A corpus of literary papyri online:
the pilot project of the medical texts via SoSOL

Nicola Reggiani
Università degli Studi di Parma
Ruprecht-Karls Universität Heidelberg

Since a long time, the study of ancient Greek papyri, with its fundamental outcomes in every historical discipline, has developed a wide, more and more sophisticated network of electronic tools, useful – not to say unavoidable – for the analysis and publication of the multifarious data that such documents offer to the scholars. From digital graphics to the integrated data banks, the scenario of the so-called “digital papyrology” is always in progress.

The recent developments of the project “Integrating Digital Papyrology”, which through the Papyrological Navigator is moving towards a more and more comprehensive database of Graeco-Roman texts written on papyrus, pottery, wood/waxed tablets, parchment (both documentary and literary pieces: see below), have offered the best frame for the project “Corpus dei Papiri Greci di Medicina [CPGM] Online” [“Corpus of the Greek Medical Papyri Online”], started in December 2011 at the University of Parma (Italy) by a team led by Prof. Isabella Andorlini with the aim of creating a textual database of ancient Greek texts on papyrus dealing with medical topics. The encounter of Prof. Andorlini’s long-lasting interest in ancient medical texts with the new, powerful, yet easy potentialities provided by the SoSOL platform on which the Papyrological Editor is based (see below) has been the starting point of this project, together with the awareness that a textual database of Greek medical papyri currently does not exist.

Due to their mainly “borderline” nature, most medical texts on papyrus are indeed included neither in literary nor in documentary databanks: and not even in repositories such as the

---

1 Cf. Andorlini 2012, 131-134; Reggiani 2012b.
3 For an overview on Greek medical papyri cf. e.g. Marganne 1981; Hanson 1985; Andorlini Marcone 1993; Marganne 1995. For a general bibliography on this subject, see the web page Medici et Medica edited by M.-H. Marganne at the address http://promethee.philoulg.ac.be/cedopal/Bibliographies/Medica.htm.
4 Cf. Andorlini Marcone 1993; Andorlini 1997a; Andorlini 2001; Andorlini 2004; Andorlini 2009. On the previous project “Corpus dei Papiri Greci di Medicina” see Andorlini 1997b.
Catalogue of Paraliterary Papyri\textsuperscript{5}, which shows – after a search for the key-string “medical” – no more than five results. Yet medical papyri form a corpus of about 300 (published) items of different size and complexity, the study of which as a whole is of the utmost importance to shed light on many aspects of the ancient cultures – and of our own world. They spread from the category of documentary texts (anything written without literary purposes, such as private letters discussing medical matters, annotations by doctors, and above all recipes or collections of recipes – the so-called receptaria\textsuperscript{6}) to that of literary works (chiefly medical treatises, of different subject – e.g. anatomy, pathology, surgery, pharmacology – and structure, both by known authors and adespota\textsuperscript{7}), through that of para-literary or subliterary pieces, mostly didactic texts like technical-practical handbooks, herbals, catechisms, school-books\textsuperscript{8}. A systematic and annotated tool for this kind of documents will therefore be necessary and useful.

The framework architecture of the project is that offered by the Papyrological Editor, the powerful collaborative, peer-reviewed editing interface lying behind the Papyrological Navigator, a tool which is already well-known to any papyrologist\textsuperscript{9}. In the Papyrological Editor, the digital text of any papyrus (either taken from the previous Duke Data Bank of Documentary Papyri or entered directly in the Editor itself) is linked to the metadata provided by the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV), and often also by the Advanced Papyrological Information System (APIS), and possibly to a commentary and/or a translation, in order to provide an essential “digital edition” of the document. As is known, the structure of such digital editions is based on the TEI/EPIDOC XML standard, which marks any meaningful papyrological feature of the text (lines, gaps, supplements, errors, etc.) with particular machine-readable tags; the contribution of anyone is made possible and easier thanks to the SoSOL platform, which deploys the special markup language “Leiden+”, much closer to the usual way in which papyri are published in print editions.

