of low representativeness which should be assigned to flaws in CGL's sampling. Similarly, the fact that TLG – a relatively large collection – includes a small number of texts, from a small range of text types, does not seem to be irrelevant for the unique linguistic patterns observed in this corpus. The point is that size alone cannot guarantee representativeness. In what concerns EMG corpora, our study indicates that linguistic representativeness is straightforwardly connected with variables, such as period, form, geographical region, and text type. Future investigations, as well as future EMG corpora, should address these associations. Other factors not examined in our study, such as author identity, the text tradition (autographs, manuscripts, editions), and the degree of the author's creativity (original work vs translation), are very likely also to affect EMG corpora representativeness and deserve scrutinization in future studies.

References

Anscombre, Jean Claude & Oswald Ducrot. 1977. Deux 'mais' en français? *Lingua* 43. 23–40.

Auroux, Sylvain. 1994. *La révolution technologique de la grammatisation*. Liège: Mardaga.

Baggioni, Daniel. 1997. *Langues et nations en Europe.* Paris: Payot.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243-257.

Blakemore, Diane. 1987. *Semantic constraints on relevance*. Oxford: Blackwell.

Blakemore, Diane. 1989. Denial and contrast: a relevance theoretic analysis of 'but'. *Linguistics and Philosophy* 12. 15–37.

Blakemore, Diane. 1993. The relevance of reformulations. *Language and Literature* 2(2). 101–20.

Blakemore, Diane. 1994. Relevance, poetic effects and social goals: a reply to Culpeper. *Language and Literature* 3(1). 49–59.

Blakemore, Diane. 1997. Restatement and exemplification: a relevance theoretic re-assessment of elaboration. *Pragmatics and Cognition* 5(1). 1–19.

Blakemore, Diane. 2002. *Relevance and Linguistic Meaning.* Cambridge: Cambridge University Press.

Blakemore, Diane. 2007. 'Or'-Parentheticals, 'That is'-Parentheticals and the Pragmatics of Reformulation. *Journal of Linguistics* 43. 311-339.

Brinton, Laurel J. 1996. *Pragmatic Markers in English: Grammaticalization and Discourse Functions.* Berlin: Mouton de Gruyter.

Burke, Peter. 2004. *Languages and Communities in Early Modern Europe.* Cambridge: Cambridge University Press.

Conceiçao, Manuel Celio. 2005. *Concepts Termes et Reformulations.* Lyon: Presses universitaires de Lyon.

Culpeper, Jonathan. 1994. *Why relevance theory* does not explain 'The relevance of reformulations'. *Language and Literature* 3(1). 43-48.

Ducrot, Oswald & Carlos A. Vogt. 1979. De magis à mais : une hypothèse sémantique. *Revue de linguistique romane* 171-172. 317-341.

Fraser, Bruce. 1996. Pragmatic Markers. *Pragmatics* 6. 167-190.

Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31. 931-952.

Furko, Peter. 2014. Perspectives on the Translation of Discourse Markers. *Acta Universitatis Sapientiae, Philologica* 6(2). 181–196.

Goutsos, Dionysis. 2010. The corpus of Greek texts: A reference corpus for Modern Greek. *Corpora*, 5(1). 29–44.

Gray, Bethany, Jesse Egbert & Douglas Biber. 2017. Exploring methods for evaluating corpus representativeness. Paper presented at the Corpus Linguistics International Conference 2017. University of Birmingham, 24-28 July.

Gülich, Elisabeth. & Thomas Kotschi. 1983. Les marqueurs de la reformulation paraphrastique. *Cahiers de Linguistique Française* 5. 305-351.

Hanks, Patrick. 2012. The corpus revolution in lexicography. *International Journal of Lexicography* 25(4). 398–436.

Hatzigeorgiou, Nikos, Athanasia Spiliotopoulou, Anna Vakalopoulou, Anastasia Papakostopoulou, Stelios Piperidis, Maria Gavriilidou and Giorgos Karayannis. 2001. Ethnikos thisavros ellinikon keimenon: Soma keimenon tis Neas Ellinikis sto diadiktyo [National thesaurus of Greek Texts: a corpus of Modem Greek on the internet]. *Studies in Greek Linguistics* 21. 812-821. (In Greek)

Hinterberger, Martin. 2006. How Should We Define Vernacular Literature? Paper given at the congress: Unlocking the Potential of Texts: Interdisciplinary Perspectives on Medieval Greek, University of Cambridge, 18-19 July.
Holton, David & Io Manolessou. 2010. Medieval and Early Modern Greek". In E. Bakker (ed.) *A Companion to the Ancient Greek Language*, 539-563. Oxford: Wiley/ Blackwell.
Holton, David, Geoffrey Horrocks, Marjolijne Janssen, Tina Lendari, Io Manolessou & Notis Toufexis. 2019. *The Cambridge Grammar of Medieval and Early Modern Greek*. Cambridge: Cambridge University Press.
Kakoulidou-Panou, Eleni, Eleni Karantzola & Katerina Tiktopoulou. in press. Anthologio dimodous pezou logou tou 16ου aiona [Demotic prose texts of the 16th century]. Thessaloniki & Athens: Centre for Greek Language & MIET. (in Greek)
Karantzola, Eleni. & Alexis Kalokerinos. 2005. Discourse markers of opposition in Early Modern Greek. In Jeffreys, Elizabeth & Michael Jeffreys (eds.) *Approaches to Texts in Early Modern Greek, Proceedings of the Neograeca Medii Aevi V*, 179-196. Oxford: University of Oxford.
Lakoff, Robin. 1971. If's, and's, and but's about conjunction. In Fillmore, Charles. J. & D.Terence Langendoen (eds.), *Studies in linguistic semantics*, 115–150. New York: Holt, Rinehart & Wilson.
Mauri, Caterina. 2008. *Coordination Relations in the Languages of Europe*. Berlin & New York: Mouton de Gruyter.
McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London and New York: Routledge.
Meyer, Charles. F. 1992. *Apposition in Contemporary English*. Cambridge: Cambridge University Press.
Papaioannou, Anastasios. 2016. *Peza Keimena Logion tou 16ou kai 17ou Aiona se Dimodi Elliniki Glossa, Vasei Aftografon Xeirografon Kodikon*. [Vernacular Greek prose texts of 16th and 17th century scholars, as preserved in autograph manuscripts]. PhD dissertation, University of the Aegean. (in Greek)
Raineri, Sophie & Camille Debras. 2019. Corpora and Representativeness: Where to go from now? *CogniTextes* 19. Retrieved from https://journals.openedition.org/cognitextes/1311 (accessed 20 May 2020)
Rossari, Corinne. 1994. *Les opérations de reformulation: analyse du processus et des marques dans une perspective contrastive français – italien*. Berne: Lang.
Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.
Schourup, Lawrence. 1999. Discourse markers. *Lingua* 107. 227-265.
Sinclair, John. 1996. EAGLES preliminary recommendation on text typology. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1988&rep=rep1&type=pdf (accessed 11 June 2020).
Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3). 538–556.
Steuckardt, Agnès. 2009. Décrire la reformulation: le paramètre rhétorique. *Cahiers de praxématique* 52. 159-172.