ΜΕΛΕΤΕΣ ΓΙΑ ΤΙΣ ΝΕΟΕΛΛΗΝΙΚΕΣ ΔΙΑΛΕΚΤΟΥΣ ΚΑΙ ΤΗ ΓΛΩΣΣΟΛΟΓΙΚΗ ΘΕΩΡΙΑ

Επιστημονική Επιμέλεια:
Mark Janse, Brian Joseph, Παύλος Παύλου, Αγγελική Ράλλη και Σπύρος Αρμοστή
Φιλολογική Επιμέλεια:
Γιώργος Β. Γεωργίου

STUDIES IN MODERN GREEK DIALECTS AND LINGUISTIC THEORY

Editors:
Mark Janse, Brian Joseph, Pavlos Pavlou, Angela Ralli and Spyros Armosti
Volume Editor:
Giorgos V. Georgiou

ΚΕΝΤΡΟ ΜΕΛΕΤΩΝ ΙΕΡΑΣ ΜΟΝΗΣ ΚΥΚΚΟΥ
ΛΕΥΚΩΣΙΑ 2011

RESEARCH CENTRE OF KYKKOS MONASTERY
NICOSIA 2011

ISBN 978-9963-645-52-7

Τ.Θ. 28192, 2093 ΛΕΥΚΩΣΙΑ, ΚΥΠΡΟΣ

Corpus linguistics in historical dialectology: a case study of Cypriot

Io Manolessou & Notis Toufexis
University of Patras & University of Cambridge

> I am the very model of a user of technology
> For testing out hypotheses on grammar and morphology
> I used to do it manually, with diagrams arboreal
> But life is so much better since my research went corporeal.
>
> A language looks quite different when processed electronically
> My lab has all the software to describe it diachronically
> I have a suite of programs which equips me with facilities
> For tagging and for parsing and computing probabilities.
>
> S. Blackwell (2006). In A. Renouf & A. Kehoe (Eds.) *The changing face of corpus linguistics* (p. 1). Amsterdam: Rodopi.

Abstract

The present paper gives an overview of the branch of corpus linguistics that deals with historical corpora, i.e. electronic text compilations of of past forms of language, and discusses their applicability and availability for the study of the history of the Greek language. The methodology for constructing a historical corpus of the Cypriot dialect (Corpus of Medieval Cypriot Texts, CMCT) is presented, with discussion criteria for text inclusion and of modelling and implementation issues (mark-up languages, metadata, digital transcription methods).

1. The framework

1.1. Historical and dialectal corpora

The lyrics which open our present discussion reflect the dynamism and optimism of the scientific branch of corpus linguistics. Of particular interest for us is the expansion of corpus linguistics in the diachronic domain: for most European languages there exist corpora of varying size and degree of annotation which cover extended periods of time, and aim to document their diachronic evolution. The *terminus technicus* for this type of text compilations is *historical corpora*. English, as expected, is the most well documented language, thanks mainly to the Helsinki

In M. Janse, B. Joseph, P. Pavlou, A. Ralli, & S. Armosti (Eds.). (2011). *Studies in Modern Greek Dialects and Linguistic Theory* (pp. 305–320). Nicosia: Research Centre of Kykkos Monastery.

Corpus and its various further elaborations and expansions.[1] However, as already mentioned, historical corpora already exist, or are in the process of compilation, for most European languages, including smaller ones such as Welsh or Czech: most of them are set out in table 1 below.[2]

Language	Historical Corpus	Web address
English	*Helsinki Corpus of English Texts*	http://icame.uib.no/hc
	Innsbruck Computer Archive of Machine-Readable English Texts (ICAMET)	http://www.uibk.ac.at/ anglistik/projects/icamet/
	A Representative Corpus of Historical English Registers (ARCHER)	http://www.llc.manchester.ac.uk/ research/projects/archer/archer3_1/
	Penn Parsed Corpora of Historical English	http://www.ling.upenn.edu/hist-corpora/
	The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)	http://www-users.york.ac.uk/ ~lang22/YcoeHome.htm
German	*Deutsch Diachron Digital (DDD)*	http://www.deutschdiachrondigital.de/
	Bonner Frühneuhochdeutschkorpus	http://www.korpora.org/Fnhd/
	Bochumer Mittelhochdeutsch Korpus	http://www.ruhr-uni-bochum.de/ wegera/archiv_1.htm
	Digitales Mittelhochdeutsches Textarchiv	http://mhgta.uni-trier.de/
Dutch	*Integrated Language Database of 8th-21st-Century Dutch (ILD)*	http://www.inl.nl
French	*Consortium pour les corpus de français medieval (CCFM)*	http://ccfm.ens-lsh.fr/
Italian	*Tesoro della Lingua Italiana delle Origini (TLIO)*	http://www.ovi.cnr.it/
Spanish	*Corpus Diacrónico del Español (CORDE)*	http://corpus.rae.es/cordenet.html
	Corpus del Español	http://www.corpusdelespanol.org/
Portuguese	*Tycho Brahe Parsed Corpus of Historical Portuguese*	http://www.tycho.iel.unicamp.br/ ~tycho/corpus/en/
	Corpus informatizado do Portugues Medieval (CIPM)	http://cipm.fcsh.unl.pt/

[1] On English historical corpora see mainly Rissanen (2000) and the links available at http://tiny.cc/corpora (Section "Historical Corpora").
[2] A short description of these, up to 2001, is available in Decorte (2003). For Romance languages, more details are available in Pusch, Kabatek & Raible (2005).