As already mentioned above, the CPGM project relies upon the SoSOL framework: thanks

\textsuperscript{5} http://cpp.arts.kuleuven.be, ed. by M. Huys and Collaborators.
\textsuperscript{6} On the recipes see e.g. Gazza 1955; Andorlini 1981; Harrauer - Sijpsteijn 1981; Andorlini 2007. On other kinds of documentary medical texts see Marganne 2006.
\textsuperscript{7} Cf. e.g. Andorlini 1984; Marganne 2008.
\textsuperscript{8} Cf. Andorlini 1999b.
\textsuperscript{9} On these new electronic resources see Sosin 2010; Andorlini 2012, 134-138; Bagnall 2012; Reggiani 2012b, 98-101.
to Prof. Josh Sosin, one of the leaders of the SoSOL project, we have been provided with a special environment (“ParmaMed Community”) in which it is possible to conduct the work of editing medical papyri via SoSOL, submitting them to our own editorial board, and managing the first step of the Papyrological Editor peer-reviewed process: its dashboard currently displays 29 fully verified and refined texts. Since SoSOL platform was initially designed for documentary texts only, more complex textual features such as, e.g., critical and diacritical signs (see below), cannot be fully represented through the actual Leiden+ markup tags, though being extremely relevant for the comprehension of the texts. Enlarging the database to literary papyri in general, and medical texts in particular, means considering some improvements in Leiden+, as well as in the metadata (the latter easily obtainable from literary databanks such as the Leuven Database of Ancient Books – LDAB – and Trismegistos)10. SoSoL is currently developing to include also (para)literary texts11, and one task of our pilot project is to point out which categories may be useful to take in consideration in order to integrate the actual markup.

Thus the construction of the CPGM Online requires both the digitisation of the texts and a re-thinking of the whole architecture of the database, in order to fit the requirements of medical papyri as technical texts. The latter task may be performed by means of the tools provided by corpus linguistics, that is the study of language as expressed in (electronic) corpora (or samples) of texts12. Integrating SoSOL with corpus-specific information means adding more annotation levels as (XML?) tagging layers, where an annotation (that is the

10 Cf. Reggiani 2012a, 142.
12 This might arise some problems, such as the question whether medical papyri can be considered as a corpus or not. According to corpus linguistics, a corpus is “a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language” itself (Sinclair 1996). It should be a body of texts carefully sampled to be representative of a language, or of a language variety: in other words, a corpus is used as a sample of a larger population to do statements on it. To fulfil this task, the sample should not be random (producing wrong generalisations) nor stratified (based on the knowledge of the variables that influence the population), but representative, based on the known percentage of each variant of the population (e.g., in the corpus “surviving classical Greek texts”, we know the percentage of each literary genre: cf. Biber 1993). As many historical corpora, medical papyri have survived randomly, thus forbidding any authoritative generalisations – yet from the perspective of their whole they can be considered not as a “random sample”, but rather as the whole “population” of the (extant) texts with which we are dealing. Therefore they are not taken as a representative sample of a larger group (such as literary or paraliterary texts), but as a population itself, that we use as data for our research questions. In general, on corpus linguistics, see Lüdeling - Kytö 2008-2009.
application of metadata to a text) is intended as an interpretative enrichment of the available data (in this case, the papyrus texts) through categories\textsuperscript{13}. So far, the following categories have been pointed out as potentially useful sources of information concerning Greek medical papyri: *Parts of speech; Kinds of errors; Technical terms; Abbreviations; (Dia)critical marks and punctuation; Marginalia; Formulaic features; Cross-references; Quotations* (for a sample case see at the end of the article).

**Parts of speech (PoS) and Syntactic annotation.**

This annotation layer is very common in corpus linguistics and marks the syntactical value of each word in the texts. It is helpful to perform basic syntactical and lexical analysis of the entire corpus\textsuperscript{14}. Building a syntactically annotated corpus (a so-called *treebank*) would be very useful to a deep analysis of the textual content of medical papyri.

