Philipp Roelli (University of Zurich)
Jan Ctibor (University of Prague)
LiLa workshop
Milano, Università Cattolica, June 2019

# *Corpus Corporum*: current state and planned further development

**www.mlat.uzh.ch**

UNIVERSITAS TURICENSIS MDCCC XXXIII

# CORPUS CORPORUM

*repositorium operum Latinorum apud universitatem Turicensem*

(i) Background of *Corpus corporum* (CC)
(ii) Present state of CC with live examples
(iii) Some technical background
(iv) Future outlook

# (i) Idea

A collection of Latin text corpora from various sources, each usable by itself but also searchable as a whole, and thus a full-text repository. With 163 M words the largest one for Latin.

Software and text collections are strictly kept apart: texts can be easily added in standardised TEI xml format (and downloaded in that format).

Free and open platform for Latin texts at University of Zurich. Strictly non-commercial.

Project begun 2012, run by Philipp Roelli (Zurich) and Jan Ctibor (Prague).

# (ii) Present state

Main features

- Texts displayed and readable online
- Dictionary entries for words obtainable by clicking words
- Queries (incl. complex searches) on various levels
- A direct link to any of the texts is available
- A synoptic Bible (Hebrew, Greek, Latin, English).

Besides

- Bibliographic information about the edition is displayed, authors and texts are linked to other databases like VIAF, mirabile (SISMEL Firenze), DNB or Wikidata
- Display of entries of critical apparatuses (where present)
- Display of images (where present)
- Our assessment of the edition used, its OCR and the richness of its tag-set (experimental).

# (ii) Present state

At present 26 *corpora*, among which:

| no | name | authors | texts | from–to | words | source |
|---|---|---|---|---|---|---|
| 0 | *Libri sacri* | 2 | 6 | | 2,833,328 | var. |
| 2 | *Patrologia Latina* | 1,807 | 5,259 | 230–1617 | 95,513,977 | cf. OpenGreekAndLatin |
| 4 | *Auctores scientiarum varii* | 46 | 51 | -149–1952 | 6,474,669 | var. (often our own OCR) |
| 5 | *Latinitas antiqua* | 46 | 212 | -184–735 | 5,384,080 | Perseus (Tufts University) |
| 6 | *Rinascimento* | 10 | 67 | 1202–1600 | 1,786,416 | Biblioteca Italiana (Uni Sap. Roma) |
| 7 | *Richard Rufus Project* | 5 | 10 | 1245–1294 | 250,628 | Richard Rufus (Indiana University) |
| 8 | *Croatiae auctores Latini* | 102 | 216 | 1059–2004 | 3,423,000 | CroALa (Universitas Zagrabiensis) |
| 9 | *Neolatinitas* | 14 | 27 | 1502–1706 | 6,807,185 | CAMENA (Universität Mannheim) |
| 13 | *Grammatici Latini* | 51 | 103 | 79–1450 | 1,032,635 | CGL (Université de la Sorbonne) |
| 15 | *Poetica* | 270 | 1,052 | 950–1956 | 2,925,081 | var. (mostly Perseus) |
| 16 | *Antiquitas posterior* | 101 | 189 | 31–1450 | 2,991,072 | digilibLT (Uni Piemonte Orientale) |
| 19 | *Scriptores Ecclesiastici* | 56 | 271 | 200–1150 | 5,657,589 | cf. OpenGreekAndLatin |
| 21 | *Latinità Italiana del Medioevo* | 189 | 346 | 731–1522 | 4'283'397 | ALIM |
| 22 | *Monumenta* | 135 | 407 | 395–1505 | 6'844'635 | dMGH |
| 23 | *Mathematica* | 14 | 16 | 1150–1855 | 509'070 | mostly Università Pisa |
| 24 | *Mirabile Digital Library* | 2 | 4 | 1220–1298 | 303'972 | SISMEL Firenze |
| 25 | *noscemus* | 0 | 0 | n/a–n/a | 0 | Nova scientia (Universität Innsbruck) |
| 99 | *Graeca miscellanea* | 9 | 78 | -700–1260 | 2,302,718 | mostly Perseus (Tufts University) |
| Total | | 2'834 | 8'167* | | 156'614'723 | |

* among which ca. 215 in two or more different editions.

