Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
T. Geelhaar – *The* Latin Text Archive. *A Platform for Historical Semantics and Text Mining*
DOI: https://doi.org/10.60923/issn.2532-8816/23548

The *Latin Text Archive.* A Platform for Historical Semantics and Text Mining.

A long-term Project as Part of the Text Archive Series at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)

Tim Geelhaar

Historisches Seminar, Goethe Universität Frankfurt am Main, Germany
tim.geelhaar@gmail.com

Abstract

The *Latin Text Archive* (*LTA*) is an online platform hosted by the Berlin-Brandenburg Academy of Sciences (BBAW) since 2020 (https://LTA.bbaw.de). Its primary objective is to facilitate computer-assisted semantic analysis of Latin texts and corpora spanning various epochs and genres. The *LTA* collaborates with prominent text providers and related projects in this field. Its core activities center on post-philological editorial text preparation, which is essential for implementing text mining techniques in corpus-based historical semantics. The archive lemmatizes and stores Latin texts, augments them with relevant metadata, and organizes them within thematic or genre-specific corpora. These texts can be also read online and downloaded in various formats. Currently in a beta version, the *LTA* offers already 12,960 curated texts authored by 1,280 identified individuals, amounting to 54 million words. Furthermore, the *LTA* supplies access to its morphological lexicon, which supports the lemmatization process. Through the 'Latin Universe', users may also access additional texts not yet fully curated. Both texts and corpora are searchable via third-party tools such as *Voyant-Tools* or through integrated functionalities like the 'Time series query' – which allows for diachronic comparison of keywords and lemmas – and 'Diacollo', which analyses co-occurring lemmas over time.

Keywords: Text Mining; Corpus Building; Historical Semantics; Medieval Latin; Ancient Latin; Lemmatization.

Latin Text Archive *(*LTA*) è una piattaforma online ospitata dalla Berlin-Brandenburg Academy of Sciences (BBAW) dal 2020 (https://LTA.bbaw.de). Il suo obiettivo principale è facilitare l'analisi semantica di testi e* corpora *latini appartenenti a epoche e generi differenti. LTA collabora con importanti provider di testi e con progetti affini nel settore. Le sue attività principali riguardano la preparazione editoriale post-filologica dei testi, fondamentale per l'applicazione di tecniche di text mining nella semantica storica basata su* corpora. *L'archivio lemmatizza e archivia testi latini, li arricchisce con metadati pertinenti e li organizza in* corpora *tematici o legati a specifici generi. I testi possono essere letti online e scaricati in diversi formati. Attualmente in versione beta, il LTA offre già 12.960 testi, prodotti da 1.280 autori identificati, per un totale di 54 milioni di parole. Inoltre, LTA mette a disposizione il proprio lessico morfologico a supporto del processo di lemmatizzazione. Attraverso*

il 'Latin Universe', gli utenti possono anche accedere a testi curati solo parzialmente. Sia i testi sia i corpora *sono interrogabili tramite strumenti di terze parti come* Voyant-Tools *e tramite funzionalità integrate, come la query a serie temporali – che consente il confronto diacronico tra parole chiave e lemmi – e 'Diacollo', che analizza la co-occorrenza dei lemmi nel tempo.*

Parole chiave: Text Mining; costruzione di corpora; semantica storica; latino medievale; latino antico; lemmatizzazione.

1. Introduction

The primary goal of the *Latin Text Archive* (*LTA*)[1] is to support computer-assisted, corpus-linguistic analysis of Latin texts across different historical periods and literary genres. The concept for the *LTA* originated in the late 1990s with the emergence of the first digital text collections. At that point, it became feasible to pursue historical semantics as envisioned by the '*Begriffsgeschichte* approach' established in the 1960s and 1970s, notably through the influential encyclopedia *Geschichtliche Grundbegriffe* [1]. While previous generations of researchers faced challenges in adhering to a key principle of this approach – systematically analysing semantic patterns within diverse discourses using predefined corpora – recent advancements in technology, methodology, operability, and access to textual resources have ultimately enabled the effective implementation of this central rule [6].

