The Corpus Corporum,  
a new open Latin text repository and tool*

I. Aims and background

This article introduces a new online resource: the Corpus Corporum – reposito-
rium operum Latinorum apud universitatem Turicensem (www.mlat.uzh.ch). Some
three years ago the idea was born from my linguistic research to create an open and
non-commercial Latin text meta-collection. Now it displays Latin texts I can find in
an adequate format from Latin’s entire life-span from antiquity to (occasionally) the
20th century. At the end of 2014 there were some 130 Million words on display from
nearly 6’000 texts. Experimentally some Greek texts are also included.

The project stands out by its methodology: it makes a clear distinction between the
texts themselves – which are encoded in a subset of the standard format TEI xml and are
freely distributed – and the actual database into which such texts are loaded1. The latter
is being developed by Max Bänziger, who specialises in the development of web applica-
tions in philology and palaeography2. Once we will have implemented our basic ideas,
the software will be released under a GNU-license. Such an approach will guarantee
the long-term availability of these digitised texts, in contrast to distributing them within
closed-source software which will become unreadable quickly. Another novelty in our
approach is that in order to be able to study thematically defined subgroups of texts, our
collection consists of several single corpora which can be searched and studied sepa-
rately, but of course also collectively (hence the project’s name). The start page of each
sub-corpus shows data about the authors and the origin and scope of its texts in general.
These days many institutions are developing textual corpora and it seems important to
coordinate international collaboration in order to avoid doing work several times. Thus
I am currently in contact with the SISMEL led initiative for a Mediaeval Latin Digital
Library, with Gregory Crane who is working on a new large collection of Mediaeval
Latin texts in Leipzig3, and the Comphistsem at the University of Frankfurt a. Main4.

* This project has been funded by the Swiss SBFI within Cost action IS1005.
1 This methodology is inspired by Gregory Crane’s pioneering Perseus project (http://www.
perseus.tufts.edu).
2 Cf. www.monumenta.ch.
This project is very low-key: a small grant by the Swiss state through COST-action IS 1005 enables me to finance the programming, whereas I am preparing and formatting the text files in my spare time. This way I intend to show how much can be reached by few people and little money – but on the downside this means that I cannot hand-check our texts’ formatting and have to rely on automated methods to deal with texts throughout.

People with TEI xml Latin text corpora are invited to have them displayed in our collection (according to their wishes as far as technically possible for us). This makes them searchable in complex ways and helps secure their long-term availability. First examples of recently edited texts in our collection are those of the Richard Rufus Project at Stanford University (led by Rega Wood) and three of my own editions. More texts, such as the Corpus Grammaticorum Latinorum (http://kaali.linguist.jussieu.fr/CGL/), or Loris Petris’ (University of Neuchâtel) new edition of Michel de L’Hospital’s Carmina, will follow soon.

II. Basic functions

The texts are presented in five frames on our web-page (www.mlat.uzh.ch, cf. ill. 1). On the left, corpus, author and text names are always displayed in the top frame which serves as main orientation for the user, the middle frame shows the further breakdown
of the work one has navigated to (down to chapters and paragraphs), and on the bottom there are search and dictionary look-up tools, which are always applied to the level one has navigated to. On the right-hand side in the main frame the actual Latin text is displayed and below it further data that can be requested by the user: apparatus and reference entries, whose presence is indicated by asterisks or superscript numerals in the text, and by clicking a word it gets grammatically resolved and its possible lemma(ta) are presented in some dictionaries. This will make reading texts much faster for people who are not yet very advanced in Latin.

Ill. 2 shows a typical page view inside a text (here one from the work Philosophiae recentioris versibus traditae libri sex by the Croatian author Benedictus Stay). The first word in the text (inscius) has been clicked and dictionary entries for it are displayed at the bottom. Instead of navigating through corpus and author to a work, the user may also directly navigate to a list of all works by an author (which may be scattered over different corpora) or a specific work by using the Ajax-powered search boxes on the main page (cf. ill. 1).
III. Standards and some technical details

In this project we use exclusively free and open software (mostly GNU General
Public License): Ubuntu Server 14.04 LTS runs on a virtual machine hosted by the
University of Zurich, the data is stored in a MySQL database, Sphinx is used as search
server, Apache2 as web server, and PHP as programming language.