**Kinds of errors.**

Errors in writing are a very common feature of many Greek papyri, of both literary and documentary type, mainly due to the influence of the spoken language on the scribe. We may distinguish *vowel exchanges* (e.g. $\epsilon \leftrightarrow \eta$ or $\alpha, \iota \leftrightarrow \upsilon$ or $\epsilon, \omega \leftrightarrow \sigma$ or $\omicron$), *consonant exchanges* like $\tau \leftrightarrow \delta$, $\kappa \leftrightarrow \gamma$, or $\lambda \leftrightarrow \rho$; *broken diminutives* such as $\sim \gamma$ for $\sim \omicron$; *grammatical mistakes* consisting in wrongly declined nouns or wrongly conjugated verbs; *missing letters*. A good example, among many others, can be provided by GMP II 10, a letter requesting medical tools of the 6\textsuperscript{th}-7\textsuperscript{th} century AD\textsuperscript{15}. After having defined a *Normalization* layer (*i.e.* the correct word spellings), this level of annotation may provide plenty of information about many relevant aspects of the writing habits or of the cultural environment in which medical texts were produced. However, error annotation is problematic, because the definition of an error itself is problematic. First of all, it is clear that an error can only be annotated if a

\textsuperscript{13} Cf. Lüdeling 2011.

\textsuperscript{14} On part-of-speech tagging (and tokenization) see Schmid 2008.

\textsuperscript{15} As an example of error annotation (GMP II 10, 1-5): the first row is the *Tokenization* layer (which divides the single words); the second row is the *Normalization* layer (correct word spelling); the third row is the (multiple) *Kinds of error* layer [VE = vowel exchange; BD = broken diminutive; ML = missing letter] (see also below). On the features of this document see the edition by C. Fischer Bovet, in Andorlini 2009, 157-165.

| πέμψον | μοι | τῷ | ὀθόνι | μου | καὶ | τῷ | κυάθι | μ[σ] | καὶ | τῷ | τρω[σ] | σφραγί | μ[σ]ν |
|--------|-----|-----|-------|-----|-----|-----|-------|-----|-----|-----|-------|-------|-----|-----|
| τῷ     | ὀθόνιον | τῷ    | κυάθιον | τῷ    | σφραγίδιον | Μοι | 
| VE     | BD  | VE  | VE+BD | VE   | ML+VE+BD |
“correct” version of the utterance is assumed, and it is often difficult to state which may be
the “correct” form of a variant. Then, error annotation may lead to define multiple layers, on
the basis of multiple kinds of errors – which means further databanking issues16.

Technical terms.
After having defined a Lemmatization layer (that is the reduction of each word to its
lemma), this kind of annotation will pinpoint the technical terms recurring in Greek medical
papyri, to be distinguished among the main branches of medicine: anatomy (parts of the body,
both internal and external); botanics (mostly plants used as ingredients17); mineral items (used
as ingredients as well); pathology (diseases and symptoms); pharmacology or materia medica18 (pharmaka, that is remedies); regimen (terms related to diet); surgery (instruments
and other terms related to surgical operations19); containers; instruments20; quantity (units of
measurement or terms related to the quantity of the ingredients)21. The study of the technical
terms in medical papyri is part of the related project Medicalia Online, coordinated by Prof.
Anastasia Maravela at the University of Oslo, and will be developed also as an online
interactive glossary connected to the main database22.

16 On this topic see Reznicek - Lüdeling - Hirschmann 2013. I am very grateful to Prof. A. Lüdeling for having
kindly given me advice and reference about this topic on the occasion of the Summer Institute “Working with
Text in a Digital Age” (Tufts University, Medford, MA, July-August 2012).
17 For a general bibliography about ingredients in Greek-Egyptian pharmacology see the web page
Pharmacopoea Aegyptia et Graeco-Aegyptia by M.-H. Marganne and P. Koemoth, at the address
http://promethee.philo.ulg.ac.be/cedopal/Bibliographies/IPHARMEG.htm. On botanical vocabulary, in
particular, see Fausti 1997.
18 On materia medica in Graeco-Roman Egypt see e.g. Arnold 1986.
19 On this topic see Marganne 1998; some specimina of surgical vocabulary can be found in Ghiretti 2010.
20 Some specimina of instrumental vocabulary can be found in Ghiretti 2010.
21 The difference between containers and units of measurement is sometimes very little, in particular as regards
liquid capacities (cf. Reggiani, forthcoming): this might arise some further issues in annotating such specific
cases.
22 This part of the project will be presented by Prof. Maravela at the 7th Papyrology Workshop “Papyracea Nova
Reperta” at the University of Parma (May 22nd, 2013) and by Dr. Isabella Bonati (who is working on a PhD
dissertation about technical terms in Greek medical papyri, at the University of Parma) at the 27th International
Congress of Papyrology (Warsaw, July-August 2013). For the moment see the web page
http://www.papirologia.unipr.it/CPGM/glossary.html.
**Abbreviations.**