The Leibniz Projekt *Political Language in the Middle Ages: Semantic Approaches (2008-2014)* developed the *Historical Semantics Corpus Management* (*HSCM*) database for computer-assisted, corpus-linguistic research, in close collaboration with Goethe University's Text Technology Lab. As part of the *eHumanities Desktop*, the database formed part of a larger digital infrastructure for uploading, pre-processing, annotating and lemmatizing Latin texts. Alongside this experimental expert tool, the Leibniz Project and the Text Technology Lab developed a second, user-friendly web-based platform expertise, *CompHistSem* (*Computational Historical Semantics*) that required little to no technical expertise ([2], [6]).

The database and web platform solved key challenges such as text preparation, corpus building and ensuring that the analysis tools were as user-friendly as possible. However, further tasks arose: How could the long-term availability of the service to the scientific community be guaranteed? How could partners be found to expand the number of texts and thus achieve a higher representativeness of the corpora? How could individual researchers be enabled to prepare their own texts and create corpora? To accomplish these tasks, it was important to gain an institutional partner that could ensure the long-term availability of the database and promote its further development.

In 2019, the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) agreed to include the database in its series of text archives. Since then, the Academy has been hosting the database in a new format and under a new name, namely *Latin Text Archive* (*LTA*), to highlight its proximity to the *German Text Archive* (*DTA*) and the *Patristic Text Archive* (*PTA*). The concept also changed. The *LTA* is now considered as a tool for basic scientific research in the field of digital text analysis. It offers historical source editing enterprises and individual researchers to collaborate on the digital processing and archiving of their source editions. Especially historians who wish to employ a corpus-based historical semantic approach still have to invest a huge

[1] https://lta.bbaw.de/.

amount of time and resources in digitizing, annotating and corpus building ([4], [11]). The *LTA* now aims to relieve text providers and researchers of most of that burden of digital processing while at the same time expanding and improving the text corpora as a central working tool for digital text analysis. This will benefit philologists, linguists and historians from a wide range of disciplines, such as legal history, history of ideas and church history. The *LTA* is deliberately designed to be an open and connected archive. This is because the texts collected through cooperation are in turn to be linked to other digital resources to achieve the greatest possible benefit for research. Currently in beta, the *LTA* already provides the main archiving and analysis functionalities for individual texts and predefined corpora as the core element of the platform.

2. Texts and Corpora

The *Latin Text Archives* provides two databases: the *LTA*, which includes annotated texts, and the 'Latin Universe', which features texts awaiting curation. The *LTA* comprises 12,960 texts by 1,280 authors and contains approximately 54 million word forms. The 'Latin Universe' has an additional 45,102 texts with roughly 52 million word forms. Both databases have similar analytical features, except that the 'Latin Universe' lacks a detailed text list and corpus affiliation. The aim is to annotate all texts in the 'Latin Universe' similarly to those in the *LTA*, enabling integration of the two databases in the future.

The *LTA* retrieves and processes texts from several text providers and partners such as the *Corpus Corporum* database at Zurich university,[2] the *Digital Monumenta Germaniae Historica* (*dMGH*),[3] the Institut de Recherche et d'Histoire des Textes (IRHT),[4] the *Archivio della Latinità Italiana del Medioevo* (*ALIM*),[5] and other research projects like the *Cartae Europae Medii Aevi*[6] or the *Corpus de la Bourgogne du Moyen Âge*.[7] Among the collections that have been integrated into the *LTA* are the *Patrologia Latina*, the *Cluny Charters*, and more than 120 volumes of the *MGH* like the *Scriptores* and *Epistolae series*, as well as the *Concilia* and *Capitularia series*.

Texts received from partner institutions and projects undergo comprehensive pre-processing to ensure compliance with established standards. This process may involve structuring the text in accordance with XML/TEI guidelines, as well as conducting tokenization, sentence segmentation, part-of-speech tagging, and lemmatization. Following upload, further steps include metadata enrichment, text type classification, and assignment to a specific corpus. The *LTA* utilizes authority records such as Wikidata and VIAF identifiers and documents the text type classification[8] independently on its platform. Moving forward, the complete *LTA* base format (in XML/TEI) and all metadata decisions will also be fully documented, aligning with best practices in data management. Generally, the *LTA* is committed to providing its materials

[2] https://mlat.uzh.ch/home.

[3] http://www.dmgh.de/.

[4] https://www.irht.cnrs.fr/.

[5] https://alim.unisi.it/.

[6] https://cema.lamop.fr/#aimsoftheproject.