Czech	*Bank of diachronic Czech (ČNKDIA)*	http://ucnk.ff.cuni.cz/english/diakorp.php
Bulgarian / OCS	*Sofia-Trondheim Corpus of Old Bulgarian*	http://www.hf.ntnu.no/SofiaTrondheimCorpus/
	Corpus Cyrillo-methodianum helsingiense	http://www.slav.helsinki.fi/ccmh/
Welsh	*Historical corpus of the Welsh language*	http://people.pwf.cam.ac.uk/dwew2/hcwl/menu.htm
Old Scandinavian	*Medieval Nordic Text Archive (MENOTA)*	http://www.menota.org/
	Studér Middelalder på Nettet (Danish Medieval Texts)	http://smn.dsl.dk/index.php
	Fornsvenska textbanken (Old Swedish Textbank)	http://www.nordlund.lu.se/Fornsvenska/Fsv%20Folder/

Table 1. Major historical corpora

Historical corpora are also expanding in the domain of dialectology, again with English at the forefront:

English	*Helsinki corpus of Older Scots (1450–1700)*	http://icame.uib.no/olderscotseks.html
	A corpus of Irish English	http://www.uni-due.de/CP/CIE.htm
Bulgarian	*An Electronic Archive of the Bulgarian Dialects*	http://www.bultreebank.org/veda/indexeng.html
Scandinavian	*Nordic Dialect Corpus*	*http://www.tekstlab.uio.no/nota/scandiasyn/*

Table 2. Major dialectal corpora

The Scottish and Irish varieties of English possess corpora of their own (see table above), and many of the above mentioned corpora of other languages contain dialectal data. The aim of the present paper is to discuss the status of Greek within the context of historical and dialectal corpus compilation and to give a concrete example on the basis of the Medieval Cypriot dialect.

Let us begin first by comparing the resources discussed above to those available for Greek. To begin with, the Greek case is hardly comparable to the above languages, because a) its language history is much longer b) its earlier stages are much better documented c) it has a double tradition, learned and non-learned and d) its earlier stages have been the object of international scholarly research for decades if not centuries. As a result, the electronic data coverage for the earlier phases of Greek is very good, thanks to the efforts of large and well-funded teams of (mainly non-Greek) scholars. Table 3, below, gives an overview.

Period	Historical corpus	Web address
Classical	*Perseus Digital Library* *Thesaurus Linguae Graecae (TLG)*	www.perseus.tufts.edu/hopper
Ancient Greek dialects	*Searchable Greek Inscriptions*	http://epigraphy.packhum.org/inscriptions/
Hellenistic-Imperial (learned)	*Thesaurus Linguae Graecae*	www.tlg.uci.edu
Hellenistic-Imperial (non-l.)	*Duke Databank of Documentary Papyri (DDBDP)*	http://www.papyri.info/navigator/ddbdpsearch
Medieval Greek (learned)	*Thesaurus Linguae Graecae* *Οι δρόμοι της πίστης-Ψηφιακή Πατρολογία*	www.tlg.uci.edu http://patrologia.ct.aegean.gr/patrologia.htm
Medieval Greek (non-learned)	*Thesaurus Linguae Graecae*	www.tlg.uci.edu
Modern Greek	*Εθνικός Θησαυρός Ελληνικής Γλώσσας (ΕΘΕΓ)*	http://hnc.ilsp.gr
	Σώμα Ελληνικών Κειμένων	http://www.sek.edu.gr

Table 3. Greek electronic corpora

The Greek situation, as described above, is an almost ideal one for the student of Greek literature or history. But is it so for the student of historical linguistics? The answer is, unfortunately quite emphatically, no. The two reasons for this rather disappointing answer are the following:

a) **Coverage**: Whereas for the earlier periods of Greek and the learned literature (Classical, Hellenistic, Byzantine) the coverage is full, i.e. all the extant works exist in electronic form in the totality of their length, later Greek, as represented by vernacular Medieval literature, exists in such form only in unsatisfactory quantity. In 2009 and 2010 a number of Medieval and Early Modern Greek vernacular texts (including among other have been added to the TLG.[3] The publicly available electronic representation of non-literary medieval and early modern texts (legal documents, private letters etc.) is very limited.[4] Zero is also the representation of the vernacular literature roughly from 1453 up to our days (Cretan Renaissance, Enlightenment, 19th c.). This is a considerable gap of several centuries, as the Hellenic National Corpus (ΕΘΕΓ) and the Corpus of Greek texts contain texts only from contemporary Greek, i.e., from 1990 onwards (see e.g. Goutsos, 2003). Finally, the representation of Modern Greek dialects, whether medieval or modern, is also

[3] See http://www.tlg.uci.edu/authors/post_tlg_e.php for details. For more details on the structure and history of the TLG see Pantelia (2003). For a constantly updated list of further vernacular Medieval and Early Modern texts available in electronic form, see http://www.early-modern-greek.org/archives/28.

[4] For example, several volumes of the medieval monastery archives of Athos (Archive de l' Athos) are included in the online TLG. A small number of medieval texts is available in a CD-ROM produced by the institution Θησαυρός της ελληνικής Γλώσσας for the purposes of the Frankfurt Book Fair 2001. More texts are provided by the same institution after subscription, at www.thesavros.gr

zero (excepting again, the projected CGT, which will contain 10% of Cypriot material).

b) **Corpus construction**: corpora for linguistic purposes are constructed following strict guidelines which ensure i) the homogeneity and reliability of the text ii) the representativeness of the sample and iii) the ease of performance of linguistic searches thanks to grammatical and syntactic annotation. These guidelines are not observed for the most part in the Greek case, as the initiative behind the creation of the Greek corpora, with the sole exception of the Modern Greek corpus, was not linguistic, and no corpus linguists were involved in their construction.