All our texts are encoded in UTF-8 unicode and stored in xml-files. We use a subset
of the standard format TEI (http://www.tei-c.org/) for our text encoding within xml.
A TEI file consists of a header which contains meta-information about the text, its
digitising and formatting, and the actual text with an xml mark-up. From the header
the server takes information about the following facts it displays and uses internally
for searches: (1) the author and his lifetime, e.g. Lambertus de Monte, whose year of
birth is unknown and who died in 1498. (2) The text’s name, e.g. De salvatione Aris-
totelis, optionally its date of composition, its editor(s), bibliographic information about
the book it was published in (or in case of online editions, the online source), occasion-
ally the manuscript(s) or early prints it was transcribed from and similar data. (3) A
unique ID which we assign to each text made up of six letters for the author’s name, a
dot and up to seven letters for the work’s name. So the above text has the ID: LamMon.
DeSaAri. In case of several editions of the same text, they are distinguished by numbers
at the end. This ID enables third parties to link directly to a text, e.g. as http://mlat.uzh.
ch/?c=4&LamMon.DeSaAri (the number « c=4 » indicates the corpus the text belongs
to). (4) A statement about the text’s orthography as a normalization tag. This is of use
for the text’s lemmatisation. Depending on whether the text is in « Classical Latin ortho-
graphy » or not, a form like recte may or may not also be interpreted as rectae.

The text itself must have its internal structure defined by divN-tags (N being a posi-
tive integer, the lower the number, the higher-level the structure) in order to be load-
able into the database, optionally also by milestone-tags for small units (paragraphs or
biblical verses). If present, their attributes n="N" or id="X" for numbers or abbreviated
names (such as prol. for a prologue) of the respective books, chapters, or paragraphs
are used to display them in search results. The div-tags must be followed by a head-
tag which defines the title to be displayed in the middle left frame. Besides this, the
following information is rendered as well, whereas all other tags in the xml files are at
present ignored upon loading a text into the database:

- <p>: new paragraph.
- Text highlighted as italic <i> or bold <b> (to save space we use the html
  encoding, not the lengthy TEI standard).
- <pb n="X"/> : page breaks in editions or manuscripts are rendered by displaying
  the page number X in brackets.
- <l>: line breaks for metric or rhythmic verse. We offer an option to search only
  within these. Indentation, e.g. in the case of pentameters, may be indicated as an
  attribute.
- <app>: critical apparatus entries, optionally with a number attribute, cf. below.

5 Encoded as <author>Lambertus de Monte</author><date>-1498</date></author>.
6 To delimit sentences, <s> tags were automatically introduced into most xml files.
Copyrighted texts

The current copyright situation, granting authors and editors and even their heirs or publishers exclusive rights for a very long time, seems to be to the detriment of the advancement of digital text studies. Especially for editions of texts that stem from long past epochs its applicability itself is controversial. What further complicates things is the fact that the legal situation differs significantly even among European countries and it is hard to be certain about it for a layman. In case we have accidentally posted copyrighted texts we invite the copyright holder to inform us and we will suppress open access for the text in question. The collections do also host a few texts that seem to be subject to copyright and which are thus not displayed publicly. They can only be fully viewed within our university’s IP range, and in the future also by registered external collaborators. But they can be searched by anyone and the results will be shown as snippets. For context the non-local (or not registered) user will then have to consult the book or the website hosting these texts. As a test case the important text *De nuptiis Philologiae et Mercurii* by Martianus Capella edited 1983 by James Willis in the Teubner series is now included. Entire corpora may also be incorporated like this if their legal owners are opposed to open access.