[7] http://www.cbma-project.eu/.

[8] https://lta.bbaw.de/d/text-type-classification.

under a Creative Commons license, thereby contributing to a broader semantic web of Latin resources, and addressing data silo issues. Consequently, all texts will be additionally lemmatized by the *Linking Latin* project (*LiLa*)[9] to enhance interoperability with other Latin resources available online.

Until then, the *LTA* utilizes lemmatization conducted through the Historical Semantics Corpus Management. A high-performing tagger, developed at the Text Technology Lab, was used to automatically lemmatize all texts, after which results were validated against the morphological lexicon [3]. Following this automated process, a substantial number of texts underwent manual review and correction using a Post-Lemmatization Editor, which also facilitated updates to the lexicon which in turn improved the automated lemmatization. One particular useful feature was to colour-code the lemmatization status of each word in a text so that readers are able to identify ambiguous tagging and not identified words immediately. This feature still exists in the *LTA* to inform about a text's lemmatization status. It also was adapted by others like the *LiLa* project with its 'Text Linker (β)'.[10]

The central concept of a corpus, distinct from that of a repository, originates in computer science and corpus linguistics, and is employed here within the context of historical semantics. Typically, a repository constitutes an unstructured compilation of texts that does not aim to meet specific research requirements. As a result, discoveries made within such repositories are generally not representative and do not facilitate generalizations regarding particular concepts or word usage. Many online digital collections of Latin texts function in this way, with notable exceptions such as the *Cartae Europae Medii Aevi*, which benefits from years of dedicated corpus development and provides the *NoSketchEngine* – a tool for corpus analysis [9].

The *LTA* bases its corpus building on essential criteria that have already been discussed in corpus linguistics. John Sinclair defined a corpus as a «collection of texts in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research» [13]. However, this definition needs to be adjusted for digital historical semantics and historical corpora [5]. First, since 'electronic form' is not specific enough – it could also be PDFs – this aspect needs to be specified as 'machine-readable texts', i.e. in a format that the computer understands, such as TXT or, better, XML. Second, the texts selected must be coherent according to a certain aspect, e.g., genre, author, region, or discourse community which requires also a thorough metadata annotation of each textual item. Third, to achieve representative results, a corpus must be as complete as possible. This requires overcoming earlier disciplinary preconceptions about canonical texts and taking a broader view by including neglected texts too. Fourth, a corpus must be balanced in terms of genres, time periods, and geographical distribution. The last two conditions are particularly challenging for historical corpora, as these are always limited by surviving text transmission. Fifth, a corpus must be finite, reproducible, and accessible, so that the results can be verified. This represents another challenge because the corpus building process can take quite some time before it can be considered as accomplished. For this reason, it is highly advisable to work with versioning of corpora. Sixth, unlike linguistic corpora, historical corpora can be centered on a specific historical event or institution, such as the charters from the Benedictine monastery of Cluny in France, to pursue a historical rather than linguistic research question. Seventh, historical corpora are diachronic by nature, which creates linguistic challenges as orthography is never homogeneous.

[9] https://lila-erc.eu.

[10] http://lila-erc.eu:8080/LiLaTextLinker/.

This requires a specific approach to lemmatizing the corpus, which will be explained later regarding the lexicon section.

The importance of corpora within the *LTA* becomes visible already in its landing page. In the left section on texts, users will find not only the main database itself but also several entry points by author, text type, text type class and by corpus. Next to the 'predefined diachronic base corpora' are the 'predefined exemplary special corpora'. The *Corpus of Frankish Capitularies* is based on the *MGH* edition by Alfred Boretius and Victor Krause. This corpus is the result of the collaboration with the long-term *Capitularia project*[11] that is preparing a new edition of the Frankish Capitularies. The current corpus was presented and discussed at the conference on *Die Sprache des Rechts* at the German Historical Institute Paris in 2017 [7]. The second 'predefined exemplary special corpus' is a revised and augmented version of the edition of Cluny Charters by Auguste Bernard and Alexandre Bruel that also contains the content of those texts that have been only mentioned by the editors. The additional editorial work and the metadata annotation for these more than 5,600 charters took more than a year to complete. Currently, both corpora are available for analytical purposes. Meanwhile, a third corpus is in progress as part of a collaboration between the *LTA* and the *MGH* working group *Constitutiones* at the *BBAW*.[12] This project concentrates on late medieval constitutions from the MGH series of the same title; two volumes containing 1,416 texts are already accessible through the *LTA*.

