

Νότης Τουφεξής • Notis Toufexis  
University of Cambridge  
nt262@cam.ac.uk

“Creating a database for the ‘Grammar of Medieval Greek’ project”,  
στο: Γ. Μαυρομάτης (εκδ.), *Πρακτικά του διεθνούς συνεδρίου Neograeca Medii Aevi VI: Πρώιμη νεοελληνική δημόδης γραμματεία, Γλώσσα, παράδοση και ποιητική*, Σεπτέμβριος 2005, Ιωάννινα (in press)

(This paper reflects the status-quo of 2005. Since then the database described here has been considerably changed. Please contact me for a more up-to-date description of the database.)

## Creating a database for the “Grammar of Medieval Greek” project\*

### *Abstract*

The main goal of the “Grammar of Medieval Greek project” is to produce a comprehensive Grammar of Medieval Greek in book form;<sup>1</sup> an electronic publication of the material collected in the process, or in the Grammar itself, is not planned for the time being. However, from its beginning the research project relies heavily upon the use of electronic resources; this is a reasonable decision when one has to collect and organize large amounts of data. Nowadays it is also often considered as a prerequisite for funding a large-scale research project. This paper aims at describing all issues that are related to the creation of a custom-built electronic database and tries not to concentrate on technical aspects (as the interested reader can find a full description of technical matters elsewhere<sup>2</sup>) but on issues concerning modelling of data and research methodology.

### 1 Objectives and main modelling decisions

One of the main objectives of the Grammar of Medieval Greek project was the compilation of a bibliography of as many texts as possible (both literary and non-literary) from the period under examination, their location and acquisition as well as their classification according to specific categories. All collected published texts form a corpus of texts on the analysis of which the “Grammar of Medieval Greek” will be based (see Holton and Lendari & Toufexis in this volume for more details). All project materials relating to the compilation of this Grammar of an extinct language are hence “document oriented”,<sup>3</sup> i.e. they relate to a written text of some sort and annotate it (“markup”) according to selected principles.

Several projects concerned with compiling similar grammars of other European languages go one step further in adopting what we can call the electronic corpus approach: building as a first step an electronic corpus of texts which is then consequently analysed.<sup>4</sup> A review of several descriptions of contemporary research projects has shown that most of these build upon already available

---

\*This paper is the outcome of research conducted for the research project “A Grammar of Medieval Greek” at the University of Cambridge. The project is funded by a grant from the Arts & Humanities Research Council. I am grateful to David Holton and Tina Lendari for their suggestions and corrections on matters of style.

<sup>1</sup> See the paper by David Holton in this volume.

<sup>2</sup> A detailed description of the databases can be obtained by contacting the author of this paper.

<sup>3</sup> For this terminology see J. Bradley, «Documents and Data: Modelling Materials for Humanities Research in XML and Relational Databases», *Literary and Linguistic Computing* 20 (2005) 133-51 and L. Hunter, «Fact--Information--Data--Knowledge: Databases as a Way of Organizing Knowledge», *Literary and Linguistic Computing* 5 (1990) 49-57; cf. W. McCarty, «Humanities Computing: Essential Problems, Experimental Practice», *Literary and Linguistic Computing* 17 (2002) 103-25 for a more general discussion of relevant issues.

fundamental research: abundance of reliable editions and linguistic bibliography, cartography of dialectal regions and studies on the diatopic distribution of texts and in some cases electronic editions of key texts.<sup>5</sup> On top of these resources specialized databases or electronic corpora can be created relatively easy, so that specific phenomena or subjects can be analysed in greater detail.<sup>6</sup>

Given the lack of similar basic research in the field of Medieval Greek studies and the fact that for some areas there was practically no groundwork available, our approach had to be different right from the beginning. At a very early stage it was determined that it would not be feasible to create an electronic corpus that would meet even the minimal standards required for the envisaged linguistic description. The reasons for this decision are many and would require a special paper for their presentation.<sup>7</sup> We have nevertheless been able to compile a relatively small electronic corpus (based on material supplied by editors) that does not yet meet all requirements but is valuable in some areas of research.

In the course of our project all members of the research team would thus be involved in reading through texts of the established corpus and collecting samples of text (excerpts), which document or illustrate all linguistic features that need to be discussed or analysed in the Grammar; this list of features or phenomena has been precompiled for all levels of linguistic analysis (accompanied by a review of avail-

---

<sup>4</sup> See for instance the approaches of the ‘Middle English Grammar project’ (“A machine-readable and diatopically-ordered corpus of texts from the late Middle English period, consisting of some 3.5 million words directly transcribed from either the original manuscripts or good-quality microfilms”, <http://www.arts.gla.ac.uk/sesll/englang/ihs1/projects/MEG/MEG.htm#Objs> accessed 18/06/2006) and the “Projekt Mittelhochdeutsche Grammatik” (“Die neue mittelhochdeutsche Grammatik wird von Grund auf neu aus den Quellen, d.h. den mhd. Handschriften, erarbeitet. ... Grundlage ist vielmehr ein Quellenkorpus, das in folgender Weise strukturiert ist: Für jeden Zeitabschnitt von 50 Jahren enthält es aus jeder mhd. Sprachlandschaft zwei Vers- und zwei Prosatexte, ab 1250 auch geeignete Gruppen von Urkunden“, [http://www.mittelhochdeutsche-grammatik.info/DE/korpus\\_bonn.html](http://www.mittelhochdeutsche-grammatik.info/DE/korpus_bonn.html) accessed 18/06/2006).

<sup>5</sup> S. Horobin and J. Smith, «A database of Middle English spelling», *Literary and Linguistic Computing* 14 (1999) 359-74 describe, for instance, available resources in the field of Middle English studies; almost all similar research projects set about creating new or enhanced versions of already existing reference Grammars based on contemporary methods. The situation in the field of Medieval Greek studies is of course totally different.