For the rest of our discussion, we will leave aside Ancient and learned Byzantine Greek, and concentrate on Medieval and later Greek and its dialects, which forms the focus of the present conference. We will only note that almost no Greek subcorpus of any period possesses any kind of linguistic annotation,[5] and all were constructed using heterogeneous critical editions (with the exception of the inscriptional and papyrical corpora, which are by definition diplomatic).

1.2. Medieval Greek linguistics and Medieval Greek dialectology

We begin therefore with the assumption that for the investigation of the history of later Greek (Medieval and Early Modern), and therefore inevitably for the research on the historical dialectology of Modern Greek, the available corpus resources are very limited. To compound the problem, in comparison with other European languages Greek is also lacking sophisticated reference works: there is no grammatical description of Medieval and Early Modern Greek and there are no atlases recording the geographical distribution of linguistic features in any period (with the exception of Crete). Furthermore, historical dialectology as a branch of modern linguistics is non-existent in Greece, although making rapid progress abroad, precisely due to the lack of technological infrastructure which nowadays is a necessary presupposition.[6]

One recent positive development is the work conducted at the University of Cambridge towards an urgently needed *Grammar of Medieval Greek* (1100–1700).[7] While this descriptive Grammar cannot fill all gaps in our knowledge of the relevant period, it is hoped that it will provide the necessary groundwork for more detailed studies on the Medieval Greek dialects in the future. The Cambridge grammar is based on the most comprehensive to date collection of published textual sources

[5] The New Testament is available in electronic form with full Part of Speech Tagging (see the various available (commercial) editions at www.logos.com. Both Perseus and, until recently, the TLG offer mechanisms for on the spot morphological analysis (while browsing the text) and lemmatized search, respectively. It is not yet possible to conduct searches involving combination of text strings and part of speech annotation.

[6] For a discussion of this issue and an overview of historical dialectology, see Manolessou (2008).

[7] The aims, scope and methodology of the whole project are described in Holton (forthcoming), Lendari & Toufexis (forthcoming). Additional information is available at the project website, at www.mml.cam.ac.uk/greek/grammarofmedievalgreek

(both literary and non-literary) of this period; the descriptive framework of the Grammar takes heavily into account parameters of geographical distribution and variation and will offer a comprehensive account on all linguistic levels of analysis.

Of great assistance to such linguistic descriptive work is without question any kind of digital infrastructure; similar linguistic projects analysing the medieval period of other European languages rely heavily on electronic corpora and customised software for their exploration, on line thesauri, syntactic and morphological parsers or lexical databases and digitised lexica, for which, as mentioned in the beginning, there is a serious lack in Medieval Greek. This paper aims at providing a first feasibility study for the creation of such a specialized electronic corpus of Medieval Cypriot texts.

Our choice of the Medieval Cypriot dialect for this case study does not relate only to the venue of this conference; the main motivation behind it is the fact that it constitutes a closed, restricted corpus, for which one can achieve full representation. That is, the Medieval and Early Modern documents of Cypriot provenance are limited in number and size (in comparison, say, to Cretan ones), which means that they can be integrated in a corpus in their totality, not simply sampled. This corpus, furthermore, should be of sufficient size to be of practical use for linguistic research. Cypriot texts also cover a wide variety of genres (literary prose, literary verse, non-literary prose of various types). Additionally, Cypriot possesses distinct dialectal characteristics at all levels of analysis (phonology, morphology, syntax, lexicon) which can both be used as criteria of inclusion and as targets of research (we will come back to the inherent danger of circularity of this issue).

For the remainder of this paper we will discuss the particulars of a potential Corpus of Medieval and Early Modern Cypriot texts (CMCT).

2. A potential CMCT

2.1. Criteria for inclusion of texts to an electronic CMCT

One major difficulty with defining criteria for inclusion of texts in a potential electronic corpus of CMCT is exactly the definition of the term "Cypriot" in this context: It can refer both to the geographical region of Cyprus and the specific linguistic form spoken in this region. We will discuss briefly these two different categories in order to highlight some theoretical problems —this discussion is of course valid for the research on any historical dialectal form, not only Cypriot.

Cyprus as a geographical region is clearly identifiable from ancient times until today. We possess medieval texts that have been written on the island: our information is provided explicitly by authors and scribes or can be inferred from the content of the texts themselves. Many of these texts show systematic evidence of Cypriot dialectal features, as we know them from modern dialectological research, such as geminate consonants, strong palatalisation, loss of morphological genitive plural in masculine nouns, etc. The inclusion of such texts in the corpus poses no problems.

Several texts on the other hand, for which direct or indirect evidence suggests with some certainty that they have been written in Cyprus, show little or no evidence of Cypriot dialectal features as we know them today or even as they are attested in other texts of the same period. These are in the first place texts written in predominantly learned registers,[8] which are of no true interest for historical linguists, except in the domain of the lexicon. Another category is constituted by texts written in vernacular registers, judged of Cypriot origin through non-linguistic criteria, but possessing linguistic features that are not restricted to Medieval Cypriot but also occur in other regions of the Medieval Greek speaking world (for example, the verse work *ο Πρέσβυς Ιππότης*).[9]

Such texts may of course have been written in Cyprus by non-Cypriot authors; one should not exclude the possibility though, especially for texts of an earlier date, that they were written in a period when distinctive dialectal features of Cypriot had not yet emerged (for example, a version of *Spaneas* produced in Cyprus). Since the potential corpus of MC texts, discussed here, aims at providing a resource for the study of the emergence and development of Medieval Cypriot, such texts need to be included. Their exclusion could lead to a distorted picture of what is linguistically common and what particular to a specific area in medieval times and to the creation of circular arguments for the definition of dialectal features.

