Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
P. Roelli – Corpus Corporum. *An Overview of the Current Development*
DOI: https://doi.org/10.60923/issn.2532-8816/23668

Corpus Corporum

An Overview of the Current Development

Philipp Roelli

Department of Greek and Latin Philology, University of Zurich, Switzerland
roelli.sglp@yandex.com

Abstract

The *Corpus Corporum* project hosted by the University of Zurich is the largest structured digital collection of Latin texts. The texts span from antiquity to the twentieth century, currently totalling approximately 226 million words across 30 corpora. Conceived as an open-access research infrastructure, it provides philologists, linguists, historians, and scholars of Latin with a unified environment for reading, searching, and analysing texts encoded in standardised TEI XML format. Important Latin dictionaries are integrated into the site. The platform, built on open-source technologies including BaseX, Sphinx, and TreeTagger, maintains a distinction between corpus, author, work, and edition levels, and integrates persistent identifiers (VIAF, Wikidata) and external resources such as geschichtsquellen.de. Recent advancements are discussed in the article, especially two major new analytical tools. The 'Text Reuse' module enables configurable intertextual analysis based on *k*-skip-*n*-gram algorithms, while the Metrical Analysis module automatically identifies Latin poetic metres. These innovations allow large-scale, reproducible investigations of textual transmission and poetic structure. An example concerning the sources of Isidore of Seville's *Etymologiae* is briefly discussed. Future developments envision AI-assisted translation, semantic indexing, and synonym-based search, thereby enhancing the platform's potential as a comprehensive, interoperable resource for digital Latin philology and the broader field of computational humanities.

Keywords: Corpus linguistics; Latin linguistics; Text reuse; Metrical analysis; Latin dictionaries.

Il progetto Corpus Corporum*, ospitato dall'Università di Zurigo, costituisce la più ampia raccolta digitale strutturata di testi latini (dall'antichità al XX secolo) e comprende attualmente circa 226 milioni di parole distribuite in 30 corpora. Concepito come infrastruttura di ricerca ad accesso libero, il progetto offre a filologi, linguisti, storici e studiosi di latino un ambiente unificato per la lettura, la ricerca e l'analisi dei testi, che devono essere codificati in formato TEI XML. Importanti dizionari di Latino sono integrati nel sito. La piattaforma, basata su tecnologie open source quali BaseX, Sphinx e TreeTagger, mantiene una chiara distinzione tra i livelli di corpus, autore, opera ed edizione e integra identificatori persistenti (VIAF, Wikidata), nonché risorse esterne come geschichtsquellen.de. Nell'articolo vengono presentati gli sviluppi più recenti del sito, in particolare due nuovi strumenti di analisi: 'Text Reuse' e 'Metrical Analysis'. Il modulo 'Text Reuse' consente un'analisi intertestuale basata su algoritmi k-skip-n-gram, mentre il modulo 'Metrical Analysis' identifica automaticamente i metri dei versi latini. Tali innovazioni rendono possibili nuove indagini sulla trasmissione testuale e sulla struttura poetica*

dei testi. Viene brevemente discusso un caso studio tratto dalle Etymologiae *di Isidoro di Siviglia. Gli sviluppi futuri prevedono la traduzione assistita dall'intelligenza artificiale, l'indicizzazione semantica e la ricerca basata sui sinonimi, accrescendo così il potenziale della piattaforma come risorsa completa e interoperabile per la filologia latina digitale e, più in generale, per il vasto ambito delle scienze umane computazionali.*

Parole chiave: Linguistica dei corpora; Linguistica latina; Riutilizzo di testi; Analisi metrica; Dizionari latini.

1. *Corpus Corporum* project

The *Corpus Corporum* project[1] hosted at the University of Zurich provides scholars and the public with free access to a large collection of Latin texts from antiquity, the middle ages, and the early modern period (some of them even from the twentieth century). We are a small two-men project and see the project's added value primarily in aggregating freely accessible data (digital texts, dictionaries, models, software) and making the results useful for Latin philologists, linguists, historians, and students of Latin. *Corpus Corporum* enables users to search and read these texts as well as access scholarly information about the sources and external links about the authors. As the name suggests, the collection is composed of multiple text corpora from various scholarly sources, each usable individually but also searchable globally. There is a hierarchy of corpus, author, work, and edition.[2] With 226 million words in 30 corpora it is the largest structured Latin full-text collection currently available.