<sup>6</sup> See for instance S. R. Parkinson and A. H. A. Emiliano, «Encoding Medieval Abbreviations for Computer Analysis (from Latin-Portuguese and Portuguese Non-literary Sources)», *Literary and Linguistic Computing* 17 (2002) 345-60 for a specialized electronic transcription method for the analysis of the graphematic system of Old Portuguese or Horobin and Smith, «A database of Middle English spelling», for a database dedicated to the analysis of Middle English spelling.

<sup>7</sup> We list just some of the most problematic issues: codification of manuscript variants in texts with multiple versions / manuscripts; normalization of different editorial practices (and spelling in diplomatic editions) so that universal searches are possible; automatic or semiautomatic lemmatization, etc. Technical solutions are of course available for most if not all of these issues; their implementation however is not possible within the time limits of our project. See also D. M. L. Philippides, «Ο υπολογιστής στη σύγκριση κρητικών αναγεννησιακών κειμένων: εκδοτικές απορίες με αφετηρία το λεξιλόγιο της ρίμας», in E. Jeffreys and M. Jeffreys (eds.), *Αναδρομικά και Προδρομικά. Approaches to Texts in Early Modern Greek. Papers from the conference Neograeca Medii Aevi V, September 2000* Oxford 2005, 101-14 for a discussion of similar problems from the perspective of the “computer philologist”.

able bibliography). The primary material (text samples but also bibliographical references of primary sources and secondary bibliography) will be collected from different locations but has to be accessible to all members of the team at all times; furthermore it should be available in a form compatible with the design of the publication, i.e. according to several specific criteria (date, provenance, genre/type of text etc.).

Collection of linguistic material can only be selective, the aim being the collection of enough instances for the illustration of each particular linguistic phenomenon or feature, or, for the analysis of less well studied phenomena,<sup>8</sup> the compilation of a well structured set of representative instances as a basis for detailed linguistic description.<sup>9</sup> The research team is involved in two different but interconnected kinds of activity: codification of linguistic theory expressed through annotating labels as well as collection of text excerpts along with their linguistic annotation, which makes use of these labels.

From this discussion it is probably evident that electronic resources developed for our project should therefore not model the primary texts themselves but the task of composing the specific “Grammar”.<sup>10</sup> Their role would be to provide an electronic tool for the collection, organization and annotation of material that would eventually be used for the composition of the “Grammar of Medieval Greek” in book form.

Data collected during that process fall generally into three different categories:

- Bibliographical references of different kinds (primary sources, secondary bibliography etc.) and their annotations
- Information on different categories of text, classified according to a particular scheme to suit linguistic research
- Linguistic/metalinguistic data (excerpts from texts and metalinguistic annotations).

---

<sup>8</sup> For a discussion of deficiencies in the specific field of linguistic research on Medieval Greek see the paper of Lendari & Toufexis in this volume.

<sup>9</sup> The process of collecting samples of texts (or full texts) and annotating them for linguistic analysis is called linguistic annotation or markup in the relevant literature. There is abundant bibliography on this issue especially when this involves the use of computers. For a more general introduction and as a starting point see H. v. Halteren, *Excursions into syntactic databases* [Language and computers 21], Amsterdam 1997 and M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford 2005 (also available under <http://ahds.ac.uk/linguistic-corpora/>, accessed 27/06/2006) (with more information on electronic linguistic corpora).

<sup>10</sup> On the decisive question of modelling in relation to the creation of relational databases for the humanities see Bradley, «Documents and Data: Modelling Materials for Humanities Research in XML and Relational Databases», J. Bradley and H. Short, «Texts into Databases: The Evolving Field of New-style Prosopography», *Literary and Linguistic Computing* 20 (2005) 3-24; for a different modelling approach concerning historical documents of the same period see P. Costantopoulos, M. Doerr, M. Theodoridou, and M. Tzobanakis, «Historical documents as monuments and as sources», paper given at the conference *Computer Applications and Quantitative Methods in Archaeology, CAA2002*, Herakleion, Crete, 2-6 April 2002 (<http://www.ics.forth.gr/isl/publications/paperlink/caa2002.pdf>, accessed 29/06/2006).

The author would like to acknowledge the help of Michael Jeffreys in questions relating to modelling of data and database construction at an initial stage.

These data will be collected in three consecutive stages:

- i) Heuristic and corpus compilation (accumulation of bibliographical references, verification of their reliability and relevance and primary classification of contents)
- ii) Collection of primary linguistic data based on an established corpus (collection and annotation of text excerpts for each identified phenomenon/item)
- iii) Grouping and analysis of collected data / Compilation of the Grammar (searching and browsing of collected material, compilation of lists and reports, revision of problematic cases, composition of Grammar)

Following a phase of conceptual analysis of sample data (a small corpus of published letters) the decision was made to create an electronic database based on the relational model.<sup>11</sup> The source (the actual texts) would only be represented in the form of (relatively short) excerpts: issues pertaining to the use of relational databases for the representation of linguistic data<sup>12</sup> were therefore not considered as problematic. Much more importantly the data in question can be conceived as a series of linked hierarchies and classifications which can be very adequately represented with the use of the relational model (see section 3.2 for more details). The collected data should be open to aggregation and some sort of quantitative analysis although this is not considered as the main objective of the overall study.<sup>13</sup> In short, many of the criteria for the use of relational databases in our field were met.<sup>14</sup>

The objective behind the use of such a database was to construct an electronically available, well organized, verified and semi-interpreted set of truly relevant

---

<sup>11</sup> It should be stressed here that decisions of this kind often take into consideration practical matters and availability of resources and personnel competence. Our main focus lies therefore with the clarification of the consequences that these decisions have both for the methodology of the project and the overall research.

<sup>12</sup> For a discussion of such problems see J. Lawler and H. A. Dry, *Using computers in linguistics : a practical guide*, London - New York 1998 (also available under <http://www.routledge.com/linguistics/using-comp.html>, accessed 21/06/2006), esp. the examples provided in the online version of the book.

<sup>13</sup> This is a major issue that was also discussed during the Neograeca Medii Aevi VI conference. Quantification in the sense of actual statistical analysis is of course not possible with this kind of data collection and organization. The big issue of using a quantitative approach in this kind of research needs to be discussed thoroughly, as the merits of quantification are not always immediately evident.