Remarks on xml coding

While checking the tagging of available texts, I noticed that the encoding is really the bottleneck, not only concerning labour (and thus cost !) but also for the quality of the texts and their display. Two examples of text I encountered show well how not to code. When transcribers of Latin texts know no Greek, they may type a Greek quotation as: `<foreign>Epei ...</foreign>`: the first word in not one-to-one transcription, the rest left off. Thus the text is basically rendered unreadable and I have usually decided not to include such texts, and sometimes added the missing text myself. Another example are apparatus entries, for instance:

```
<note>68 69 70 71 Illis : illius Cedant : cedam Fortasse regna et Ad unius : adutus</note>
```

This transcriber was too lazy to transcribe every footnote as one entity in xml (as one is supposed to). Although such an entry is human readable, it will not be displayed in a database in the appropriate place. In some cases such problems could be resolved automatically (using *sed* and similar Unix commands), but in others things had to be left the way they are.

For our project I changed such `<note>` tags to `<app>` tags, as we use note tags differently.

7
IV. Current Features

READING: LEMMATISATION, INCORPORATED DICTIONARIES

A first aim is to present texts in a form to peruse them online. A great help for the reader is the possibility to click words and get their morphological analysis, lemma(ta) and dictionary entries. For the morphological analysis we use Helmut Schmid’s TreeTagger with Latin data prepared by Gabriele Brandolini. This software attempts to disambiguate in case of more than one morphological possibility, e.g. it will try to tell whether *Deo* is in the current context a dative or an ablative. Obviously there are cases for which this produces mistakes, and the user should retain caution. Besides we present the non-disambiguated morphological analysis from the Perseus Latin word-list and, in case these two tools cannot resolve a form, also from the very extensive word-list currently elaborated by the research-group led by Bernhard Jussen at the Comphistsem in Frankfurt. An annotated TEI xml file containing the morphological information gained by TreeTagger and improved by the two mentioned word-lists can be downloaded for each text. TEI-tags of this form are used: `<w type="N:abl" lemma="Deus">deo</w>`.

On our website, lemmata found in these lists are used as input for the following dictionaries whose entries are displayed in the bottom right frame:

- Karl Ernst Georges, *Ausführliches lateinisch-deutsches Handwörterbuch*, Hannover, 1913, for Latin-German translations. It is still the German standard dictionary.

---

8 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
9 The same software was used by Gregory Crane and David Bamman in 2008 for the Perseus Library and its success was assessed in http://dl.tufts.edu/catalog/tufts:PB.001.002.00003. The accuracy that can be reached with the rather small hand-tagged data available is relatively good as can be seen there.
10 This list was generated by expanding dictionary entries and contains many rare and possibly unattested forms: in order not to fill the rather small text-frame too much the stated restriction was chosen.
11 If the TreeTagger list does not know a word, but it gets resolved unambiguously by one of the other two (or both) tools, it is included in the annotated file.
12 Brandolini’s data lack some grammatical information, in this case the fact that *Deo* is singular. We hope to improve this in the future. The full POS tag-set is on display in the «about» section.
We hope to be able to incorporate the *Mittellateinisches Wörterbuch* from the Bayerische Akademie der Wissenschaften (Munich) in the future. The following dictionary is only on display within our local IP-range:

- Jan Frederik Niermeyer & Co van de Kieft, *Mediae Latinitatis lexicon minus* = *Lexique latin médiéval* = *Medieval Latin dictionary* = *Mittellateinisches Wörterbuch*, Darmstadt, 2002, for mediaeval Latin to English, German and French.

The morphological and dictionary tool can also be used directly by typing input into the lower box in the bottom left frame. This also works for Greek words, for which we display the entries from:

- Wilhelm Pape, *Handwörterbuch der griechischen Sprache in vier Bänden*, Braunschweig, 1863-1880\(^{14}\).