3. Tool Presentation

The *Latin Text Archive* contains access to both text databases, a morphological lexicon, and various search functions that are outlined in the following chapters. The landing page consists of three main sections [Fig. 1]. The left section allows users to browse texts within the *LTA*, the middle section provides access to the lexicon, and the right section links to the 'Latin Universe'. Two primary search tools are positioned above these sections. The search bar supports multiple methods for querying the *LTA*, with details about the query language available in the instructions next to it. The 'Time Series Query' button leads to a second tool, which will be described later.

[11] *Capitularia*. Edition der fränkischen Herrschererlasse: https://capitularia.uni-koeln.de/en/project/.

[12] *MGH Constitutiones et acta publica*: https://www.bbaw.de/forschung/monumenta-germaniae-historica.

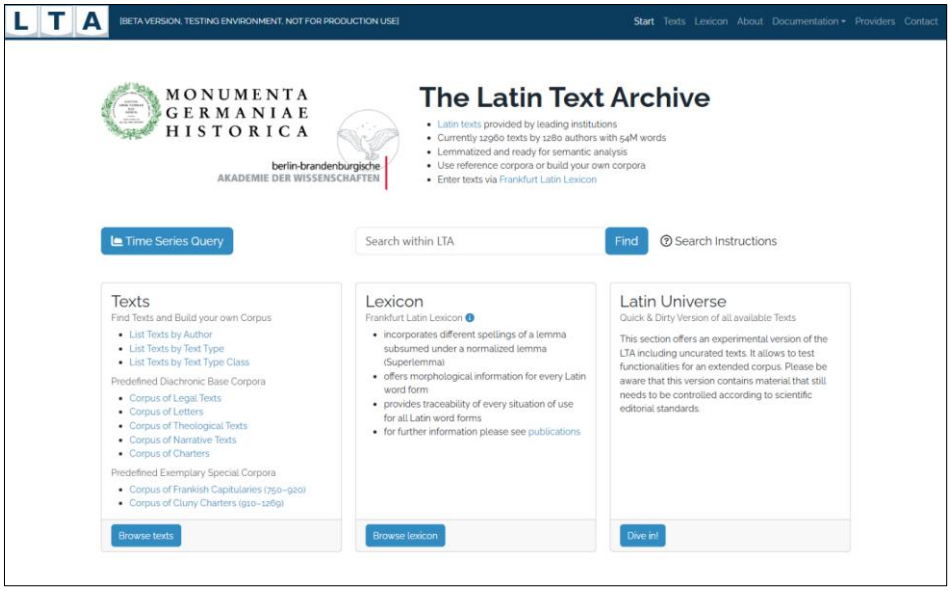

Fig. 1 – The *LTA* landing page with its key features: search within the *LTA*, 'Texts', 'Lexicon', 'Latin Universe'

Texts

In this section, users can browse the complete set of texts or select from various collections or corpora. The following page displays a result list containing data such as the author, title, date, text type, text type class, edition on which the digital document is based, and size. An information icon located behind the list's heading provides further details about the list.

Users may perform a full-text search within the corpus or apply filters using the feature at the top right-hand side of the list. Selecting a title opens the reading view for the chosen document [Fig. 2]. This view presents metadata in the front section, including linked data for the author and the text (when available), as well as bibliographical references to both the source edition and any external provider. Expanding the front section reveals additional information regarding text classification, linked data, and technical annotation, along with availability details. Information icons offer extra details and reference documentation concerning text type classification.

The main panel displays the text in full-text view. The left panel allows switching to a lemmatization info view, where the text reloads with additional lemmatization data for each word. A legend above the text explains the varying lemmatization statuses and their corresponding colours. Clicking on a word opens an extra panel containing detailed morphological data about the selected word.

Fig. 2 – The reading page displays a single text with metadata annotations and various functionalities available on the right panel.

The left panel includes several additional functions. Users can report errors, download the file in different formats, and access the metadata. The statistics section provides information about word count and sentence number in the document. Word clouds and frequency lists can be generated and will replace the main text in the central panel. Additionally, users can send the text – either in its original form or as lemmatized text – to *Voyant tools* for further analysis.