This layer is thought to provide a specific categorization of the texts with relation to the use of abbreviations or sigla, in the framework of a classification proposed by two members of our group\textsuperscript{23}. Abbreviating a word or a phrase – that is reducing it to a conventional grapheme – is a very common feature in Greek medical papyri (see also below), and it would be very interesting to state if regular patterns are followed. Abbreviations mostly pertain to the final letters of a word, because they could be inferred by the context, but we have instances of words abbreviated in the middle, or just of symbols. While the semantics of the word(s) involved in the abbreviative phenomena may be retrieved by performing a combined search within the *Abbreviations* and the *Technical terms* or *PoS* layers, the former will be devoted to distinguish the different kinds of abbreviation, and in specific: *suspension* (omission of one or more letters at the end, without any marks); *horizontal stroke* (omission of one or more letters at the end, marked by a horizontal stroke above the word); *vertical stroke* (omission of one or more letters at the end, marked by a vertical stroke at the end of the word itself), *slanting stroke* (omission of one or more letters at the end, marked by a diagonal stroke at the end of the word itself); *sinusoid* (omission of one or more letters at the end, marked by an S-shaped sign); *superscription* (omission of one or more letters at the end, marked by one letter written above the end of the word itself); *discontinuous abbreviations* (omission of one or more letters not – or not only – at the end of the word); *monograms* (superimposition of two or more letters, showing various degrees of overlapping or merging); *symbols* (graphic mark in which there are no recognizable letters); *brachigraphy* (very common, non-technical words, such as conjunctions or articles, abbreviated by short marks).

**(Dia)critical marks and punctuation.**

Critical and diacritical signs are defined as graphic markers that enhance the comprehension of the text; generally speaking they can be compared with punctuation, the purpose of which is to divide words or parts in a text\textsuperscript{24}. To be tagged within this layer there may be the *paragraphos* (a horizontal bar, usually put between two lines, dividing two different sections of a text), the

\textsuperscript{23} Centenari - Iori, forthcoming. In general on abbreviations in Greek papyri see Blanchard 1974 (documentary texts) and McNamee 1981 (literary texts).

\textsuperscript{24} On punctuation and (dia)critics in Greek papyri, both documentary and literary, see Turner 1987, 7-13. This particular issue will be developed also within a research plan about punctuation in ancient documents (inscriptions, papyri, manuscripts) recently started by the author of this article together with Prof. Michèle Brunet (Université Lumière - Lyon 2) and Dr. Giuseppe G.A. Celano (Università degli Studi di Pavia).
stigme (a dot, usually marking the beginning of a new sentence), the dicolon (double dot), the diple (arrow-shaped sign marking relevant features in the text), the filler mark (usually a horizontal stroke, put at the end of a line in order to fill blank spaces and to keep the line aligned to the right), the diaeresis (often used over a ι or a υ to mark the beginning of a word), the supraline (very often used to mark figures expressed by alphabetic characters, e.g. η̅ for 8 – non to be confused with the abbreviations marked by a horizontal stroke: see above). We can include in the critical marks also ekthesis (outdented line) and eisthesis (indented line) and even blank spaces, since they are all graphical devices used to highlight particular lines or parts of a text (e.g. the first line of a recipe, or the question in the erotapokrisis catechisms – a good example of this is P.Ross.Georg. I 20, 2nd century AD) 25. The layout of the document itself, together with any other kind of graphical devices such as marks and abbreviations (see above), is an essential part of the medical text as a technical work 26: whence the need to have one or more annotation layers devoted to this.