Based on this discussion, the potential CMCT should cover all texts belonging to the following categories:

Texts written or copied in Cyprus (as established on the basis of non-linguistic, e.g. philological evidence) in a predominantly vernacular register; following subcategories are possible

(i) A text written or copied by a Cypriot[10] author or scribe in Cyprus
(ii) A text written or copied by a Cypriot author or scribe outside Cyprus
(iii) A text written or copied by a non-Cypriot author or scribe in Cyprus

Texts showing clear evidence of linguistic features presumed to belong to Medieval Cypriot as inferred from comparison with Modern Cypriot according to the premises of historical linguistics; most text of this category will also belong to the first category. There are texts, however, which have been identified by past scholars as "Cypriot" based solely on linguistic criteria without external evidence (e.g. the *Ανακάλημα τῆς Κωσταντινόπολης*), although they very rarely present features which are exclusive to Cypriot- mainly due to our insufficient knowledge of the areal distribution of linguistic features in Medieval Greek. These also need to be included in the corpus, if only with a warning about their disputed status.

This discussion has also shown the necessity for a mechanism that will provide information to the potential corpus user about the characteristics of each text that

[8] For the linguistic features of learned ecclesiastical documents from Cyprus, see Kalli 1997.
[9] For this and all subsequent references to editions of literary medieval texts, we refer the reader to the list of primary sources in Kazazis et al. (2001), also available on-line at http://www.greek-language.gr/
[10] "Cypriot" here and in similar uses equals "a native speaker of Cypriot".

allow its placement to one of the above categories.[11] This information together with relevant references to secondary bibliography has to be attached to the text as part of its metadata and should be editable, so that it may be updated if new evidence (philological or linguistic) is found concerning the provenance of the specific text.

2.2. Preliminary overview on available texts

Based on research conducted so far by the Grammar of Medieval Greek project and on secondary bibliography (e.g. Kechagioglou, forthcoming, Kitromilidis, 2002, Grivaud, 1996), we can present a fist overview of available texts, ordered according to date and text-type or genre: the information provided here has a provisional character, and includes only published material.

We organise our material firstly in two broad categories: non-literary and literary texts. We include one large non-literary (legal) text, the Ασσίζες in the second category, as it has been transmitted to us in two different manuscripts, thus bearing more resemblance in terms of its annotation and transcription to literary texts. Individual texts are grouped together within these categories according to their century of composition and their text-type or genre.

2.2.1. Non-literary texts

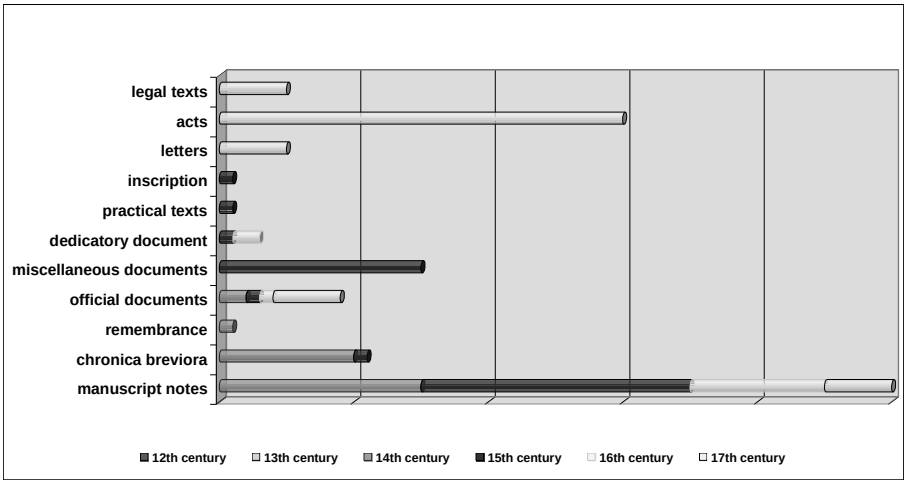
By far the most productive text-type in this category are manucript notes and marginalia, which provide us with a considerable sample of small texts from all centuries. Many of the earlier texts show little or now dialectal features, while in notes from the 16th century Cypriot dialectal features are attested more regularly. Two similar text-types, remembrances and short chronicles offer some of the earliest samples of written language in Medieval Cyprus (but are sometimes difficult to date precisely).

The text-type "miscellaneous documents" is covered by a single publication, containing official documents of the French rulers of Cyprus (Richard-Papadopoullos, 1983). The existence of acts only from the 17th century must relate to specific socio-historical developments in the island- all other Venetial ruled areas have provided us with rich archival material.[12] Notable is also the scarcity of private and official letters from the 16th and 17th century. This is probably due to the Ottoman conquest of the island in 1571, although further research has to be conducted for unpublished archives etc.

An overview of the available non-literary sources can be seen in the graph below:

[11] On the necessity of encoding provenance and genre parameters in a historical corpus see Rissanen (1992) and especially Meurman-Solin (2001).

[12] The largest document collections are Kyrris (1987) and Perdikis (1998), both after the 17th c. For a full list of the non-literary publications used for the compilation of the CMCT, please contact the authors of this paper.



2.2.2. Literary texts

Under literary texts we include for the time being only texts for which we are certain that they have been written in Cyprus or by Cypriot authors. We have to stress that our research in this domain has not been as extensive as for non-literary texts.

Prose texts

The overall picture is the reverse of that from in other areas in Medieval times: we have a large corpus of prose texts and much fewer verse texts. Two large historiographical works (the Chronicles of Machairas and Voustronios), both available in more then one manuscript versions, build the main part of the corpus of prose texts.