<sup>14</sup> For a discussion of these issues see J. Bradley, «Relational database design and the Reconstruction of the British Medical profession: Constraints and Strategies», *History and Computing* 6 (1994) 75 and S. Townsend, C. Chappell, and O. Struijvé (eds.), *Digitising history: a guide to creating digital resources from historical documents* [AHDS guides to good practice in the creation and use of digital resources], Oxford 1999 (also available under <http://www.ahds.ac.uk/history/creating/guides/digitising-history/index.html>, accessed 21/06/2006). Although both publications relate to historical research, they are very relevant to issues described here, as they discuss serious conceptual and methodological questions.

primary data; this dataset can also be conceived as a “proto-narrative”<sup>15</sup> of the actual “Grammar of Medieval Greek”; in the final “product”, the Grammar in book form, and with the help of linguistic theory and philological interpretation this material will be formed into a single cohesive narrative, a comprehensive “Grammar of Medieval Greek”. Because of its structure, the collected dataset will be useful for further studies of other aspects of Medieval Greek and could easily be used as basis for future research.

## 2 Available electronic resources in the field of Medieval Greek studies

As we have already said, the decision has been made, right at the beginning of our project, that the creation of an electronic corpus of texts from our period was not feasible. In this section we aim at providing an overview of relevant electronic resources which have either been born “digital” (i.e. have been conceived and realized with the objective of electronic publication or use) or have been secondarily transferred/converted to the digital medium. After this overview the specific reasons for our decision will hopefully become even clearer.

Electronic texts and digital libraries:

Unlike other European literatures and languages Greek lacks a dedicated electronic, fully searchable corpus of texts of its medieval period written in the vernacular. After an early initiative in the nineties with some success<sup>16</sup> there has been no similar initiative in this field and there are to our knowledge no initiatives planned for the near future. So far, only the Thesaurus Linguae Graecae (<http://www.tlg.uci.edu>) is actively pursuing the task of producing a full-scale electronic corpus of Medieval Greek texts as part of its broader digital library programme of texts in Greek. Because of its broader scope the TLG is understandably not capable of meeting all demands of the “Neograeca Medii Aevi” research community or our project;<sup>17</sup> it has nevertheless proved extremely valuable, particularly in some areas of our research.

There have been some other initiatives to create electronic versions of (vernacular) texts of the period but almost none of them have their results generally available.<sup>18</sup> The only ver-

---

<sup>15</sup> I am using the term “proto-narrative” in the sense of Bradley and Short, «Texts into Databases: The Evolving Field of New-style Prosopography», .

<sup>16</sup> The “Medieval Greek Database” at King’s College London, see R. Beaton’s introduction in R. Beaton, J. Kelly, and T. Lendari, *Πίνακας συμφοραζομένων του Διγενή Ακρίτη, Σύνταξη Ε*, Herakleion 1995 for some details of the project and J. Kelly, «*Digenis, Livistros* and the computer: technical aspects of the King’s college research project», in N. M. Panayotakis (ed.), *Origini della letteratura neogreca: Atti del secondo Congresso Internazionale «Neograeca Medii Aevi» (Venezia, 7-10 Novembre 1991)* Venice 1991, vol. 1, 129-35 for a description of technical issues. The project was suspended due to lack of funding; the electronic transcriptions and concordances of manuscripts which were produced are nevertheless partly available and in use by our project.

<sup>17</sup> The inclusion of the apparatus criticus is the main outstanding desideratum for linguistic analysis like ours. The TLG excluded the apparatus in its early stages purely on the basis of technical problems of the time (1970-1980); although the technical problems have now been overcome, this policy has not changed, as updating all texts with the apparatus would occupy all available resources. The TLG would be prepared to include the apparatus in its digital library if this is feasible, although copyright restrictions apply and might require negotiations with rights-holders (personal communication by Maria Pantelia, TLG director).

<sup>18</sup> The “Θησαυρός της ελληνικής γλώσσας” foundation (<http://www.thesavros.gr>) published in the year 2001 a CD ROM for the 53rd Frankfurt International Book Fair that includes some rel-

vacular text of our period to our knowledge currently available in a reliable electronic form/edition is *Erotokritos*<sup>19</sup>.

The project “Hellinonmimon” was a research project at the University of Athens that created, at the end of the nineties, a digital library of Greek philosophical and scientific books and manuscripts (1600-1821).<sup>20</sup> This digital library (the texts are only available as digital facsimiles) is of limited relevance for our linguistic description but proves by its mere existence that similar projects are feasible.

Concordances:

The existence of just three published concordances of texts from our period<sup>21</sup> proves in a way that computers have definitely won out over traditional methods in this area of philological activity: there is no need to print concordances when everyone can easily generate them on his or her computer. Editors of texts in Medieval Greek seem not very keen to explore the possibilities of hypertextual concordancing programs<sup>22</sup> and publishing concordances on the Internet.

Register of authors and texts:

In the 1990s the “Greek vernacular manuscript project” at the University of Sydney conducted the most exhaustive survey to date of vernacular Greek texts, manuscripts/scribes and authors.<sup>23</sup> Thanks to the kindness of Elizabeth and Michael Jeffreys the raw databases of the

---

evant texts (without the apparatus and in a proprietary format). The foundation is apparently concentrating its efforts for the time being on other digitisation projects.

The research project “Forms of Written discourse in Byzantine and Modern Greek Diglossia” of the Research Centre 538: Multilingualism at the University of Hamburg compiled (with the active participation of the author of this paper) between 2002-2005 an electronic corpus of 16th and 17th c. narrative texts which has unfortunately not been published due to several reasons.

Eleni Karantzola described an electronic corpus of notarial texts in Jannina and at a more recent conference (E. Καραντζόλα, Α. Τικτοπούλου, and Κ. Τ. Φραντζή, «Γλωσσική Πληροφορία και Ημι-αυτόματη Αναγνώριση», paper given at the conference *Νέες Τεχνολογίες και Φιλολογικές Σπουδές*, Τμήμα Φιλολογίας ΑΠΘ, 24-25/11/2005 (<http://www.lit.auth.gr/index.php?page=13770> [conference programme], accessed 29/06/2006)).