The software and text collections are kept strictly separate: texts can be easily added in standardised TEI XML format (and downloaded in that format). These texts mostly come from other projects, a list of which is provided in the 'Home' tab. The software is developed in open access[3] and the site runs on a virtual Ubuntu Linux server. The code is written mostly in XQuery, Python, and PHP (backend) as well as Javascript (frontend), and integrates external (free and open) tools such as BaseX (XML database engine), Sphinx (search engine), and TreeTagger (PoS tagging and lemmatising).[4] The project was initiated in 2012 and is managed by myself and Jan Ctibor (University of Prague) who joined in 2015. Since 2021 there is a completely remade and improved version [2].

The webpage features several tabs, currently the following ones (each with a brief description of its main functions):

- 'Home' with information about the project, novelties, and related publications.
- 'Browser' for the actual texts with two sub-tabs:

[1] https://mlat.uzh.ch/.

[2] There is some overlap: there are authors with works in several corpora. We do not mind having several editions of one texts. It is possible to restrict searches to the most recent edition of each text.

[3] https://github.com/CtiborJan/Corpus_Corporum.

[4] The two last named are dated and will be changed in the future.

  - 'Bibliography' containing bibliographic data[5] about the current level (corpus, author, text, or edition) including external links and download options (mostly in TEI XML[6]).
  - 'Text Viewer' displaying the actual texts; words are clickable for dictionary entries; extra information from the TEI files is displayed: book and chapter organisation, page-breaks, line breaks for verses, apparatuses, images; some new functions (such as text reuse and metrics) are now available by mouse-over on the texts' left-hand margin.
- 'Dictionary' is dedicated to working with the dictionaries that can also be queried by clicking words in the texts; inflected words can be searched too. Details about the dictionaries are provided below.
- 'Synoptic Bible' offering the Bible texts we have in a synoptic view.
- 'Metrical Analysis' is a new feature analysing user provided verses.
- 'Help' with basic information about the use of the site.

These tabs will be partly redesigned in the future. We intend to expand the capabilities of the synoptic Bible viewer to display any two (or more) editions of the same text[7] and a new tab 'Text Reuse' will be added. This tab will enable users to apply the text reuse feature to their own (short) texts.

[5] Especially VIAF, DNB, Wikidata, mirabile (SISMEL, Firenze), geschichtsquellen.de (BAdW, München).

[6] This includes the original input files, but also files including the images (if there are any) as zip-files, and PoS tagged and lemmatised versions.

[7] The main outstanding problem is the automated alignment.

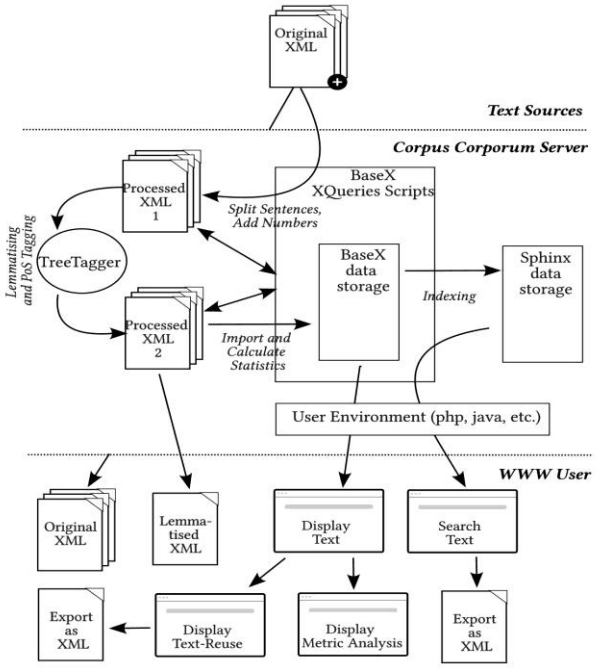

Fig. 1 – Data flow.