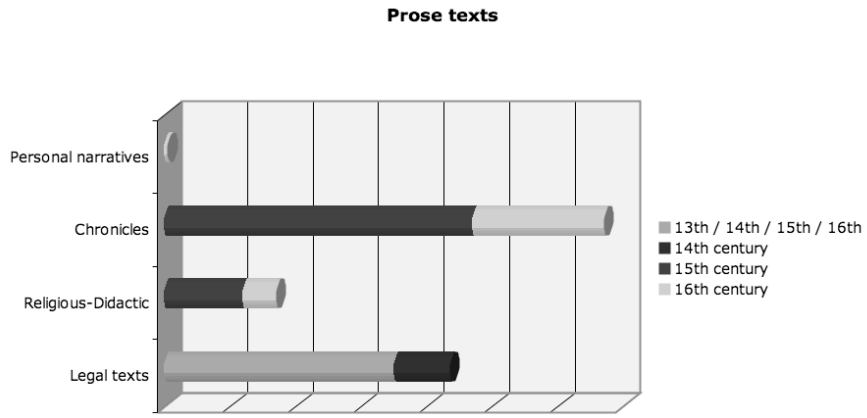
The Ασίζες του Βασίλειου των Ιεροσολύμων και της Κύπρου, available in two manuscript versions, a long legal text currently not available in a reliable edition, can be considered as a key-text for research on Medieval Cypriot, provided that it is possible to date it more precisely. There exist also interesting prose texts from the 16th century with strong dialectal features in the form of intralingual translations of religious texts and a personal narrative (a sample of which will be provided below).

Verse texts

We have only been able to trace two verse texts written in Cyprus in the 16th and 17th century. A collection of Love poems of various authors (to which the editor has inserted dialectal features not present in the manuscript)[13] and a long Verse Chronicle written by a Cypriot author in Bucarest with few dialectal features.[14]

[13] Siapkaras-Pitsillidès Th., Le Pétrarquisme en Chypre, poèmes d'amour en dialecte chypriote d'après un manuscrit du XVIe siècle, Paris ²1975.
[14] Kaplanis A., "Ioakeim Kyprios' Struggle (Mid-17th Century). A study of the text with an edition of selected passages", PhD Thesis, University of Cambridge 2003

An overview of literary texts can be seen in the graph below



2.3. Modelling and implementation issues

2.3.1. The priority of manuscripts in "text languages"

The creation of an electronic resource that will be primarily used by historical linguists has to take into consideration one of the predominant assumptions of the discipline. Historical linguistics is not working with speakers or with spoken language, despite the fact that these are better indications both of the synchronic system of the language as well as of its diachronic evolution. It is working with 'text languages' whose native speakers are the manuscripts (Fleischman, 2000).

While it may be argued that a critical edition is the only way for the literary reception of a medieval literary text today, according to our experience so far many editors are often normalizing, unfortunately sometimes even tacitly, the spelling of witnesses even in cases of diplomatic editions. Converting such an edition to electronic text adds only limited value to it as source for linguistic research.

This is not to say that modern editions of literary or non-literary texts are not suitable for linguistic analysis, when they adopt a conservative editorial approach that preserves the true linguistic features of the text. While this is many times the case, even the slightest editorial intervention (like adopting a unified spelling system with normalization of orthography, word-division, analysis of abbreviations etc.) is a form of interpretation, based on general assumptions on the linguistic form of the text. This layer of interpretation may obstruct some form of linguistic research (for instance the relationship between grapheme and phoneme in the writing system or the form of inflectional morphemes in the case of abbreviations).

An electronic historical corpus should therefore ideally model, i.e. represent, the only trustworthy witnesses of a "text language" available: the manuscripts or, in the case of texts transmitted only in print, the early editions themselves. In such an ideal corpus every manuscript of a text belonging to the electronic CMCT would

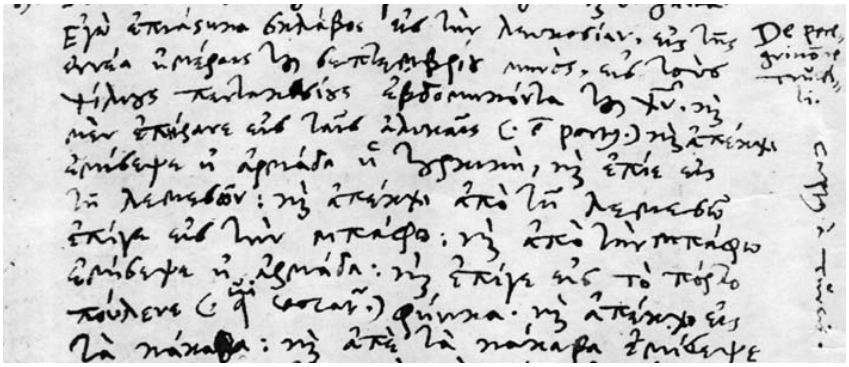
be available in the form of a transcription in electronic form, preserving as many of the original characteristics of the text as possible; each page of the manuscript should also ideally be linked to a digital picture of the actual manuscript.

We could consequently think that if we create a "one to one" electronic representation of such a diplomatic text in some generally acceptable and easily available format we automatically possess a versatile and valuable methodological tool for linguistic description. This is unfortunately not the case: preserving the original spelling of the manuscript (which in most, if not in all, cases does not follow any standards) makes searches much more complicated, as it is not always possible to foresee all possible spellings of a word, especially in the case of Greek, where historical orthography is so far distant from phonetic reality; the existence of a general reference schema is important, for quoting and verification purposes; absence of normalized spelling makes the use of external tools (like digital lexica or morphological parsers) only possible through complicated interfaces that take into consideration all possible spellings; sorting and concordancing mechanisms can be applied with greater difficulty and do not provide the researcher with a "transparent" set of data, suitable for linguistic interpretation.