<sup>19</sup> D. M. L. Philippides, D. Holton, and J. L. Dawson, «Του δίσκου τα γυρίσματα. Ο Ερωτόκριτος σε ηλεκτρονική ανάλυση», [CD-ROM], Athens in preparation. Texts relating to our project are certainly available in electronic form with their editors. In some case vernacular texts are available online as part of private/institutional web pages. The website <http://www.early-modern-greek.org> hosts a list of available electronic texts which is being regularly updated.

<sup>20</sup> <http://echo.mpiwg-berlin.mpg.de/content/greeksources> (accessed 18/6/2006) and more exhaustively under <http://www.lib.uoa.gr/hellinonmimon/main.htm> (accessed 20/6/2006).

<sup>21</sup> Beaton, Kelly, and Lendari, *Πίνακας συμφραζομένων του Διγενή Ακρίτη, Σύνταξη Ε*; D. M. L. Philippides, D. Holton, and J. L. Dawson, *Του κύκλου τα γυρίσματα. Ο Ερωτόκριτος σε ηλεκτρονική ανάλυση*, 4 vols., Athens 1996-2001; D. M. L. Philippides, *The SacriŌe of Abraham on the computer. A concordance, word-indexes and stylistic remarks*, Athens 1986.

<sup>22</sup> Like the ones described in Ν. Τουφεξής, «Ο υπολογιστής στην υπηρεσία του εκδότη: σημερινές δυνατότητες και προοπτικές για το μέλλον», in Η. Eideneier, U. Moennig, and Ν. Τουφεξής (eds.), *Θεωρία και πράξη των εκδόσεων της υστεροβυζαντινής, αναγεννησιακής και μεταβυζαντινής δημόσιας γραμματείας. Πρακτικά του διεθνούς συνεδρίου Neograeca Medii Aevi IVa. Αμβούργο 28-31.1.1999* Herakleion 2001, 271-88. The only such electronic concordance available is Philippides, Holton, and Dawson, «Του δίσκου τα γυρίσματα. Ο Ερωτόκριτος σε ηλεκτρονική ανάλυση»,

<sup>23</sup> For some details on the project see M. Jeffreys and V. Doulavera, *Early Modern Greek Literature: General Bibliography* (4.000 items) 1100-1700, Sydney 1998. Cf. also two online prototypes that the project has produced: “Early printing in Greek (1469-1700) (<http://babel.mml.ox.ac.uk/neograeca/index.htm> (accessed 18/6/2006)) and “Early vernacular Greek” (<http://babel.mml.ox.ac.uk/neograeca2/index.htm> (accessed 18/6/2006)).

“Sydney project” have been made available to us; several technical issues have not allowed us to incorporate material from these databases in our system; nevertheless their use as reference material remains extremely valuable for our project.

Electronic versions of lexica:

The online version of the epitomized Lexicon of E. Kriaras has proved, despite some limitations,<sup>24</sup> to be valuable as it allows us to access the material of the lexicon in ways that are not possible in the printed edition.<sup>25</sup>

Online bibliographies:

The Institute for Modern Greek studies (Manolis Triandafyllidis Foundation) maintains and hosts an online database of linguistic bibliography relating to Modern Greek that also covers several areas of Medieval Greek.<sup>26</sup> An extensive online bibliography on post-Byzantine law is maintained by I. N. Arnaoutoglou.<sup>27</sup>

The situation has improved considerably in recent years.<sup>28</sup> Nevertheless this brief review reveals the limitations of electronic texts and databases in the field of Medieval Greek studies. Building an electronic corpus to facilitate linguistic description under these circumstances would have been an extremely complicated endeavour and would require for its execution resources not available at present.

### 3 Database design

#### 3.1 Entities modelling data relating to the Grammar of Medieval Greek

One of the major challenges for the actual database design is the identification of those entities<sup>29</sup> that are of true relevance for the actual research, the definition of attributes that allow for exact and consistent description of these entities and the

---

<sup>24</sup> The main problem is the fact that the second volume of the lexicon (*λαβαίνω – παραθήκη*) is apparently not included in the database while the list of editions and bibliography has not been made available as part of the online database (but see also the next footnote).

<sup>25</sup> As part of the “Ηλεκτρονικός κόμβος για την υποστήριξη των διδασκόντων την Ελληνική Γλώσσα” (<http://www.komvos.edu.gr/dictionaries/dictonline/DictOnLineKri.htm> [accessed 18/6/2006]). This entire site of the hosting “Centre for the Greek language” is at present (June 2006) under active redevelopment; in the course of this the Kriaras lexicon will be upgraded to cover the whole of the printed epitomized version. (Δ. Κουτσογιάννης, Μ. Αραποπούλου, Τ. Γιάννου, Α. Σακελλαρίου, and Κ. Τικτοπούλου, «Η πύλη για την ελληνική γλώσσα του Κέντρου Ελληνικής Γλώσσας», paper given at the conference *Νέες Τεχνολογίες και Φιλολογικές Σπουδές, Τμήμα Φιλολογίας ΑΠΘ*, 24-25/11/2005 (<http://www.lit.auth.gr/public/syn/koutsogiannis.pdf>, accessed 18/6/2006)).

<sup>26</sup> <http://ins.phil.auth.gr/Database/Introduction.htm> (accessed 20/6/2006). Bibliography on Greek Linguistics (covering some aspects of Medieval Greek) can also be found at BL Online. The bibliographical database of Linguistics (<http://www.kb.nl/blonline/>, accessed 3/7/2006).

<sup>27</sup> <http://www.geocities.com/ekeied/> (accessed 30/06/2006).

<sup>28</sup> Many of the resources described here for instance were not available at the time Τουφεξής, «Ο υπολογιστής στην υπηρεσία του εκδότη: σημερινές δυνατότητες και προοπτικές για το μέλλον» was written.