Fig. 1 shows the data flow from TEI XML files containing the Latin texts to information and data the user can view or download. The XML downloads are especially useful for users who wish to use the data in ways we have not implemented on the server. Before focusing on the main recent innovations in the remainder of this article, I present an overview of the dictionaries currently displayed and their sizes in characters and lemmata (bibliographic details about them can be found on the site).

Dictionary	From	To	Time period	Current range	M characters[8]	k entries
Georges	Latin	German	BC 200-600 AD	A-Z	25.0	54.7
Lewis and Short	Latin	English	BC 200-400 AD	A-Z	24.8	51.6
Gaffiot	Latin	French	BC 200-600 AD	A-Z	13.4	72.2
Dvoreckij	Latin	Russian	BC 200-400 AD	A-Z	9.6	49.5
DuCange	Latin	Latin	500 AD-1500 AD	A-Z	48.2	90.4
Mittellateinisches Wörterbuch	Latin	German/Latin	500 AD-1200 AD[9]	A-C	18.8	21.7

[8] Not counting XML markup tags.

[9] Plus the works of Albertus Magnus (ca.1200-1280).

Novum Glossarium Mediae Latinitatis	Latin	French	500 AD-1500 AD	L-P	7.4	9.1
Bohemorum Lexicon	Latin	Czech/Latin	500 AD-1500 AD	A-M	24.8	48.4
Schütz, *Thomas-Lexikon*	Latin	German	13th c. AD	A-Z	3.2	1.9
Koebler[10]	Latin	German	BC 200-1500 AD	A-Z	34.2	190.8
Ramminger[11]	Latin	German	ca. 1300-1700 AD	A-Z	8.0	19.6
Graesse, *Orbis Latinus*	Latin	German	ca. BC 100-1800 AD	A-Z	1.0	23.5
LSJ (1940)	Greek	English	ca. BC 700-500 AD	A-Ω	28.8	141.0
Pape	Greek	German	ca. BC 700-100 AD	A-Ω	19.0	97.4
Authenriet, *Homeric Dictionary*	Greek	English	ca. 700-500 BC	A-Ω	1.1	10.5

Table 1 – Dictionaries available on Corpus Corporum, size in characters, and number of lemmata.

These dictionaries comprise many of the most important tools for Latin lexicography that exist. Some of them are still incomplete, with additional letters being added as they become available. Unfortunately, some others are not freely available and cannot be used in our project. Additionally some, like Lampe's *Patristic Greek Lexicon,* have not yet been digitised.

Thanks to funding by University of Zurich we were able to implement new features and improve some existing ones in the past two years.[12] The most important recent improvements of the site are:

- A search API to make our data usable for outside resources (currently used by Mirabile, SISMEL Firenze).
- A reorganisation of the data: now we strictly differentiate between the mentioned four levels: corpus, author, work, edition. For some 250 works we currently have more than one edition.
- Direct links to authors through their VIAF or Wikidata authority numbers (e.g. https://mlat.uzh.ch/browser?path=Q316083 for William of Rubruck using his Wikidata number).
- Permalinks on all levels for easy accessing and sharing (for example https://mlat.uzh.ch/browser/cps_16.Boetiu.DeCoPh2#:5.1;2 for Boethius, *De consolatione Philosophiae*, the edition in corpus 16, book 5, chapter 1, sentence 2).

[10] https://www.koeblergerhard.de/Latein2/LAWVorwort2.html, compiled by Gerhard Köbler from various dictionaries, especially all TLL headwords. Used with the author's permission.

[11] http://nlw.renaessancestudier.org/. Used with the author's permission.

[12] Funding by 'Digitale Lehre und Forschung' for 2023 and 2024.

- Flags to exclude orthographical variants in searches (*uirtus* / *virtus*) and to use medieval spelling in searches (*hyemps* and *hiems*).
- Users can download automatically PoS-tagged XML-files for linguistic research.
- Links for authors and works to the project geschichtsquellen.de (in collaboration with Bayerische Akademie der Wissenschaften, Munich).
- A TEI standard scheme.[13]
- The possibility to limit full-text searches to the most recent edition of each work.
- Biblical references in the *Patrologia Latina* can now be clicked and lead to the synoptic Bible text.[14]

But the two major novelties we have been working on are text reuse and metrical analysis which are now discussed in some detail.