The compiler of a corpus is therefore confronted with a dilemma: if s/he decides to normalize the text of the manuscripts, saving much trouble and time, he loses vital information. Creating double and aligned versions (diplomatic and normalized) of the same text is time expensive and difficult to manage. For such an electronic corpus to meet the high standards needed for cutting edge linguistic research with a design that is not imposing restrictions, one has to make use of more advanced methods for representation of texts in a digital environment with the use of extensible markup languages.

2.3.2. The use of XML in digital transcriptions: transcription and annotation

We would like to demonstrate what exactly XML (extensible markup language) is and how it can be effectively used in producing an electronic corpus of CMCT with the help of small sample of an as yet unpublished Cypriot text of the 16th century: a narrative about the adventures of Alexander Trucello, a Cypriot from Nicosia who visited the German humanist Martinus Crusius in 1582. Trucello himself was illiterate but narrated the Battle of Lepanto (1571) from an eye witness perspective to Crusius, who wrote them down in a manuscript (today Cod. Tybingensis Mb 37 of the University Library of Tübingen). A small sample of the manuscript can be seen in the slide:



Cod. Tyb. Mb 37, part of p. 80

A diplomatic transcription of the first few lines of this text looks like this (slide):

> Ἐγὼ ἐπιάστηκα σκλάβος εἰς τὴν λευκοσίαν, εἰς τῆς |[2] ἐννέα ἡμέραις τοῦ σεπτεμβρίου μηνὸς, εἰς τοὺς |[3] χίλιους πεντακοσίους ἑβδομηκόντα τοῦ χ(ριστο)ῦ. καὶ μὲν |[4] ἐπήρανε εἰς ταῖς ἀλυκαῖς (e(st) portus) καὶ ἀπέκχι |[5] ἐμίσεψε ἡ ἀρμάδα ἡ τουρκικὴ, καὶ ἐπίε εἰς |[6] τῆ λεμεσῶν: καὶ ἀπέκχι ἀπὸ τῆ λεμεσῶ |[7] ἐπίγε εἰς τὴν μπάφω: καὶ ἀπὸ τὴν μπάφω |[6] ἐμίσεψε ἡ ἀρμάδα: κ(αὶ) ἐπῆγε εἰς τὸ πόρτο

This diplomatic transcription follows some broadly accepted editorial conventions that allow the representation of a text written in a manuscript as printed text. These symbols represent a kind of declarative metalanguage that is inserted into the text in order to identify items of interest for understanding or analyzing the text that cannot otherwise be denoted in a compact manner.

Such declarative metalanguage about some information or data is commonly called "markup" when it refers to computer-related methods of representation of physical objects in the computer. It is possible to represent all symbols used in the above transcription with the help of an extensible markup language and so-called semantic tags. Our example from the narrative of Trucello looks in XML (following the guidelines of the Text Encoding Initiative[15]) as follows (slide):

```
<pb
    n="80"
    xml:id="GH80.jpg"/>
<p>Ἐγὼ ἐπιάστηκα σκλάβος εἰς τὴν λευκοσίαν, εἰς τῆς <lb/>ἐννέα ἡμέραις τοῦ σεπτεμβρίου
    μηνὸς, εἰς τοὺς <lb/> χίλιους πεντακοσίους ἑβδομηκόντα τοῦ χ<abbr>ριστο</abbr>ῦ. καὶ
    <lb/> μὲν ἐπήρανε εἰς ταῖς ἀλυκαῖς (<seg
        xml:lang="lat">e<abbr>st</abbr> portus</seg>) καὶ ἀπέκχι <lb/>ἐμίσεψε ἡ ἀρμάδα ἡ
    τουρκικὴ, καὶ ἐπίε εἰς <lb/> τῆ λεμεσῶν: καὶ ἀπέκχι ἀπὸ τῆ λεμεσῶ <lb/> ἐπίγε εἰς
    τὴν μπάφω: </p>
```

The editorial conventions used in the diplomatic transcription have been converted to XML-tags (or elements), that wrap the text they annotate: the tag <abbr> denotes an abbreviation, an empty tag <lb> (empty because it does not wrap

[15] See Sperber-McQueen & Burnard (1990). The most recent version of the encoding guidelines is available electronically at http://www.tei-c.org/cms/Guidelines/P5/get_p5.xml

some text in it) is placed where a line break occurs in the manuscript and as a marker of a page break (<pb>) together with the number page and the name of the computer file holding a picture of the manuscript. Paragraphs are not marked with simple line breaks but with an opening ( ) and a closing ( ) tag. With the use of so called "attributes" it is possible to add further information about a specific tag: Crusius' commentary in Latin (*est portus*) has been marked as a different segment of text (with the element <seg>), written in Latin (with use of the attribute "xml:lang" inside the <seg> tag).