<sup>29</sup> “An entity is usually defined as being an object of interest concerning which information is held. Each distinct category of the information held about a particular entity is regarded as an attribute of the entity.” L. Burnard, «‘Principles of database design», in S. Rahtz (ed.), *Information Technology in the Humanities* Chichester 1987, 54-68.

construction of relationships that link entities meaningfully to each other in a way that truly facilitates the aims of the research. The populated database can be conceived as a set of tables with rows and columns, in which each row corresponds to the entities and each column to the attributes defined in the database.<sup>30</sup> Beyond that, the construction of a relatively complex database tool also means the creation of automatic mechanisms for linking data, formatting and presentation of fields on the computer screen in combination with labels, buttons and other design elements for efficient data input.

– Bibliographical references and textual units

The first objective of our research was a detailed search for all published non-literary and literary texts that, according to our methodology, belong to the corpus of the “Grammar of Medieval Greek”.<sup>31</sup> Table 1 summarizes the entities that have been identified as relevant for this task and have been implemented in our database.

Non literary texts <sup>32</sup>		
<b>PUBLICATION</b>	<b>DOCUMENT</b>	<b>[PERSON]</b>
Author Reference Type Notes ...	Type Date Provenance Notes ...	Name Origin Notes ...

Literary texts		
<b>EDITION</b>	<b>TEXT</b>	<b>[AUTHOR]</b>
Editor Reference Type Notes ...	Title Date Genre Form Notes ...	Name Origin Notes ...

Table 1: Entities relating to text types (simplified for the needs of this presentation)

Journals and other related bibliographical material has been inspected and all relevant references inputted in the database as instances of the entities **Publication** and **Edition** respectively. Characteristics<sup>33</sup> that are relevant to linguistic analysis are recorded with the help of specific attributes.<sup>34</sup>

The distinction between **Edition** and **Publication** for literary and non-literary texts respectively requires some further clarification. Non-literary texts can be found in many different kinds of publications while literary texts are in their ma-

<sup>30</sup> S. Ramsay, «Databases», in S. Schreibman, R. G. Siemens, and J. Unsworth (eds.), *A companion to digital humanities* [Blackwell companions to literature and culture 26] Malden, MA - Oxford 2004, 177-97.

<sup>31</sup> See Lendari & Toufexis in this volume.

<sup>32</sup> For criteria that we apply to distinguish between literary and non-literary texts and related problems see the paper by Lendari & Toufexis in this volume.

<sup>33</sup> Such as type (diplomatic, critical etc.), integrity (full or fragmentary), availability etc.

<sup>34</sup> Held in “fields” according to the terminology of most database software packages.

majority edited according to a specific methodology and for their own sake. Non-literary texts are normally transmitted in a single witness (with the exception of copies, for which see Lendari & Toufexis in this volume) and almost always edited diplomatically with minimal interventions by editors. The exact date and provenance of most non-literary texts is in most cases known and can be recorded accordingly. Because of this last characteristic and also for reasons relating to their linguistic form, non-literary texts are considered to be of high relevance for the “Grammar of Medieval Greek”.<sup>35</sup>

The totally different characteristics of most literary texts need not be listed at great length here. Altogether different categories are required for their description. This fact makes the distinction between **Publication** and **Edition** an obvious one: bibliographical references (and their description) need to be stored in separate tables of the database as they clearly represent different entities.

Editions and publications are not represented only as bibliographical references kept in an electronic database. A large proportion (more than half) of all the 941 publications of non-literary texts so far recorded has been physically acquired and is available in various forms on location in Cambridge. The database functions in this respect also as a catalogue of the archived material and allows us to manage all collected texts efficiently.

The main objective of our research is, however, the collection of excerpts, i.e. text “pieces” illustrating linguistic features of Medieval Greek. Each such piece of text has been identified as a separate instance of the entity **Excerpt**. This entity and its exact manifestations will be described in more detail below, its introduction at this stage is however necessary in order to have a better understanding of the entities describing the individual textual units analysed by our project: **Document** and **Text**.<sup>36</sup>

The entities **Document** and **Text** have been primarily defined from the perspective of database design and in accordance with the aims of the actual research: an **Excerpt** is contained within a **Document** or **Text** as edited in a **Publication** or **Edition**.<sup>37</sup> A consequence of this relationship is that important (for linguistic analysis) characteristics (date, provenance, genre etc.) of **Document** or **Text** are inherited by all excerpts contained within them: a text excerpt (a word or phrase recorded from a document or text) is thought of as not having a date per se but bearing the date of its source (the textual unit from which it has been excerpted).

A further consequence of this relationship, which at the same time represents a main methodological decision for our research, is the way instances of **Document** or **Text** are identified, recorded and described. Excerpting can only be performed once the textual unit (source) has been described and the respective record

---

<sup>35</sup> See the contributions of D. Holton and Lendari & Toufexis in this volume as well as I. Μανωλέσσου, «Οι μη λογοτεχνικές πηγές ως μαρτυρίες για τη γλώσσα της μεσαιωνικής περιόδου», *Λεξικογραφικόν Δελτίον* 24 (2003) 61-88.

<sup>36</sup> For the rest of this paper the term “Document” refers exclusively to non literary documents and “Text” to literary texts.

<sup>37</sup> This basic relationship which is crucial for the organization of data in our database is explained further below.

in the database created. Faced with the abundance of published non-literary documents we made the decision to input in the database – which ultimately means analyse – only those non-literary documents that, according to our criteria, contain enough truly interesting excerpts for the description of Medieval Greek.<sup>38</sup> We will try to be more exhaustive in the field of literary texts, where some significant progress has already been made by the “Sydney project” (see above p. 6).

For each non-literary document we analyse we record, among other characteristics, the type of document, based on a two-level customized classification scheme,<sup>39</sup> its date and its provenance. A verbose identifier is automatically compiled for each document that looks something like this:

— 15th c. (1436), private will; Chandakas/Crete, ed. MANOUSAKAS 1960/61: 146-147 (no. 2)  
[DOC\_25]

Note that the document is linked to the publication it is contained within. Each excerpt recorded from this document remains always linked with this identifier and with all other information about it.