2. Text Reuse Analysis

The idea behind our approach is to utilise the speed of the Sphinx search engine to conduct real-time full-text searches for all passages of entire sections or books of Latin works. This allows us to identify similar phrasing in older texts (ideally quotations) and more recent texts (ideally text reception). The optimal search criteria will vary depending on the specific research question, as the desired level of strictness will differ. If the goal is to identify all quotations, a looser set of criteria with later manual screening may be more suitable. Conversely one may wish to avoid trivial matches such as *ille autem dixit*[15] and find only substantial quotations. To provide users with flexibility we have chosen to offer a range of search criteria that can be adjusted in the cogwheel menu when initiating a text reuse query. It can be made visible by mouse-over on the left-hand side of the text in question. In general we search for *k*-skip-*n*-grams, that is, *n* words must be identical with a maximum of *k* different words between each of the *n* words. The default settings are k=1 and n=3. Additionally, users can choose whether the words must occur in the same order or not. They can also select an offset value *m;* that is, the *m* most common words in *Corpus Corporum* are skipped in the searches (default = 50; range: 0-200). We are currently assessing the addition of additional options:

- Lemmatised searches instead of word searches (in order to also find quotations in another case or verb form).
- Omitting function words, thus only querying verbs, adjective, and nouns.

[13] Details in [3]: 363. It is as yet only implemented in some of the corpora.

[14] This feature is in a beta stadium and does not always function as yet.

[15] This exact phrase occurs 187 times in *Corpus Corporum* 24.10.2025.

- The use of frequency classes in searches. By definition, a text corpus's most common word belongs to frequency class 0, those that are 2^n times less common to the *n*th.[16] In order to avoid trivial matches a setting 'match only words of frequency class *n* or higher' could be added.[17]

In Roelli [3] I conducted some initial testing of the currently available text reuse options. For the current report, I have conducted a larger scale experiment and searched for quotations in Isidore's *Etymologiae* (Lindsay's edition) using 2-skip-4-grams and an offset of 200. These settings are meant to minimise false positives and find only substantial quotations by requiring at least four identical words to appear without any of the 200 most frequent Latin words. I downloaded the resulting XML files for all twenty books and used a script to remove all hits except the oldest one for each passage if it predated Isidore's text.[18] This process yielded a total of 3,191 such passages from 336 works by 131 authors. Table 2 lists authors from whom at least three potential quotations were found.

Servius	446	Cicero	34
Augustinus Hipponensis	397	Donatus	30
Cassiodorus Vivariensis	288	Hegesippus Ps.	28
Hieronymus Stridonensis	265	Lucanus	28
Plinius maior	208	Pompeius	27
Anonymus	154	Diomedes Grammaticus	26
Solinus	149	Eusebius Caesariensis	24
Vergilius Maro	116	Eucherius Lugdunensis	23
Biblia	75	Iunius Philagrius	23
Lactantius	74	Marcus Iunianus Iustinus	23
Tertullianus	74	Columella	21
Boetius	72	Benedictus Nursiae	20
Martianus Capella	66	Quintilianus	14
Orosius	49	Hilarius Pictaviensis	12
Ambrosius Mediolanensis	46	Martialis	12
Gregorius I	35	Vindicianus Afer	12

[16] A more precise definition: https://en.wikipedia.org/wiki/Word_list#Statistics, accessed 24.10.2025.

[17] For the above example we get ille (5) autem (3) dixit (6). For instance excluding all classes down to and including 6 the 235 most common words would be excluded from searches.

[18] Thus I kept hits with <hit n='1' chronologial_relation='--older--'>.

Eusebius Pamphilus	10	Pseudo-Priscianus	5
Gaius	10	Pseudo-Probus	5
Caelius Aurelianus	9	Rufinus Aquileiensis	5
Grillius	9	Aulus Gellius	4
Sallustius	9	Consentius	4
Sacerdos	3	Eusebius Vercellensis	4
Suetonius	3	Florus	4
Flauius Sosipater Charisius	8	Lucretius	4
Ovidius Naso	8	Prudentius	4
Pseudo-Sergius	8	Sulpicius Victor	4
Aphthonius	7	Athanasius Alexandrinus	3
Audax	7	Cledonius	3
Eugyppius Africae	6	Cyprianus Carthaginensis	3
Leo I	6	Eutropius	3
Persius	6	Faustus Rhegiensis	3
Petronius	6	Iustinianus	3
Terentius Afer	6	Iuvenalis	3
Gargilius Martialis	5	Marcus Cetius Faventinus	3
Horatius Flaccus	5	Paulinus Nolanus	3
Lactantius Placidus	5	Priscianus Caesarensis	3
Origenes	5	Pseudo-Cyprianus	3
Probus Marcus Valerius	5	Velius Longus	3
Pseudo-Augustinus	5	Venantius Fortunatus	3