Furthermore, XML allows for multi-modal text annotation, thus allowing for representation of text in different ways, based on a sole annotated file. With the use of the element <choice> and its children (<orig> and <reg>) it is possible to encode both the original and the regularized form of a word, as shown on the second example on the slide:

```
<p>Ἐγὼ ἐπιάστηκα σκλάβος εἰς τὴν <choice>
        <orig>λευκοσίαν</orig>
        <reg>λευκωσίαν</reg>
    </choice>, εἰς <choice>
        <orig>τῆς</orig>
        <reg>τὶς</reg>
    </choice>
    <lb/>ἐννέα <choice>
        <orig>ἡμέραις</orig>
        <reg>ἡμέρες</reg>
    </choice> τοῦ σεπτεμβρίου μηνὸς, εἰς τοὺς <lb/> χίλιους πεντακοσίους ἑβδομηκόντα τοῦ <choice>
        <orig>χ<abbr>ριστο</abbr>ῦ</orig>
        <reg>Χριστοῦ</reg>
    </choice>. καὶ <lb/> μὲν ἐπήρανε εἰς <choice>
        <orig>ταῖς</orig>
        <reg>τὲς</reg>
    </choice>
    <choice>
        <orig>ἁλυκαῖς</orig>
        <reg>ἁλυκὲς</reg>
    </choice> (<seg
        xml:lang="lat"><choice>
            <orig>e<abbr>st</abbr></orig>
            <reg>est</reg>
        </choice> portus</seg>) καὶ <choice>
        <orig>ἀπέκχι</orig>
        <reg>ἀπέκχει</reg>
    </choice>
</p>
```

From such an electronic file it is then possible with the use of so-called stylesheets, i.e. mini-computer programmes that can select specific tags and display them in the required order, to create both a normalized and a diplomatic version of the text, as in the second example on this slide:

The use of TEI-XML for the creation of an electronic historical corpus is not limited only to the annotation of structural elements of the text but can also be used for annotation of linguistic features. The linguistic annotation is stored within the transcription file. In the above example, the writing <κχ> in the adverb απέκει that appears systematically in this text, can be considered as evidence of stop germination consisting not of consonant doubling but of aspiration [k^h].

One way of representing this would look as follows:

```
<choice>
        <orig>
          <w
                lemma="ἀπέκει"> ἀπέ<c
                   function="aspiration"
                   n="kʰ"
                   type="phoneme">κχ</c>ι </w>
        </orig>
        <reg>ἀπέκχει</reg>
</choice>
```

The boundaries of the word are marked with the element <w>, while the characters of interest are marked up with the use of the element <c>. With the use of attributes is possible to provide linguistic annotation (lemma of the word, phonetic realisation, labelling of phenomenon etc.), which is saved within the same file as the text and can be summarized or queried automatically.

We have to stress here that the "vocabulary" of XML can be easily extended in order to cover more specialized annotations or special encoding needs.[16] This example was chosen on purpose, as we consider of great interest for the historical investigation of the Cypriot dialect: it may constitute a first indication that the modern pronunciation of Cypriot geminate stops was aspirated also in the medieval period. Medieval Cypriot documents written by Greeks invariably spell geminate stops as double or single consonants and it is a piece of luck for us that this document is a transcription by a non-native, who, furthermore, has little knowledge of the language. There are some indirect indications of the aspirated status of medieval Cypriot stops (see Davy & Panayotou, 2004): the fact that the aspirated pronunciation appears also in other Southeastern dialects (Rhodes, Cos, Carpathos etc.), which suggests an early date of appearance and spread, and the adaptation of stops contained in medieval loanwords from Italian, French and Turkish which is similar to modern loan adaptations. But until now, there were no actual spellings of an aspirated stop. However, one must never be hasty with the interpretation of medieval spellings: <κχ> preceding a front vowel, as in the case of απέκχι here, could also be just a variant way of recording the palatalized (with so called "tsitakismos") pronunciation of the velar stop [k].

In general geminated /k/ is very rarely represented as such in Cypriot manuscripts (whereas for nasals, liquids and /s/ the evidence is much more abundant, and there is also some evidence for the other stops, /p/ and /t/). A few available examples (from the database of the Grammar of Medieval Greek) are the following:

- ἐτζακκίσαν τους MACH., *Chron.* V 40.25 (Dawkins). In Pieris – Konnari 89.20 ἐτζακ|κίσαν τοὺς
- ετσακκίσεν VOUSTR., *Chron.* A 20.15 (Kechagioglou).

[16] Examples of customised schemata.

- σακκουμάνος MACH., *Chron.* V 440.19 (Dawkins). In Pieris – Konnari 324.50 σακκου|μάνος
- τω αυτὼν. / ττεμεσού κκιν / τὸ αὐτὸν ττεμεσούκκιν (1704, deed of sale/Cyprus, ed. KYRRIS 1987: 2, 54.7–8)

This example clearly demonstrates one of the dangers of a linguistically annotated corpus: any kind of analysis by definition involves interpretation, and the more the layers of analysis/annotation and the amount of descriptive details/tags provided, the greater the subjectivity of the resulting text(s). Indeed, the design of many corpora includes only structural or "uncontroversial" linguistic annotation under the premise that the user of an electronic corpus can develop tools that will suit his own research needs. In that sense, an open access policy that allows for users to create new versions of the corpus with added annotation of their own could facilitate collaborative research among researchers and research groups.

2.3.3. Compilation process: work stages, timetable, personnel

From what has been said until now it is clear that such corpus has partly an interdisciplinary character: it builds upon the expertise of historical linguists for defining the aims and objectives of the corpus in general and the characteristics of its design; an extensive part of its creation has to be carried out by philologists trained in palaeography, who nevertheless need to have a minimum of linguistic training both to be able to communicate with the linguists and to understand the scope of the whole endeavour; finally, computer specialists have to offer advice, support and training both for choosing suitable methods and annotation schemes but also for creating the final, published and distributed product.

References

Davy, J., & Panagiotou, A. (2004): Phonological Constraints on the Phonetics of Cypriot Greek: Does Cypriot have geminate stops? In G. Catsimali, A. Kalokerinos, E. Anagnostopoulou, & I. Kappa (Eds.), *Proceedings of the 6th International Conference on Greek Linguistics*. Rethymno: Linguistics Lab. CD-ROM. [available at http://www.philology.uoc.gr/conferences/6thICGL]

Decorte, S. (2003). Taalkundige verrijking in historische corpora in relatie tot de Geintegreerde Taalbank. *INL Working Papers, 01*, 1–65.