# (ii) Present state

Complex search options:

- Wild-card search (* only at the end of a string), phrase, proximity, quorum searches
- Search in verse only (currently slightly more than 2.0 M)
- Lemmatised search (TreeTagger based)
- Time-constrained search (“*universitas* between 1150 and 1220 AD”)
- Syntactic search (“*cum* + ABL”, “*pro* + SUP”, still experimental)

# (ii) Dictionaries

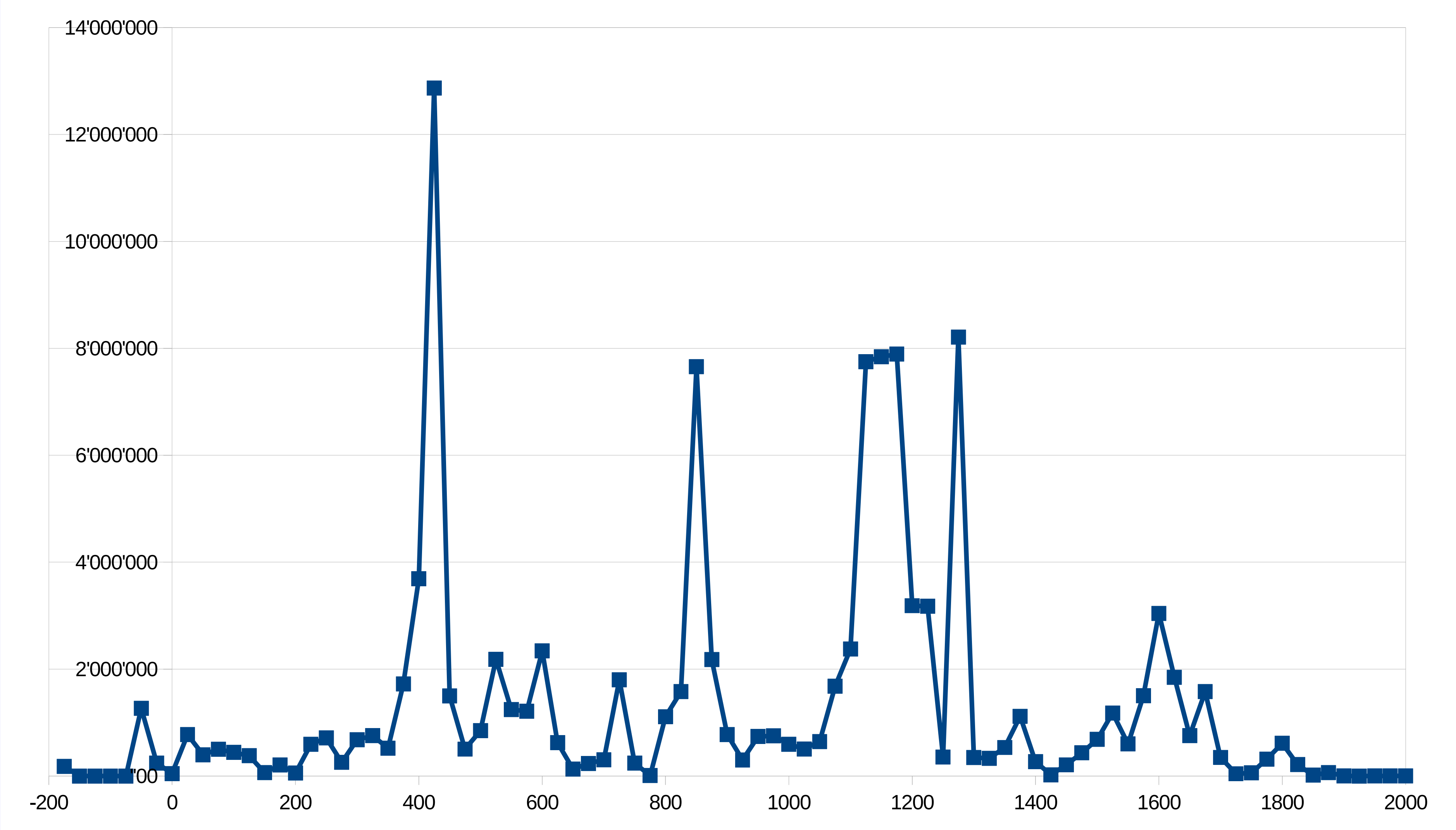
The following dictionaries are integrated:

- Classical Latin: Georges, Lewis & Short, Gaffiot, Dvoreckij
- Mediaeval Latin: DuCange, Schütz, Bohemorum Lexicon
- Neo-Latin: NLW (Ramminger)
- Greek: Pape, LSJ (1940), Authenriet
- Other: Graesse, *Orbis Latinus*, Köbler, *Abkunfts- und Wirkungswörterbuch*