**Word-lists and Concordances**

On all levels from a single work to the entire meta-corpus the user can generate lists of words beginning with a certain combination of letters and their number of occurrences by using the upper box in the bottom left frame. For instance *scienti* globally yields\(^{15}\):

| 35247 hits | 34 scientiamque | 2 scientificis |
| 315 scienti | 673 scientiarum | 4 scientifico |
| 12799 scientia | 12 scientiarumque | 14 scientificum |
| 2 scientia » | 442 scientias | 3 scientificus |
| 9662 scientiae | | |
| 3 scientiae » | 527 Scientibus | 4 scientissique |
| 38 scientiaeque | 135 scientie | 7 scientiola |
| 3 scientiales | 2 scientieque | 9 scientiolar |
| 3 scientialia | 13 scientifica | 8 scientiolam |
| 2 scientialibus | 2 scientificae | 2 scientiolas |
| 2 scientialiter | 4 scientificam | 14 scientior |
| 2 scientialium | 9 scientifice | 10 scientiores |
| 9716 scientiam | 5 scientifici | 2 scientiori |

\(^{13}\) Within our IP-range the 1996 version is also displayed. All dictionaries will still contain formatting errors here and there.

\(^{14}\) Many of these dictionaries can also be downloaded from our institute's homepage: [http://www.sglp.uzh.ch/it/mls.html](http://www.sglp.uzh.ch/it/mls.html) for offline use on computers and android devices through the free software *dictan* by Dmitry Viktorov ([http://www.softex.info/](http://www.softex.info/)).

\(^{15}\) Results with only one hit are here not shown: in fact they are mostly typographical or formatting errors. The following list still contains some errors like *scientia >* which should, of course, be included in *scientia*. 
Word-forms in such lists may be clicked in order to get a concordance list for the word-
form in question. Such a list can also be visualised directly by choosing «concordance»
in the initial drop-down menu. E.g. a global concordance for the rare word scientificis
(dat./abl. pl. of scientificus) yields these two results (clicking these snippets jumps to the
textual locus):

1 Coluccio Salutati, De laboribus Herculis, 2, 11 ; 20
eternitate fame a viris scientificis et studiosis quedam immortalitas

2 Vicich Caietanus, Iesseidos libri XII, 1, 14 ; 3
Iuncta scientificis arbor fuit altera ramis

**SEARCH OPTIONS**

More complex searches are handled by the open-source software Sphinx. It offers
complex search functions performed within a maximum of a few seconds for the current
130 million words. The main options are the following ones. Several words in the search
window will find sentences containing all of these words, whereas putting double quotes
donate only the exact phrase. Strictly ordered searches («A before B») can
be performed using >> between search terms. Searches of co-occurrences within N
words are searched by “word1 word2...”~N. The order of the words does not matter for
this operand, so “word1 word2~1 can be used to search direct co-occurrences of two
(or more) words in any order. The characters ^ and $ may be used to indicate begin-
ning and end of a sentence, respectively. Logical operators, also within brackets, can
be used: the minus sign (-) denotes the NOT-operator, and the vertical bar (|) denotes the
OR-operator (cf. examples in chapter VI). Quorum searches display hits with at least N
out of a number of words within a sentence. The syntax is “word1 word2...”/N.
Phrase, proximity, and quorum searches do not allow the use of operators and brackets. But
proximity searches (for a number N) including operators and brackets can be performed
using the string NEAR/N between two operands. A right-hand wildcard * may be

---

16 The final number, after the semicolon, is the internal sentence number we use in the database.
17 http://sphinxsearch.com/. For more details about the search functionality, cf.: http://sphinx-
search.com/docs/archives/1.10/extended-syntax.html.
18 E.g. “deus diabolus”~3’ finds 152 instances mentioning the two words with maximally 2 other
words in between.
19 Try e.g. “dicit dominus ei”~1’ which currently yields 28 hits in all combinations.
20 E.g. ‘ergo$’ yields 1425 hits of sentences ending in ergo.
21 E.g. ‘“gula luxuria avaritia ira acedia invidia superbia vanagloria gastrimargia tris-
titia cenodoxia”/7’ yields 48 hits of lists of names for «seven deadly sins».
22 E.g.: ‘(anima NEAR/3 (beata/felix/perfecta)) immortal* Dei’ with 3 hits.
used in search strings. Searches may be restricted to verse only, e.g. for the vowel quantities of rare words. There are currently slightly over one million verses to search from.