The Lexicon

In this section, users can enter the *LTA* version of the so-called *Frankfurt Latin Lexicon* (*FLL*). This morphological lexicon was initially set up in 2009 to support the automatic lemmatization of Latin texts with the 'Text-technology Lab Latin Tagger'. The lexicon's initial version was based on *lemmata* from various web-based resources like the *AGFL Grammar Work Lab*, the Latin morphological analyser *LemLat*, the *Perseus Digital Library*, the word list *Whitaker's Words Online*, the *Index Thomisticus* and other resources. Then, over the years, the lexicon has grown continuously through the lemmatization of Latin texts for *HSCM* and *CompHistSem* [8].

The *FLL* is organized according to a four-level model comprising 'wordforms', 'syntactic words', 'lemmata', and 'superlemmata'. The LTA version retains wordforms, lemmata, and superlemmata for consistency. 'Superlemmata' offer normalized spellings of lemmata, enabling various spellings at the lemma level so that the significant orthographic variability present in Medieval Latin can be preserved. The lemmata have been morphologically expanded in line with the standard grammar of Classical Latin. This automated expansion process resulted in some overgeneration, which was partly addressed through manual lexicon revision during post-correction of automatic text lemmatization. Additionally, the direct integration between the lexicon and the text database enables users to perform keyword searches within the lexicon and retrieve all corresponding instances from the text database. As of May 2019, the FLL includes 109,000 superlemmata, 133,000 lemmata, and approximately 9.6 million wordforms. The *LTA* version has undergone further refinement, resulting in a decrease in the total number of

wordforms to 8.8 million inflected forms. Since the lexicon and the text database are closely linked, the content of the lexicon will change as further text collections are lemmatised.

The lexicon is displayed to users as a table, with its appearance determined by the selected level of the lexicon. The default display shows the 'superlemma' view. The table consists of four columns: the first includes a 'corpus search' button for finding occurrences of the 'superlemma' in the entire LTA corpus; the second features a 'wordform' button that links directly to the wordform view of the lexicon; followed by columns for the 'superlemma' itself and its part of speech (pos). Search tools for each column are available at the bottom of the table. The lemma view resembles this layout, aside from differences in the representation of the 'superlemma'. To differentiate between the 'superlemma' and 'lemma', the 'superlemma' is marked with an @ followed by its part of speech, such as ecclesia@NN where NN indicates noun, corresponding to the lemmata such as ecclesia, eclesia, aeclesia or even eclesea. The 'wordform' view utilises the same table structure and further includes morphological information within the subsequent columns. As all entries are links it is easy to jump from one level to another.

Integrated Analytical tools

Three key analytical tools will be presented here: the 'general search' function, the 'Time Series Query' and the diachronic collocation analysis, known as 'Diacollo'.[13]

Like other text databases, the *LTA* features a comprehensive search function that enables highly complex queries through its advanced query language. For example, to search for all instances of an adjective directly preceding the word *populus* within the *Corpus of Frankish Capitularies* is done in the search bar on the main page [Fig. 1] by entering '$p=ADJ* populus'. A question mark icon situated behind the input field links to a dedicated web page with a detailed explanation of the *LTA* query language,[14] including practical examples.

In the results view [Fig. 3], the search bar displaying the query appears at the top of the page. The selection field below the input allows users to choose the 'Corpus of Frankish Capitularies' or easily switch to other corpora. Results can be viewed as full text or in KWIC format, and a download button beneath the paging section offers multiple download options. Each result is associated with its source, which users may select to navigate directly to the relevant passage in the text. Additional details for each find are accessible via the three lines icon. On the left side of the interface, a graph displays the temporal distribution of the results within the selected corpus, offering further options to adjust the time range or calculation method for the graph.

[13] https://clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive.html.

[14] https://lta.bbaw.de/d/search.