Literary texts are more demanding and require a more complicated approach: this is due to the fact that a literary text may have been edited more than once (and we have to be in a position to use all different editions of a text while excerpting it). This so-called “one-to-many relationship” can be easily handled with the relational model; the association of editions to texts is achieved with the help of a specially devised layout.. For each individual instance of **Text** we record among other details information on form (verse /prose, etc.), genre, a brief overview of available witnesses and, with the help of a combination of fields, information on dating.<sup>40</sup> Each literary text is furthermore assigned a unique abbreviation which will eventually be used for the presentation of examples in the future Grammar, for instance:

<i>Velis.</i> χ	=	Abbreviations used for the four different versions of <i>Velisarios</i> <sup>41</sup>
<i>Velis.</i> ρ		
<i>Velis.</i> Ν <sup>2</sup>		
LIM., <i>Velis.</i> (Λ)		

---

<sup>38</sup> This pragmatic decision was made for practical reasons: it is simply unfeasible in the available time to input each single document in the database. The full list of identified publications will of course be made available to users of the Grammar.

<sup>39</sup> For this we make extensive use of drop-down-menu value lists that guarantee homogeneity of classification.

<sup>40</sup> Comprising a dating of the original text, different dating of eventual versions, alternative dating proposed in the literature and exact dates of composition whenever available.

<sup>41</sup> W. F. Bakker and A. Van Gemert (eds.), *Ιστορία του Βελισαρίου. Κριτική έκδοση των τεσσάρων διασκευών με εισαγωγή, σχόλια και γλωσσάριο* [Βυζαντινή και Νεοελληνική Βιβλιοθήκη, 6], Athens 1988. In creating these abbreviations we decided to give priority to easy recognition over economy of space.

We include in Table 1 two further entities, **Person** and **Author** that have also been partly implemented in our database. As instances of **Person** we record minimal information on persons related with the creation of literary documents. This database table is populated only irregularly and only in cases where an author or scribe is for some reason interesting for the overall project. Authors named by known of literary texts are recorded with minimal information as instances of the entity **Author**. An abbreviated form of the author’s name is used as part of the abbreviation of related texts as in the last example above.

– Linguistic excerpts

An excerpt is defined as an annotated “piece of text” illustrating a particular linguistic feature taken from one of the texts of our corpus. It consists of two elements, the text itself and its linguistic annotation. Text excerpts are of different length (we record in most cases a full phrase containing the feature we are interested in); the smallest linguistic unit recorded is the word. Excerpts either will be eventually used as examples in the published Grammar or are needed in large numbers for the analysis of less well studied linguistic features of Medieval Greek. The structure and organization of excerpts in the database has to follow the principles set out by the research project, in other words excerpts need to be recorded and organized in a way that ultimately provides answers to the given questions of the “Grammar of Medieval Greek” project.

Recording of actual excerpts follows as the third step of the analysis, after collection of bibliographical references and indexing of textual units. In order to meet the criteria set for our research we had to develop three sets of attributes (“fields”) for the recording and linguistic annotation of each excerpt:

- i) Referencing attributes
- ii) Text recording attributes
- iii) Linguistic labelling attributes

With the help of referencing attributes each excerpt can be followed back to its source, i.e. the publication or edition from which it has been taken. We follow the reference scheme used in the original publication or edition keeping in mind that the future user of the Grammar must always be able to verify the accuracy of all quoted excerpts.<sup>42</sup>

Each excerpt needs to be recorded in a way that represents the recorded text exactly as found in the source and facilitates its subsequent precise linguistic analysis. Several fields provide us with the possibility of different ways of quoting (diplomatic, normalised) as well as the recording of additional context. Automatic

---

<sup>42</sup> It is for this reason that we decided to provide line-numbering even in cases where the original publication/edition does not have one, as for instance the edition of *Παλαιά τε και Νέα Διαθήκη* of Ιωαννίκιος Καρτάνος in E. Κακουλίδη-Πάνου (ed.), *Ιωαννίκιος Καρτάνος. Παλαιά τε και Νέα Διαθήκη [Βενετία 1536]. Φιλολογική Επιμέλεια Ε. Κακουλίδη-Πάνου. Γλωσσικό επίμετρο Ε. Καραντζόλα*, Thessaloniki 2000 or the publication of South Italian documents by S. Cusa (ed.), *I diplomî greci ed arabi di Sicilia, Palermo vol. 1. part 1, 1868, vol. 1. part 2, 1882*, repr. Cologne – Vienna 1868-1882.

mechanisms incorporated in the database provide alternative formatting according to several criteria.<sup>43</sup>

Finally, linguistic attributes provide categories for the actual linguistic description of each excerpt; detail of description and applied labels differ according to the level of linguistic analysis. These labels are partly purely descriptive<sup>44</sup> and partly interpretative, i.e. they emerge after a codification of linguistic features of Medieval Greek and questions that need to be addressed for each feature (see more details below).

This kind of multiple data annotation, absolutely necessary for the envisaged linguistic description, makes necessary the separation of excerpts in different database tables according to their source and the level of linguistic analysis. Table 2 summarizes the relevant entities that have been implemented in our database:

Separate entities for non literary documents and literary texts			
PHONOLOGY EXCERPT	NOUN MORPHOLOGY EXCERPT	[VERB MORPHOLOGY EXCERPT]	[SYNTAX EXCERPT]
Source ID	Source ID	Source ID	Source ID
Lemma	Lemma	Lemma	Notes
Phenomenon	Declension	...	...
Notes	Paradigm	Notes	
...	Notes	...	
	...		

Table 2: Entities relating to the recording of text excerpts (simplified for the needs of this presentation)<sup>45</sup>

– Linguistic labelling: guaranteeing homogeneity of annotation

Linguistic labelling of excerpts, e.g. the annotation of each piece of text with categories meaningful for linguistic description, is a matter of linguistic theory and to great extent subject to individual interpretation. Even when the individual linguistic categories (“labels”) are pre-defined, there is in many cases room for different choices resulting from different interpretation.