**Marginalia. Formulaic features.**

Marginalia are letters, words or sentences written in the margins of a papyrus, i.e. outside the regular text layout. There could be lots of marginal notes, with very different meanings (scholia, annotations, etc.) 27; a very interesting instance is that of the numbers written in the middle of the top margin, which can refer to the “page” (in a codex, or in a parchment notebook such as, e.g., PSI VI 718 = SB XXVI 16458, 4th century AD) or the column (in a papyrus roll).

Another important trait of medical papyri is that of the formulaic expressions, well known and attested in particular – but not only – in the prescriptions. Typical headings of recipes, recurrent verbs or expressions (e.g. λάβε”, παῦϲαι, πίνειν, δὸϲ...), the question-and-answer format in the erotapokrisis catechisms (e.g. Πρόϲ τι κτλ...), are all possible instances of “formulas” in medical papyri, that tend to acquire the status of technical stamp of that specific genre 28.

**Cross-references. Quotations.**

An annotation layer about cross-references can help to pinpoint references to parallel texts

---

25 On the issue of indentations and blank spaces as punctuation see Turner 1987, 8; Cribiore 1996, 83. For the erotapokrisis or question-and-answer format see Zalateo 1964 and Ieraci Bio 1995.


27 See e.g. McNamee 1992.

in other papyri or even in literary medical texts (prescriptions, instructions for medicines, and so on), which are a very common feature in both documentary and literary medical works\textsuperscript{29}, being \textit{e.g.} also a possible tool to further investigate the theme of doctor’s literacy\textsuperscript{30}.

Unlike cross-references, quotations are exact recalls of passages taken from other texts, usually of literary kind\textsuperscript{31}. They are to be distinguished between possible citations from medical authors (either known or unknown) in other (para)literary texts (even \textit{adespota}) and “non-voluntary quotations”, consisting in the fragmentary status of most of the documents\textsuperscript{32}, which very often preserve just excerpts of a (literary) work. Such accidental “quotations” are however very important, since they may attest interesting variants: this is the case, for example, of P.Aberd. I 24 (= GMP I 1), preserving a fragment of the hippocratic treatise \textit{De fracturis} (§ 37), which attests readings divergent from codices (ll. 4-5 and 11-12) as well as the “original” use of ionic dialectal forms (l. 14).

These introductory notes show all the complexity, but also the potentialities, of a deeply annotated textual database of Greek papyri. This reflects complexity and potentialities of the medical texts as well, which makes the CPGM Online project a strong test case for the forthcoming developments of the \textit{Papyrological Editor & Navigator}: as Prof. Sosin said, “this is an exciting and somewhat scary new step for the field”\textsuperscript{33}. Some questions remain still open for the moment: for example, what to do with “broken” words such as not supplied words, or words with alternative supplements; or whether \textit{SoSOL} will be able to support multiple tagging / annotation layers or not. But the first stage will be of course the basic digitisation of the texts, to be included – together with translations and short commentaries – in the main database of the \textit{Papyrological Navigator}. Linked resources such as the interactive glossary \textit{Medicalia Online} and a thematic bibliography will follow, with the hope of providing a real “digital library” of ancient texts in the field of the studies on ancient medicine.

\textsuperscript{29} See \textit{e.g.} Andorlini 1999a.
\textsuperscript{30} On which see Hanson 2010; cf. also Andorlini 2006, 153-158.
\textsuperscript{31} On quotations in medical texts see \textit{e.g.} Marganne 2010.
\textsuperscript{32} On fragmentation and the medical papyri see Hanson 1997. On the issues of representing fragments in digital editions see Berti - Romanello - Babeu - Crane 2009 and Berti 2011, and also the web site “Fragmentary Texts” (http://www.fragmentarytexts.org) edited by M. Berti.
\textsuperscript{33} Sosin 2010.
The CPGM Online Project

Web Site
http://www.papirologia.unipr.it/CPGM

Scientific Editor, Coordinator of the Project
Prof. Isabella Andorlini (Università degli Studi di Parma)

Technical Editor
Dr. Nicola Reggiani (Università degli Studi di Parma / Ruprecht-Karls-Universität Heidelberg)