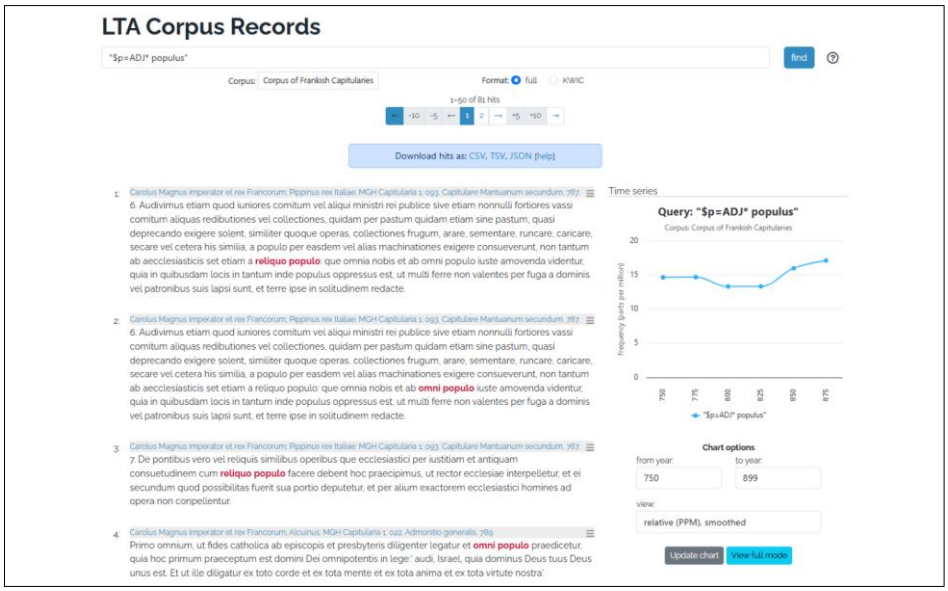

Fig. 3 – The *LTA* Corpus Records view with a cooccurrence list for the search term «$p=ADJ* populus» in the corpus of Frankish Capitularies

The time series graph is a separate tool that expands the analytical capabilities of the *LTA* by enabling comparisons of lemma usage over time within a selected corpus. When users click the 'Time Series Query' button on the main page, they are directed to a new page [Fig. 4]. On this page, up to four lemmata can be entered for comparison. Users can adjust both the time window and the length of each slice in the chosen time frame. The options for view and window influence the display and interpretation of data in the graph. In the example provided, each slice spans 25 years, and the window parameter controls data smoothing by averaging neighbouring time slices to reduce noise and clarify trends. The view option determines whether the data is shown as relative values («relative (PPM)» for parts per million) or as absolute numbers. By hovering over individual measurement points, users can access the specific data for those points.

The chart presents graphs comparing the use of *feudum* and *beneficium* within the *Corpus of Cluny Charters*. The data indicate that the frequency of *beneficium* shows a slight overall decline, followed by a renewed increase during the latter half of the twelfth century. In contrast, occurrences of the lemma *feudum* are minimal prior to the mid-twelfth century, at which point there is a marked rise in usage. Such findings are valuable for historical research, as the graphical data support Susan Reynolds' thesis regarding the history of feudalism: she contends that the feudal era commenced relatively late, in the mid-twelfth century, a claim substantiated by this linguistic evidence from this corpus. [10]

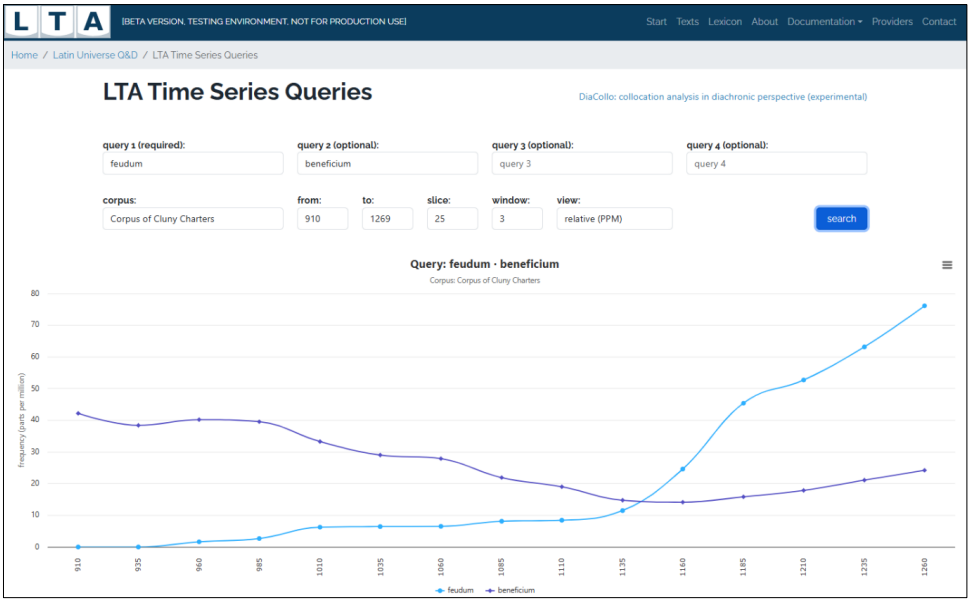

Fig. 4 – Time Series Query displaying the lemma frequency for «feudum» and «beneficium»