Ill. 3: The lower left frame presenting search tools.

Searches will display a maximum of 500 results at a time (for the sake of general server performance), but global searches can be delimited by time. Thus one may search a frequent combination e.g. only between 1150 and 1200 AD. A work’s precise date of composition is often not known – and in many cases, I just have not yet introduced it into the xml-files. Then the death year (or the end of the floruit period) of its author are taken as the year the text belongs to. Currently we have time information for 83% of all text (but only 60% of all works: many not dated works are short or fragmentary). If numbers for a chronological search are typed into the boxes in the left bottom frame, the result will by default be ordered chronologically but may be re-ordered alphabetically. Texts without a time-tag are not considered in time-dependent searches.

Lemmatised searches find the entire lemma of entered search terms, so « ego facio ago » finds sentences containing forms of these three lemmata, like ego facio actum or factum agitur mihi. Obviously the quality of these results depends on the quality of the automated lemmatisation described above.

**Rendering of critical apparatus entries**

In the relatively few cases of texts which have critical apparatuses tagged, they are indicated by asterisks after the relevant word (or a superscript number if the original publication uses numbers). Clicking the asterisk / number shows the content. Unfortunately there is a tendency among editors not to release the critical apparatus digitally, and their OCRing increases the work and thus cost significantly. Contents of critical apparatuses do not currently appear in search results.

**Greek text**

Quite a few of our texts contain single Greek words or quotations. They are encoded in UTF-8 (obsolete systems like beta-code are converted before I load a text). They can also be morphologically looked up and dictionary entries for them visualised (in the two mentioned dictionaries). For this a version of the morphological list by Perseus

---

23 Besides metrical verse this also includes mediaeval rhythmic verse.

24 Pseudonymous and anonymous texts in the Patrologia Latina remain as the main problem.
(converted to UTF-8) is used. But the database can also display entire Greek texts. We test this with the Greek bible, a Greek florilegium I edited a few years ago, and a few important classical texts from the Perseus Greek collection. Unfortunately vowels with acute accent are encoded twice in UTF-8. We defined the two forms as equivalent in order to improve search results. Greek words can be searched exactly the same way as Latin ones (they have to be typed in UTF-8 unicode). So a global word-list for αἰσχρ yields:

<table>
<thead>
<tr>
<th>Word</th>
<th>Form 1</th>
<th>Form 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>39 word-forms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>168 hits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 αἰσχροῦ</td>
<td>2 αἰσχροῦ</td>
<td>1 αίσχροκερδείας</td>
</tr>
<tr>
<td>8 αἰσχρῶς</td>
<td>1 αίσχροτέρας</td>
<td>1 αίσχροκερδείαν</td>
</tr>
<tr>
<td>11 αἰσχρῶν</td>
<td>4 αίσχρολόγος</td>
<td>1 αίσχρα</td>
</tr>
<tr>
<td>1 αἰσχρω</td>
<td>1 αίσχρολογία</td>
<td>5 αἰσχραι</td>
</tr>
<tr>
<td>8 αἰσχρὸν</td>
<td>1 αίσχρολογίαν</td>
<td>1 αίσχρας</td>
</tr>
<tr>
<td>11 αἰσχρὰ</td>
<td>1 αίσχροπραγεῖν</td>
<td>3 αἰσχήνη</td>
</tr>
<tr>
<td>1 αἰσχρᾶς</td>
<td>1 αίσχροκέρδεια</td>
<td>13 αἰσχύνη</td>
</tr>
<tr>
<td>5 αἰσχρᾶν</td>
<td>2 αίσχροκέρδειαν</td>
<td>2 αἰσχύνης</td>
</tr>
<tr>
<td>1 αἰσχρότης</td>
<td>1 αίσχροκερδός</td>
<td>8 αἰσχύνην</td>
</tr>
<tr>
<td>39 αἰσχρὸν</td>
<td>2 αίσχροκερδή</td>
<td>1 αίσχνας</td>
</tr>
<tr>
<td>1 αἰσχρᾶς</td>
<td>2 αίσχροκερδία</td>
<td>1 αίσχρόν</td>
</tr>
</tbody>
</table>

Some details still need to be improved (the alphabetical ordering or the display of words with gravis which should appear together with those with acute accent). In the future more texts may be included, but our priority for the moment is Latin.