Team Members
Dr. Andrea Bernini (Università degli Studi di Parma / Ruprecht-Karls-Universität Heidelberg)
Dr. Isabella Bonati (Università degli Studi di Parma / Universitetet i Oslo)
Dr. Margherita Centenari (Università degli Studi di Parma)
Dr. Francesca Corazza (Università degli Studi di Parma / Humboldt-Universität zu Berlin)
Dr. Giulia Ghiretti (Università degli Studi di Parma)
Dr. Luca Iori (Università degli Studi di Parma)

Partners
Dr. Rodney Ast (Ruprecht-Karls-Universität Heidelberg)
Prof. Roger S. Bagnall (Institute for the Study of the Ancient World, New York)
Dr. James Cowey (Ruprecht-Karls-Universität Heidelberg)
Prof. Ann E. Hanson (Yale University, New Haven, CT)
Dr. David Leith (University of Cambridge, UK)
Prof. Anastasia Maravela (Universitetet i Oslo)
Prof. Josh D. Sosin (Duke University, Durham, NC)

Bibliography

ANDORLINI 1981
ANDORLINI 1984

ANDORLINI 1997a

ANDORLINI 1997b

ANDORLINI 1999a

ANDORLINI 1999b

ANDORLINI 2001

ANDORLINI 2004
ANDORLINI 2006

ANDORLINI 2007

ANDORLINI 2009

ANDORLINI 2012
I. ANDORLINI, Edizione e ricostruzione digitale dei testi papiracei, con appendice di N. Reggiani, in N. PALAZZOLO (a cura di), Diritto romano e scienze antichistiche nell’era digitale (= Collectanea Graeco-Romana 10), Torino 2012, 131-146.

ANDORLINI MARCONE 1993

ARNOLD 1986

BAGNALL 2010

BAGNALL 2012
R.S. BAGNALL, The Amicitia Papyrologorum in a Globalized World of Learning, in: P.
BERTI 2011

BERTI - ROMANELLO - BABEU - CRANE 2009

BIBER 1993

BLANCHARD 1974

CENTENARI - IORI forthcoming

CRIBIORE 1996
R. CRIBIORE, Writing, Teachers, and Students in Graeco-Roman Egypt (= American Studies in Papyrology 36), Atlanta, GA, 1996.

FAUSTI 1997
GAZZA 1955

GHIRETTI 2010

HANSON 1985
A.E. HANSON, Papyri of Medical Content, Yale Classical Studies 28, 1985, 25-47.

HANSON 1997

HANSON 2010

H. HARRAUER - P.J. SIJPESTEIJN 1981

IERACI BIO 1995

LÜDELING 2011
LÜDELING - KYTÖ 2008-2009

MARGANNE 1981

MARGANNE 1995

MARGANNE 1998

MARGANNE 2006

MARGANNE 2008

MARGANNE 2010

McNAMEE 1981
K. McNAMEE, Abbreviations in Greek Literary Papyri and Ostraca (= Bulletin of the

McNAMEE 1992
K. McNAMEE, Sigla and Select Marginalia in Greek Literary Papyri (= Papyrologica Bruxellensia 26), Bruxelles 1992.

REGGIANI 2012a
N. Reggiani, Progetti digitali “in corso” all’Università di Parma, appendice ad Andorlini 2012, 141-142.

REGGIANI 2012b

REGGIANI forthcoming

REZNICEK - LÜDELING - HIRSCHMANN 2013

SCHMID 2008

SINCLAIR 1996
A very tentative example of annotation of a medical papyrus: GMP II 10.

ID = Identifier; T = Tokenization; PoS = Part of Speech; H = Syntactic head; R = Syntactic relation [H + R + ID = Syntactical annotation]; N = Normalization; E = Kinds of errors [VE = vowel exchange; BD = broken diminutive; ML = missing letter; EC = error corrected; CE = consonant exchange]; L = Lemmatization; TT = Technical terms; A = Abbreviations; M = (Dia)critical marks and punctuation [The other levels of annotation do not apply to this instance. Note that PoS and syntactical taggings are simplified for reason of space]. I am very grateful to Dr. Giuseppe G.A. Celano for his essential help in the grammatical/linguistic annotation.

“Send me my bandage, and my ladle, and my round (?) seal, and the doctor’s bag, and my scalpel, and the wooden spoon”