Fleischmann, S. (2000). Methodologies and ideologies in historical linguistics: On Working with older languages. In S. Herring et al. (Eds.), *Textual Parameters in Older Languages* (pp. 33–58). Amsterdam: Benjamins.

Grivaud, G. (1996). Ο πνευματικός βίος και η γραμματολογία κατά την περίοδο της Φραγκοκρατίας. In Th. Papadopoullos (Ed.) *Ιστορία της Κύπρου. Τόμος Ε΄, Μεσαιωνικόν Βασίλειον-Ενετοκρατία* (pp. 863–1208). Nicosia: Ίδρυμα Αρχιεπισκόπου Μακαρίου.

Goutsos, D. (2003): Σώμα Ελληνικών Κειμένων: Σχεδιασμός και υλοποίηση. In G. Catsimali, A. Kalokerinos, E. Anagnostopoulou, & I. Kappa (Eds.), *Proceedings of the 6th International Conference on Greek Linguistics*. Rethymno: Linguistics Lab. CD-ROM. [available at http://www.philology.uoc.gr/conferences/6thICGL]

Holton, D. (forthcoming). The Cambridge Grammar of Medieval Greek project: aims, scope, research questions. In G. Mavromatis (Ed.), *Πρακτικά του Συνεδρίου 'Neograeca Medii Aevi VI: Γλώσσα, παράδοση και ποιητική,* (University of Ioannina, 29.9–2.10.2005).

Kalli, M. (1997). Η ελληνική εκκλησιαστική γλώσσα στην Κύπρο κατά την βυζαντινή και μεταβυζαντινή περίοδο. *Παρουσία, 11–12*, 245–265.

Kechagioglou, G. (forthcoming). Ελληνόγλωσση γραμματεία στην Κύπρο (εποχή Κομνηνών-1570). In G. Mavromatis (Ed.) *Πρακτικά του Συνεδρίου 'Neograeca Medii Aevi VI: Γλώσσα, παράδοση και ποιητική,* (University of Ioannina, 29.9–2.10.2005).

Kitromilidis, P. (2002). *Κυπριακή Λογιοσύνη. 1571–1878. Προσωπογραφική θεώρηση*. Nicosia.

Kazazis, I. et al. (Eds.) *Επιτομή του Λεξικού της Μεσαιωνικής Ελληνικής Δημώδους Γραμματείας (1100–1669)*. Thessaloniki: Centre for Greek language.

Kyrris, C. P. (Ed.). (1987). *The Kanakaria documents 1666–1850: sale and donation deeds edited with introduction and commentary* [Texts and Commentaries of the history of Cyprus, 14]. Nicosia.

Lendari, T. & Toufexis, N. (forthcoming). Γλωσσική περιγραφή και ανάλυση με βάση τα κείμενα της περιόδου 1100–1700: μεθοδολογικά και πρακτικά ζητήματα της γραμματικής της μεσαιωνικής ελληνικής. In G. Mavromatis (Ed.), *Πρακτικά του Συνεδρίου 'Neograeca Medii Aevi VI: Γλώσσα, παράδοση και ποιητική'* (University of Ioannina, 29.9–2.10.2005).

Manolessou, I. (2008). Μεσαιωνική γραμματική και ιστορική διαλεκτολογία: παρατηρήσεις με αφετηρία την κυπριακή διάλεκτο [Medieval Grammar and historical dialectology: observations with respect to the Cypriot dialect]. In *Πρακτικά Ε΄ Διεθνούς Συνεδρίου Νεοελληνικής Διαλεκτολογίας*. Athens: Academy of Athens, 425–448.

Meurman-Solin, A. (2001). Structured Text Corpora in the Study of Language Variation and Change. *Literary and Linguistic Computing, 16*(1), 5–27.

Pantelia, M. C. (2003). The Thesaurus Linguae Graecae Project: Looking towards the 21st century. In J. N. Kazazis (Ed.), *The Lexicography of Ancient, Medieval and Modern Greek Literature*. Thessaloniki: Centre for Greek Language 31–37, 151–156 [available at http://www.greek-language.gr/greekLang/files/document/conference-1997/02-el-033-040.pdf]

Perdikis, St. (1998). *Δικαιοπρακτικὰ ἔγγραφα Ἱερᾶς Μονῆς Κύκκου, 1619–1839* [Ἀρχεῖο Ἱερᾶς Μονῆς Κύκκου, 4], Nicosia.

Pusch, C. D., Kabatek, J., & Raible, W. (Eds.). (2005). *Romanistische Korpuslingustik II: Korpora und diachrone Sprachwissenschaft*. Tübingen: Narr.

Richard, J. & Papadopoulos, Th. (Eds.). (1983). *Le Livre des remembrances de la Secrète du royaume de Chypre (1468–1469)*. Nicosia: Centre des Recherches Scientifiques.

Rissanen, M. (1992). The diachronic corpus as a window to the history of English. In J. Svartik (Ed.), *Directions in Corpus Linguistics* (pp. 185–206). Berlin: De Gruyter.

—— (2000). The world of English historical corpora: from Caedmon to the Computer Age. *Journal of English Linguistics, 28*(1), 7–20.

Sperberg-McQueen, M. & Burnard, L. (Eds.). (1990). *Text Encoding Initiative: Guidelines For The Encoding And Interchange Of* Machine-*Readable Texts.* Chicago & Oxford.