**Downloads**

We intend to share all our work for non-commercial use with a share-alike license. So at present users can download freely available texts in their root TEI xml form, as pure text (.txt without the xml mark-up), as morphologically annotated xml (by TreeTagger as described above), or as pdf files that can, for instance, be read on e-reader devices. These pdfs are currently only available as parts of works containing a dozen or so pages (they are generated on the fly and this increases the server load a lot). The option to download entire corpora or the full collection will follow.
V. Currently loaded texts

Currently we display twelve corpora of very different size, scope and origin:

- **Corpus 0**: Bible, 1.5 M words. This corpus contains Vulgate\(^{25}\), Greek text of Old and New Testament\(^{26}\) and the Hebrew text of the Leningrad Codex\(^{27}\). The user may jump directly to a locus and visualise synoptic views of the biblical text in Hebrew, Greek and Latin.

- **Corpus 1**: (Parts of) ten Latin translations of Aristotle’s *Physica*, 0.3 M words, which served me for linguistic studies, mostly transcribed by myself.

- **Corpus 2**: *Patrologia Latina*, 95 M words. A digitised version of the *Patrologiae cursus completus, series latina*, printed by J.-P. Migne in Paris between 1844 and 1864 and reprinted in the 1880s. Its hierarchical divN tag-set contains oddities as I introduced it automatically. The user may jump directly to a volume and column from this corpus’s start-page.

- **Corpus 3**: Works of St Thomas Aquinas, 8.1 M words, access restricted. Users outside University of Zurich should consult the full data on the freely accessibly *Corpus Thomisticum*: http://www.corpusthomisticum.org/.

- **Corpus 4**: Miscellaneous scientific texts (in a broad understanding), 4.6 M words. The largest of them is Francisco Suárez’s huge *Disputationes metaphysicae* transcribed by Salvador Castellote (Valencia) and Michael Renemann (Berlin). This corpus also contains two texts I have recently edited and published in book form (by Lambertus de Monte and Eberhard Schleusinger\(^{28}\)).

- **Corpus 5**: Latin texts from antiquity, 4.7 M words. The Latin collection from Gregory Crane’s Perseus Digital Library: http://www.perseus.tufts.edu/hopper/.

- **Corpus 6**: Italian Renaissance, 1.7 M words. Texts from the *Biblioteca Italiana*, a project led by Emilio Russo at La Sapienza University, Rome: http://www.bibliotecaitaliana.it/. We display their Latin texts only (the majority of their texts is in Italian).

- **Corpus 7**: Richard Rufus Project, 0.2 M Words. Texts currently edited by Rega Wood and her team at Stanford University: http://rrp.stanford.edu/index.shtml.

- **Corpus 8**: Croatian Latin texts, 3.4 M words. Texts from the CroaLa-Project, University of Zagreb: http://www.ffzg.unizg.hr/klafil/croala/. Due to our general standard and also for practical reasons (diacritics!) the authors’ names were changed from their Croatian to their Latin form.

- **Corpus 9**: Neolatin texts, 6.6 M words. It contains some of the early modern texts transcribed by the CAMENA project at University Mannheim. These texts are not proof-read and in case one wishes to quote from them, they should be double-checked on the image-files on their site\(^{29}\).

---


\(^{26}\) Rahlfs’ Septuagint and Nestle’s New Testament.

\(^{27}\) Besides these, I experimentally included the Latin translation of another holy text, the *Qur‘ān*, from the Bibliander print.

\(^{28}\) The latter together with Peter Stotz.