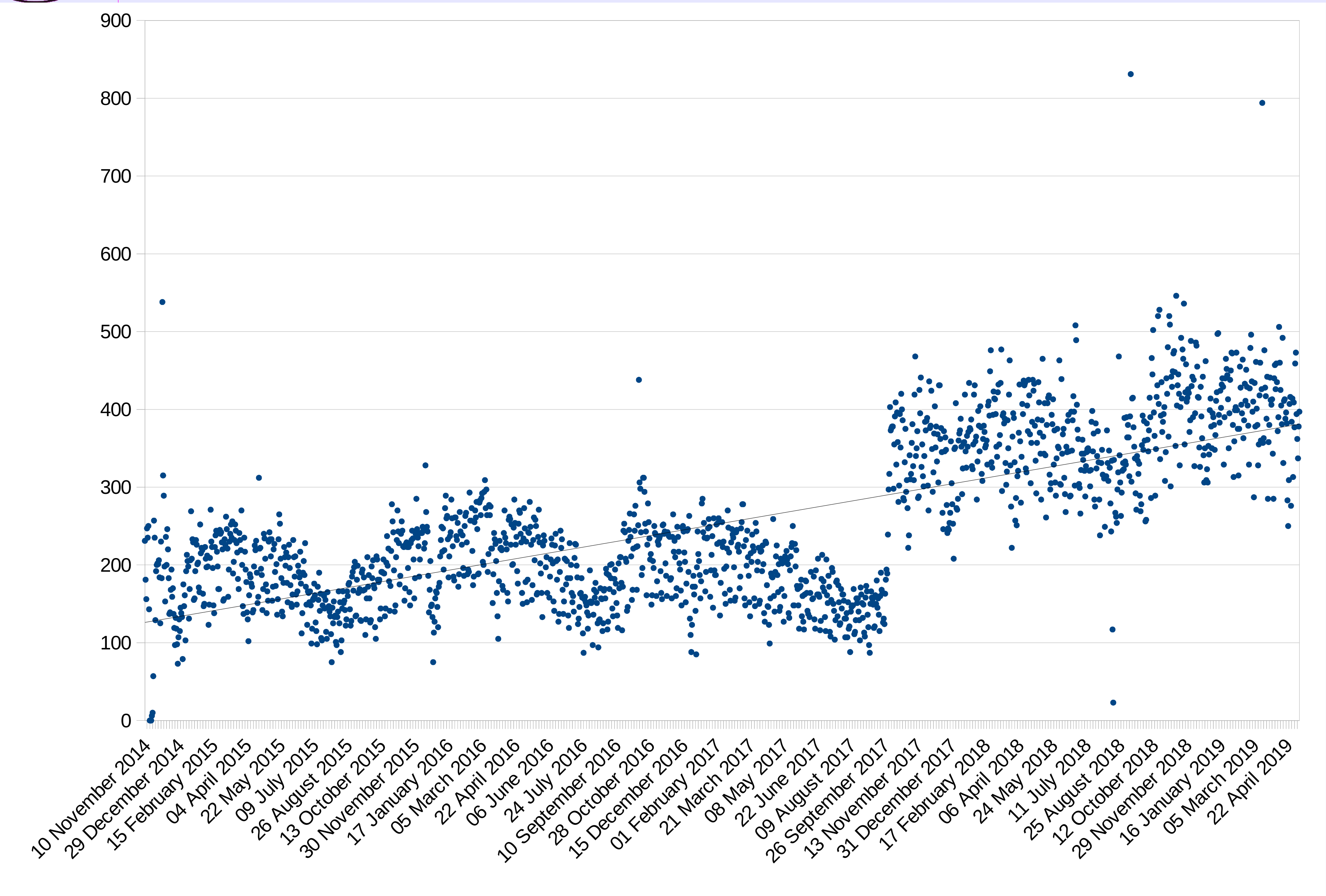
UNIVERSITAS TURICENSIS MDCCC XXXIII

# (ii) Amount of text

(Mio words per quarter-century)



# (ii) Daily users



# (ii) Live Demonstration

http://mlat.uzh.ch/?c=4&w=LamMon.DeSaAri

# (iii) Technical background

Main tools used (all free):

- Ubuntu linux virtual server
- Apache server
- the code is written in PHP (+ HTML, JS)
- xml tools
- the data are stored in MySQL databases.

Some other external tools are used (all free):

- Sphinx engine for fulltext-search
- TreeTagger (Brandolini) for lemmatising.

# (iv) CC 2.0 (2019?)

- New, more professional back-end, using modern xml tools and more efficient SQL tables.

- New interface (more intuitive, much faster, easily extendible).

- API to make our data usable for outside resources (currently: collaboration with *mirabile*, SISMEL Firenze)

- Free log-in for habitual users offering special features, especially definition of user-defined corpora.

# (iv) Planned new features

- Cross-link texts we have in more than one edition (presently 215 works) and offer corpus searches only in the most recent edition
- Mediaeval spelling in searches (finds *hyemps* and *hiems*), also lemmatised
- Downloadable automatically PoS-tagged xml-files for linguistic research
- Automatically highlight quotations from earlier texts.

# (iv) Desiderata

- Better PoS tagging (i.e. better TreeTagger data)
- Cleaner wordlists for lemmatisation
- Better ideas to automatically estimate OCR quality
- More linking to other (free and open) projects
- More texts.

Long-term:

- Publish improved software online (for Linux) enabling interested parties to set up online or offline *Corpora Corporum* of their own.

## Current TEAM

- Philipp Roelli (University of Zurich): project co-ordination, text formatting, long-term planing

- Jan Ctibor (University of Prague): programming, work on the server

- External providers of TEI xml texts and dictionaries

## SUPPORT

- Swiss government through COST action IS1005 (2012–2015)

- Carmen Cardelle de Hartmann and her chair of Mediaeval Latin, University of Zurich (server hosting)