Table 2 – Authors from whom at least three potential quotations were detected in Isidore's *Etymologiae* and their number of hits.

The next table (Table 3) lists the twenty most commonly quoted works (again in the form they are referred to in *Corpus Corporum*).

Servius, In Vergilii Aeneide comentarii	215
Plinius maior, Naturalis historia	208
Cassiodorus Vivariensis, Institutiones	156
Anonymus, Adnotationes super Lucanum supplementum	143
Augustinus Hipponensis, De civitate Dei	111
Solinus, De mirabilibus mundi	103
Cassiodorus Vivariensis, De artibus et disciplinis liberalium litterarum	99
Servius, Commentarius in Vergilii Aeneidos libros	99
Servius, In Vergilii Georgicis comentarii	94
Vergilius Maro, Aeneis	86
Biblia Sacra	75
Martianus Capella, De nuptiis Philologiae et Mercurii	66
Augustinus Hipponensis, Enarrationes in Psalmos	66
Hieronymus Stridonensis, Quaestiones Hebraicae in Genesim	51
Solinus, Collectanea Rerum Memorabilium	46
Boetius, De diffinitione	39
Orosius, Historiae	32
Servius, In Vergilii Eclogarum comentarii	31
Augustinus Hipponensis, De Trinitate	30
Donatus, Ars maior	30

Table 3 – List of the twenty most commonly detected works and the number of hits from them.

For a serious philological study, these numbers and passages would need to be checked in detail. Errors are to be expected as our data is not always clean, and some works may be attributed

differently today than in the often dated editions we use.[19] Apart from this, Isidore will often have quoted from sources now lost to us or not present in *Corpus Corporum*.[20] In the latter case, our results may show not the original author but another early author quoting from it. But even without checking all of this, it would seem to be a significant result that Isidore most commonly quotes Servius, Augustine, Cassiodorus, Jerome, and Pliny. This result could be reached in an hour's work instead of a year-long manual study of sources.

A first collection-wide application of this new text reuse tool that we are planning is to search for Bible references in the entire collection and then mention them in the apparatus including links to the Bible text. However, this may take some time to implement.

3. Metrical Analysis

This feature was announced in Roelli [3]. Now it is mostly functioning and can be used in its own tab. In a nutshell it works as follows. We downloaded the extensive (and clean) quantity data of Latin words from Wiktionary in 2024. With this data and a set of algorithmic requirements for Latin verses we convert strings given in the TEI XML files in <l> tags to strings of heavy and light syllables, then match them against known metres.[21] Here is the output of some quite randomly chosen example verses, with known long vowels known to our database indicated by macrons:[22]

Hymnidica aut quidquid cecinit laus mystica Dāvīd

- u u |- - - |- /u u |- /- |- u u |- - — *Hexameter*

ō stelliferī conditor[23] *orbis*

- - |u u - |- u u |- u — *Dimetrum anapesticum*

cum polō Phoebus roseīs quadrīgīs

- u |- - |- /u u |- /u |- - — *Sapphicus minor*

ēheu quae miserōs trāmite dēviōs

- - |- u u - |- u u - |u - — *Asclepiadeus minor*

abdūcit ignōrantia

- - u - - - u u — *Dimetrum iambicum (Archilochium)*

[19] Especially the author 'Anonymus' would have to be scrutinised.

[20] The latter should be relatively rare: the collection contains most major extant literary Latin texts from antiquity.

[21] We have refrained from searching iambic and archaic metres because of their ambiguity.

[22] Words not known to our quantity database are shown in colour on *Corpus Corporum*, here they are highlighted by underlining.