\(^{29}\) http://www.uni-mannheim.de/mateo/camenahtdocs/camena_e.html.
– **Corpus 10**: Philosophica, 1.5 M words. Some of the philosophical texts OCR’d by Peter King, University of Toronto: http://individual.utoronto.ca/pking/resources.html, more will follow.

– **Corpus G**: miscellaneous Greek texts, 0.9 M words. Currently only one text I edited myself and some experimental texts from the Perseus project.

The following two graphics (ill. 4 and 5) show the amount of text present in the Corpus Corporum at the end of 2014. A few sketchy remarks about them must here suffice. Considering that the earlier the more works will have been lost, the amount of text may still be approximately proportional to the amount produced – and thus the level of education in the Latin world – up to about 1200. Afterwards we have no representative text sample. Indeed the amount of still extant Latin texts is to be expected to rise constantly from the 12th to at least the 17th century. The peaks illustrate well: classical Latinity (from which many works are now lost), late antiquity, the Carolingian renewal, the long 12th century. The differences between counting works and words are rather minor, except that it looks as if from classical times, the 13th, and the 16th century there are relatively long works in the collection, thus making the peaks higher in the text volume count than in the work count.

![Graph showing text volume in MB per half-century. 1 MB corresponds to some 133,000 words. Ca. 150 MB lack a date tag and are not shown.](image)

30 The *Patrologia Latina* collection ends with the death of Innocent III (1216). In fact the majority of texts from the 13th century onward is still not available in modern editions.


32 The marked peak starting around 350 is likely to be caused by the fact that texts were then copied from scrolls to codices (and those considered obsolete were not copied), and not by a higher level of education than before.
VI. Examples of use

Three examples will suffice, the user is invited to experiment with the database by himself.

1. The rare word *intellectualitas*, possibly coined by Tertullian. A search ‘intellectualitas’ lemmatised, from -200 to 1000 AD, produces 3 hits. Later on in university scholasticism it becomes much more common (globally 148 hits):

   1 Tertullianus, *De anima*, PL 2, 0717B (auctor 150-230)
   Alioquin licebit animae, dilapsa domo, ex destitutione propriorum subsidiorum incolumi abire, habenti sua firmamenta, et propriae conditionis alimenta, immortali-
   tatem, rationalitatem, sensualitatem, intellectualitatem, arbitrii libertatem.

   2 Ioannes Scotus Erigena, *Versio operum S. Dionysii*, PL 122, 1133D (auctor 810-877)
   Per vero directum, cum non in semetipsam intrans, et singulari intellectualitate mota; hoc enim, ut dixi, est, quod secundum cyclum: sed ad ea circa se proveniens, et ab his extrinsecus, tanquam a quibusdam symbolis variatis et multiplicatis, in simplas et unitas reductur contemplationes.

   3 Remigius Antissiodorensis, *Enarrationes in Psalmos*, PL 131, 0816C (auctor c.841–908)

33 Notes 18 to 22 above propose some more exemplary searches.
Psalmus iste attribuitur Domino Iesu, qui ideo dicitur intellectus, quia hic designatur bona devotio et operatio ipsius, quae pertinent ad intellectualitatem, id est ad discretionem.

(2)

‘“materia elementorum”~5 (sensibilisintellectualismaterialis)’ searches sentences in which materia and elementorum appear with a maximum of 5 words distance AND where either sensibilis or intellectualis or materialis (or several of them) occur. Globally there are 2 hits, the first one of which lacks a time tag and is anonymous:

1 Incertus 040, Soliloquia animae ad Deum, PL 40, 0887

Anima enim quam creasti non de te, sed per Verbum tuum; non ex qualibet elementorum materia, sed ex nihilo; quae quidem rationalis, intellectualis, spiritualis, semper vivens, semper in motu, quam signasti lumine vultus tui, et consecrasti virtute lavacri tui: ita facta est capax maiestatis tuae, quod a te solo, et a nullo alio possit impleri.