The 'Time Series Query' guides users to the third primary analytical tool. Upon selecting 'Diacollo: collocation analysis in diachronic perspective (experimental)', users are taken to the 'Diacollo' interface. This tool was designed independently of any specific language for the *Digital Dictionary of the German Language*[15] and facilitates the visualization of word combinations, known as 'collocations', across time. Accordingly, the results field at the top features a timeline, a play button, and a slider to adjust playback speed in bubble view mode. Users can enter several types of queries in the designated query field, and the on-page help function assists with query formulation. Some parameters will be familiar from the 'Time Series Query', while others are further explained in the help section and tutorial.[16]

The example illustrated in Fig. 5 displays the five most frequent terms (kbest = 5) collocating with the search term *servus* within the overall corpus. The selected date range is 900 to 1300, with intervals set to 25 years. With the format set to bubble, the result frame shows five bubbles whose size and content vary over time. For the year 975, the most relevant terms associated with *servus* are *ancilla*, *uxor*, *nomen*, *manus*, and *infans* according to the Dice logarithm. The inserted picture within Fig. 5 presents the results for 1200, where the collocates are *Honorius*, *Innocentius*, *servus*, *deus*, and *dilectus*. This demonstrates that by 1200, the concept of *servus* was much more strongly linked to the papacy than it had been in the 10th century.

[15] https://dwds.de/.

[16] Tutorial (German): https://kaskade.dwds.de/diacollo-tutorial/#introduction.html.

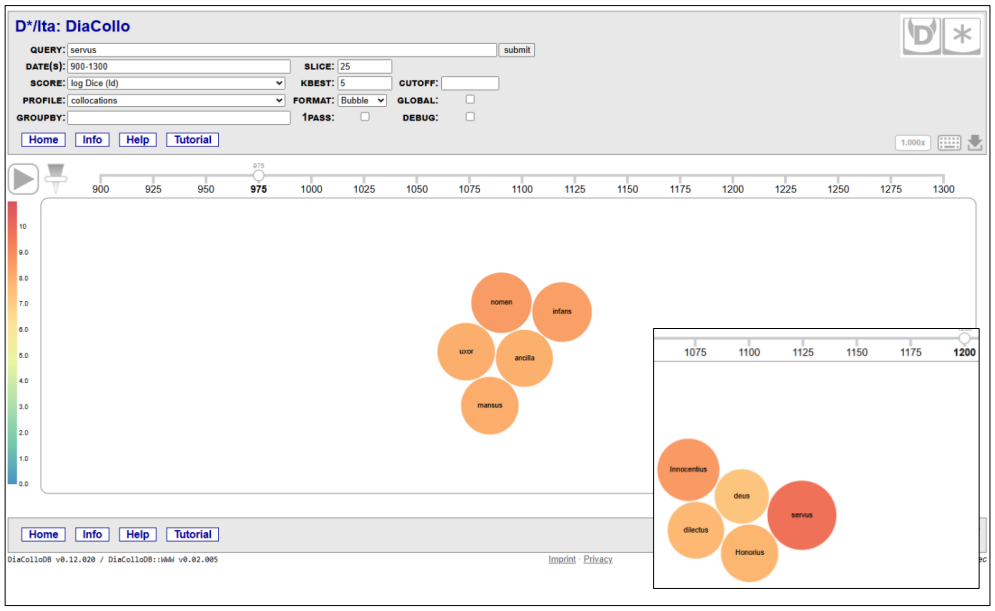

Fig. 5 – The tool 'Diacollo' displays the 5 most frequent terms that collocate with the search term servus

4. Closing remarks

The *Latin Text Archive* can already look back on a long history, yet it continues to face new challenges. The work on text corpora is far from complete. In this field, the *LTA* is dependent on cooperation with text providers. It goes without saying that only copies of the respective editions that are suitable for text mining are processed by the *LTA* and that the work of all text providers is acknowledged in accordance with good scientific practice. Furthermore, the text providers remain in possession of their texts. However, the concept of collaboration encompasses the aspiration to establish connections between the *LTA* and other research initiatives as well as digital resources. To this end, the lexicon and database will be equipped with unified resource identifiers with the help of LiLa, and automatic programming interfaces (APIs) will improve the accessibility of the material.