To address this problem we developed two different approaches. Firstly, we make sure on all database screens that the researchers can use a “notes” field, where additional information or comments can be recorded as free text. These notes are always available together with the excerpt and will be taken into consideration at the stage of composing the Grammar. Furthermore, the researchers are able to record longer statements about the nature of a phenomenon in a particular

<sup>43</sup> This is necessary mainly for literary texts since an excerpt may be found in at least four different locations in relation to a single text: the text of the edition, the apparatus criticus, a transcription of the manuscript witness (provided by an editor in published or unpublished form), or the witness itself (product of in situ research by a member of our project). We have developed different methods of quotation for each of these cases that are created automatically from the combination of several elements.

<sup>44</sup> For instance, we (manually) provide part of speech annotation and lemmatisation of the main word for each excerpt for all levels of linguistic analysis.

<sup>45</sup> Verb morphology excerpt and syntax excerpt have not been implemented in the database at the time this paper was written (June 2006).

text, the peculiarities of a textual source or any other matter. These statements are also available at all stages and can be reviewed at any time.

Secondly, codification of linguistic labels<sup>46</sup> to be used in linguistic description is conducted prior to the actual excerpting of text. The labels are structured in a way that allows their incorporation in a separate database table, together with an overview on dating, distribution, open questions etc. While excerpting primary texts the user of the database is able to consult this reference database, which is also linked to the secondary linguistic bibliography (also available in a separate database table). Through a system of on-screen lists and buttons the user can narrow down the selection of possible labels and apply the label that is most appropriate for the particular excerpt (see Figure 1). During this process he/she can review the provisional description of the actual phenomenon and modify it according to the picture that is emerging from the actual data.

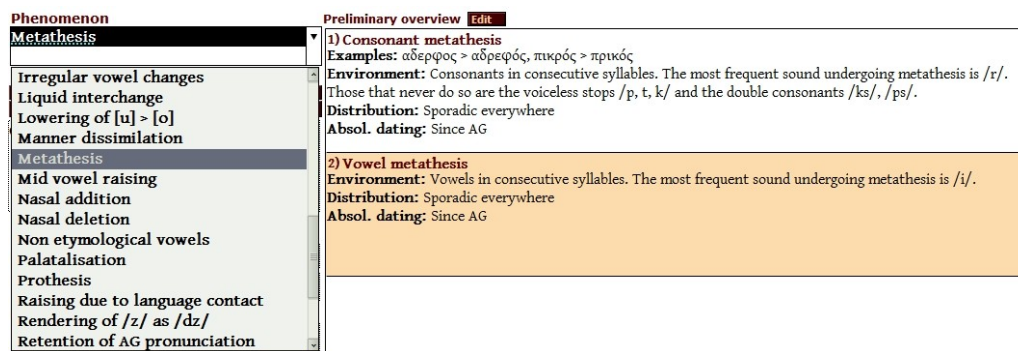


Figure 1: Selection of labels from drop down menus and linked overview of phonological phenomena

### 3.2 Relationships between entities / Structure of database

One of the main advantages of the relational database model is the possibility of linking one row in a table (say a text excerpt) to a row in a different table (a label for a linguistic phenomenon, or a source document or a different excerpt with the same lemma or the same phenomenon). In this way it is possible to store and access a large amount of information pertaining to different entities, logically linked to each other, from each side of the relationship. Such relationships in our database are presented in simplified form in Figure 2:

<sup>46</sup> These differ according to the level of linguistic analysis. Our codification of Medieval Greek phonological phenomena has resulted in 106 different labels, organized under 44 primary (more general) headings. The codification of noun morphology led to 93 different noun paradigms organized in more than (the traditional) three declensions.

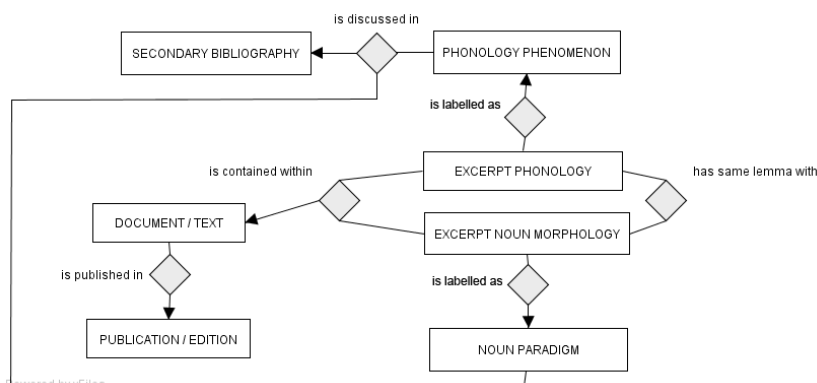


Figure 2: Simplified entity relationship diagram of the database

A phonology excerpt “is labelled as” an instance of a particular phonological phenomenon, which “is discussed in” bibliographical references from the literature. It “has the same lemma with” excerpts recorded for the documentation of noun morphology; each noun morphology excerpt “is labelled as” a particular instance of a noun paradigm, which in its turn “is discussed in” the secondary bibliography. Each excerpt “is contained within” a document or text which in its turn “is published in” a publication or edition. The prerequisite for the formation of such relationships is the existence of compatible attributes on the basis of which relationships are defined. A series of scripts (automated procedures) guarantees that records are linked automatically to each other.

With the help of such relationships it is possible to combine information from different tables in one screen; a typical example can be seen in Figure 3.

Figure 3: Snapshot from the database of non literary excerpts

In this particular case the user is editing a record of the phonology excerpts table based on non-literary documents. The buttons on the upper side of the screen allow him or her to navigate to other sections of the database. Other buttons (“Show document data”, “Show publication data”) navigate to records which are linked to this one, namely the records where the particular document (a letter

from Crete, dated 1 May 1420) or the publication (Manousacas 1962a) are described. The framed boxes with the description of the edition, the date and provenance of the document and the names of the persons involved in this letter are stored in different tables but can be accessed (viewed) in this one because they are linked to each other through a relationship. The tabs in the middle (“Environment”, “Linguistic labelling”) allow the user to access other fields described above that are necessary for recording and annotating this excerpt. He or she can also see an overview of all excerpts in this document (“Excerpts overview”) as well as all excerpts that have been annotated with the same label (“Vowel addition”) (“Phen. overview”).