2 Richardus Rufus Cornubiensis, Memoriae quaestionum in Metaphysicam Aristotelis, 6, 1 (auctor -c.1260)

Ad illud in contrarium dicendum quod in definitione elementorum cadit transmutabilis materia sed non sensibilis.

(3)

Another complex example: ‘“quatuor personae”~5 -trinit- (paterfilius)’ searches all sentences which contain the two words quatuor and personae within a maximum distance of 5 words and either pater or filius (or both), and NOT a word beginning with trinit-. Globally there are 5 hits:

1 Alcuinus, Adversus haeresin Felicis, PL 101, 0099B (auctor 730-804)

Quomodo non erunt quatuor personae, si [Forte, scilicet] Pater, et Spiritus sanctus, et Filius proprius, et filius adoptivus?

2 Innocentius III, De quadripartita specie nuptiarum, PL 217, 0926B (auctor -1216)


3 Thomas Aquinas, Quaestiones disputatae de potentia, 8, 9, 9; 49 (auctor 1225-1274)

Ergo oportet quod sit alius spiritus quo pater amat spiritum sanctum; et sic quater sunt personae in divinis.

4 Thomas Aquinas, Scriptum super Sententis, 1, 29, 1, pr.; qc. 2 co. (auctor 1225-1274)

Hic quaeruntur quatuor: 1 utrum una persona sit principium respectu alterius: et supposito quod sit; 2 utrum principium dicatur univoce de Deo respectu divinae personae, et respectu creaturae; 3 utrum eadem notione pater et filius sint principium spiritus sancti; 4 si possint dici unum principium ipsius.

5 Thomas Cisterciensis Ioannes Algrinus, Commentaria in Cantica canticorum, PL 206, 0022A

VII. Plans for the future

The future shape, development and financing of this resource are not yet clear. But the database as a platform for the presentation of TEI xml formatted Latin texts can be hosted indefinitely at the University of Zurich and will thus remain online. The data files will be shared and so will the source code. We may even be able to offer an offline version for Linux computers into which users can load their own compatible TEI xml texts or load our SQL dumps in the near future.

I am currently working on making the xml-data cleaner. Caution should be used with them at present: some of the xml-files contain syntax errors. Mostly opening tags that are not closed or some that overlap (which is forbidden in xml). I have begun a half-automated cleaning process, but it will take some time to correct – hopefully – all of these syntax errors. Our database can still load the texts with such errors, but for further use by others it is certainly advisable to have the root files free from syntax errors.

At present we are working on some more features. One of them is direct links from biblical quotations to the bible text. Another idea is to determine the quality of the texts’ OCR and tagging. Besides the mentioned syntax problems, many of them will contain typographical errors here and there. We may indicate for each text how clean the text itself and the xml mark-up are. An option for users to report such errors is a further possibility.

Philipp Roelli
University of Zurich
roelli.mls@gmail.com

Abstract. – The article presents a new online resource: the Corpus Corporum, an open Latin text repository and tool hosted at the University of Zurich. It is organised as a meta-collection of Latin text corpora from various sources which can be used separately or collectively. Currently there are some 130 million words loaded from all epochs during which Latin was used. They can be searched and studied in various ways: for instance as proximity searches, time restricted searches or lemmatised searches. Word frequency lists and concordances can also be generated – all of this on any level between single work and globally. Furthermore the tool is useful for reading texts online as the user can click words to get their grammatical form resolved and dictionary entries about them displayed. The database is fed by root texts in the standardised TEI xml format which are downloadable freely in various formats (if their owners agree). Apart from Latin texts, a few Greek texts and a version of the Hebrew bible are also experimentally loaded into the database. The article explains how this resource works and what texts it currently contains, featuring graphics showing text volume per time. It ends with a few exemplary searches.

Zusammenfassung. – Dieser Artikel präsentiert eine neue Online-Ressource, das Corpus Corporum, eine frei zugängliche lateinische Textsammlung. Es ist als Metacorpus organisiert, dessen Corpora, die aus verschiedenen Quellen stammen, einzeln...