[23] Our database knows both conditor and condītor, but only the former makes a verse in this case (and sense in the context).

saevīs cerva leōnibus

- - - u u - u u — *Glyconeus secundus*

īnfirmā perrumpere lūce

- - |- - |- u u |- u — *Tetrametrum dactylicum plenum*

- - |- - |- u u |- - [24] — *Tetrametrum dactylicum plenum*

- - u - - - u - u — *Versus Alcaicus enneasyllabus*

- - u - - - u - - — *Versus Alcaicus enneasyllabus*

per quod utriqu(e) hominī posset doctrīn(a) adhibērī

- u u |- u u |- /- |- /- |- u u |- - — *Hexameter*

Simon quem vocitant Petrum

u - - u u - u u — *Glyconeus secundus*

nūbibus ātrīs

- u u |- - — *Adoneus*

Hebraici populī et quidquid historia gessit

--- verse schema not recognized ---

A few comments about these results: Sometimes several verses or schemas are possible. In this case we show all of them. Currently the result can be exported by copying and pasting it, in the future we will also offer XML files containing more details. Strophes (such as Sapphic ones) will soon be detected as well. The last example, a verse from Alcuin,[25] contains an unusual quantity. He intended *histŏria* to be scanned as *histōria*. We plan to add a feature to identify also hexameters with one such 'wrong' syllable weight. This will enable us to query such cases, potentially in the entire *Corpus Corporum* and its currently 1.69 million Latin verses.[26]

Jacobsen and Orth [1] published an impressive manually compiled lexicon of such unexpected quantities in the middle ages. Our software may be able to enhance this work in the future by providing, for instance, data about what authors and times tend to scan *histōria* instead of *histŏria*. In this case Jacobsen and Orth already list this Alcuin verse and present some more cases for *histōria* and *histōrice*. In general, Greek loan-words have a tendency to be less fixed in their quantities.

[24] The word *luce* is usually the ablative of *lux* (- u), but it could also be the imperative of *luceo* (- -). Our software can, so far, not differentiate the two possibilities. If both produce a viable verse, both are shown, so the user can decide which is right.

[25] Alcuin. Carm. IV, 8. In codicem jussu Gerfridi episcopi scriptum, edition PL 101.730A.

[26] There are also some 178,000 Greek verses, but the metrics tool does not currently work for them. Text reuse does not work for Greek either.

4. Plans for the Future

Some plans for the future have already been mentioned in passing. In general, we plan to add more metadata about authors, works, editions, and words. We consider developing a free and optional log-in for habitual users, offering special features, especially the saving of data and the definition of user-defined corpora. *Corpus Corporum* is currently fully deterministic; in other words, we have not yet used AI for any of its tasks. We are trying to secure funding to change this by introducing it for some specific tasks, especially to offer automatic translations of texts into English, the possibility of semantic indexing using Latin Wordnet and similar freely accessible data sets. This would, among other things, make it possible to automatically identify non-literal but semantically similar allusions from older texts. The existing PoS determination and lemmatisation can also be improved with AI tools. For this we hope to collaborate with Philippe Verkerk [4].

In general, we would like to encourage users to provide feedback, including bug reports and wishes of further features.

Bibliography

[1] Jacobsen, Peter Christian, and Peter Orth. 2002. *Materialien zu einem Lexikon der irregulären lateinischen Prosodie*. Erlangen. https://kups.ub.uni-koeln.de/62924.

[2] Roelli, Philipp, and Jan Ctibor. 2022. "A New Version of *Corpus Corporum*, the Latin Full-Text Database and Tool". *Archivum Latinitatis Medii Aevi (ALMA): Bulletin Du Cange* 80 (3): 251-266. https://doi.org/10.5167/uzh-265929.

[3] Roelli, Philipp. 2025. "An Introduction and a Status-Report on the Latin Database Corpus Corporum". *Indo-European Linguistics and Classical Philology* 29 (2): 359-374. https://doi.org/10.5167/uzh-279205.

[4] Verkerk, Philippe. 2022. "Elaboration of a Practical Lemmatiser for Latin using Artificial Intelligence". *Archivum Latinitatis Medii Aevi (ALMA): Bulletin Du Cange* 80 (3): 267-294. https://hal.science/hal-04721577v1.