– Searching, selecting, ordering and grouping of records

Searching is the most obvious advantage an electronic tool has over “traditional” methods of data organization. It is possible in our database to conduct searches that select or omit records using multiple criteria, for instance: to find excerpts of phenomenon X in 16th-century non-literary documents from Crete excluding notarial documents. The search results can then be sorted and grouped (i.e. presented in a more compact form without repetitions) using different criteria. An automatic mechanism creating differently ordered or focused reports has been incorporated into the database tool.

The strengths of the relational model are probably even more apparent in a different way of using the database: browsing through the collected data and navigating between related records in search of answers for a given question. The starting point of this exploration can vary according to the nature of the problem. It is possible, for instance, to isolate excerpts from a particular geographical region or a particular period, create a report on the most frequent phenomena recorded in the database from this region, move to the description of an interesting phenomenon to review some details and move back to the particular excerpt or group of excerpts in search of similarities or differences to the existing description in the reference database (see above). Any instance of the entities displayed in Figure 2 can function as starting point for the user who can choose the path that suits him/her best for the exploration of his working hypothesis.

A further aspect of the same organizing principle relates to feedback that the system provides on similar material that has already been inputted in the database. Automated scripts count the instances sharing basic characteristics (linguistic label, grammatical markup) and make this information visible to the user on the computer screen in the form of interactive buttons. Figures 4 and 5 illustrate this principle.

**Phenomenon**

Metathesis ▼

29 non lit. with same phenomenon  
3 excerpts with same phenomenon in this text

Consonant metathesis

Figure 4: Feedback on phon. phenomenon

In Fig. 4 the researcher assigned the label “Consonant metathesis” (grouped under the subordinated heading “Metathesis”) to a particular excerpt (*σχόρφας Poulol. 559 app. cr. [CPV]*); there are 29 excerpts from non-literary documents in the database that have been labelled with the same phenomenon (“Consonant metathesis”), which can be viewed with a click on this button. Three excerpts from the same text bear the same label and can also be viewed on the same principle. shows the same principle deployed for an overview of excerpts that have been labelled under the same lexical lemma. The buttons allow access to the 19 excerpts from literary texts and the 22 excerpts from non-literary documents with the same lemma. The researcher can review the lemmatisation process at any time and check quite easily whether a particular phenomenon pertains to a single lexical item and, if so, adjust his/her excerpting strategy accordingly.

ἦνιε / εἶνιαι (1445, deed of assignation/Cyclades, ed. LAMBROS 1907b: 468.) εἶναι  
Dental palatalisation [s, z, n, l] > [ʃ, ʒ, ɲ, ʎ]

Reference & text recording Environment Linguistic labelling Excerpts overview Additional evidence Phen. overview

Text: ἦνιε Normalized: εἶναι

θ γ δ ρ κ λ ρ ρ zero

Phoneme: ɲ

Phon. real.: ɲ

Grapheme: νι

Syllable position:  initial  medial  final  monos.  Other...

Syllable stress:  yes  no  Other...

Word boundary:  yes  no  Other...

Lemma: εἶναι

22 non lit. with same lemma  
16 lit. with same lemma

Next step: Labelling

Figure 5: Feedback on lemma

The art of browsing through the material in search for answers is also essential during the excerpting process and the stage of composing the Grammar in book form. Data stored and organized in this way will in all probability allow the researcher who has build up his/her own mind about a particular feature from secondary bibliography and his/her own reading of primary texts to have a broader picture based on actual material. He or she can engage in a mental discussion with other opinions expressed in the form of notes or statements by other members of the research team, locate emerging patterns, inconsistencies or peculiarities and merge all these into the actual section of the Grammar explaining the particular phenomenon.

### 3.3 Technical specifications

The database is being implemented with the use of FileMaker Pro software. FileMaker has been chosen because of the following characteristics: cross-platform compatibility (Windows and Mac OS); full Unicode implementation and support

for text in Greek (both polytonic and monotonic); good networking capabilities; almost no restrictions on the amount of data stored in fields and tables; export of data in non-proprietary formats (XML). Furthermore FileMaker Pro provides the tools for excellent visual representation of collected data, allowing for convenient everyday work in front of the computer screen.

#### 4 Conclusions

We hope to have shown that the creation of a database for a research project like ours is not simply a technical matter; it is closely interlinked with the overall aims of research, it reflects and follows the project methodology and can only be built after a thorough conceptual analysis of the primary data and careful mapping of project workflow. It functions as an efficient research tool because it combines two characteristics: material stored in distinct tables but linked to each other into a single logical entity with the help of truly meaningful relationships and flexibility in visual presentation of the collected material on the computer screen and on paper reports. The database functions as a complex editing tool; the biggest part of our research in the first two stages described above is being, and will continue to be, conducted with its help.

Every custom-built electronic tool is in danger of becoming obsolete once the technologies on which it is based are superseded or if the people who constructed and used it have not documented it properly. Our database, although based on proprietary software, is in no immediate danger: the software we are using is still being developed and there are no signs of it being discontinued. We use industry standards for the encoding of Greek (Unicode) and we plan archiving the material stored in our databases in a non-proprietary format (XML) in the long run. Documentation, although time consuming, is being written, if nothing else in papers like this one.

The author of this paper is not strictly speaking a computer expert but has a philological background. It is debatable if a computer expert on his own could create a customised database tool like the one described here; many philologists have experienced such problems, mainly because they were not able to “communicate” properly with computer experts. While the non-computer-expert philologist might have difficulties in creating a more complex system (for instance a web-based front end for the presentation of material based on a relational database), only such a person can or should formulate research questions, conduct a thorough conceptual analysis of the primary material based on these questions, and ultimately set targets that the electronic tools have to fulfil. The next step can be created by computer experts under his/her supervision, if he or she lacks the necessary expertise. We hope that this paper describing our efforts in this area will motivate others to create similar electronic resources of their own.