After all, the idea of linked open data and a web of digital resources is not only a significant advancement from a scientific point of view, but also probably the best way to maintain and enhance the digital infrastructure developed to date. Over the past two decades, considerable progress has been made in this field, with notable contributions from Italian projects featured in this volume. It is essential that the achievements of these initiatives are preserved, even more as securing funding opportunities becomes increasingly challenging.

Looking into the future, one possible approach would be to explore how artificial intelligence could advance research in the humanities. Any reservations here would be misplaced. Artificial intelligence is already being used in our projects in the form of machine learning and deep learning. Automated text classification, automatic lemmatisation and forms of evaluation such as topic modelling are actually AI tools. Therefore, from a technological perspective, it only makes sense to explore AI for advancing natural language processing without sacrificing reinforcement learning techniques. At the humanities level, AI is already being used successfully

for automatic text recognition of manuscripts. Thanks to text recognition models such as the 'Caroline Minuscule Model' on *Transkribus*,[17] it is now possible to automatically transcribe large quantities of handwriting [12]. This allows text databases to include manuscripts alongside edited texts, enabling research into language and discourse at both levels of representation.

It would also be conceivable to have a chatbot that, like the LLM Perplexity, is trained for accuracy and reliability to interact with large amounts of text. However, the most important step would probably be the move towards multilingualism, which is only made possible by language models, to examine and analyse multilingual texts, such as how Latin has been used alongside newer philologies up to the modern era and how functional language change has taken place.

Bibliography

All links have been checked on October 30, 2025.

[1] Brunner, Otto, Werner Conze, and Reinhart Koselleck. 1972-1992. *Geschichtliche Grundbegriffe: Historisches Lexikon zur politisch-sozialen Sprache in Deutschland.* Ernst Klett Verlag.

[2] Cimino, Roberta, Tim Geelhaar, and Silke Schwandt. 2015. "Digital Approaches to Historical Semantics: New Research Directions at Frankfurt University". Storicamente 11 (7): 1-16. http://dx.doi.org/10.12977/stor594

[3] Eger, Steffen, Tim vor der Brück, and Alexander Mehler. 2015. "Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods". In *Proceedings of the 9th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105-113. https://doi.org/10.18653/v1/W15-3716

[4] Geelhaar, Tim. 2025. "Hospitalitas: A Virtue in Danger: Semantic Observations on the Use of hospitalitas in Latin Narrative Sources, 1000-1400". In *Guests, Strangers, Aliens, Enemies: Ambiguities of Hospitality in the Middle Ages, c. 1000-1350*, edited by Wojtek Jezierski, and Lars Kjaer, 39-73. Brepols. https://doi.org/10.1484/M.CURSOR-EB.5.149651

[5] Gippert, Jost. 2015. "Preface". In *Historical corpora. Challenges and perspectives*, edited by Jost Gippert, and Ralf Gehrke, 9-12. Narr Dr. Gunter.

[6] Jussen, Bernhard, and Gregor Rohmann. 2015. "Historical Semantics in Medieval Studies. New Means and Approaches". Contributions to the History of Concepts 10 (2): 1-6. https://doi.org/10.3167/choc.2015.100201.

[7] Jussen, Bernhard, and Karl Ubl. 2022. "Die Sprache der Kapitularien. Einleitung". In *Die Sprache des Rechts. Historische Semantik und karolingische Kapitularien*, edited by

[17] Caroline Minuscule Model: https://www.transkribus.org/en/model/latin-carolingian-minuscule.

Bernhard Jussen, and Karl Ubl, 9-32. Vandenhoeck&Ruprecht. https://doi.org/10.1515/hzhz-2024-1267.

[8] Mehler, Alexander, Bernhard Jussen, and Tim Geelhaar. 2020. "The Frankfurt Latin Lexicon: From morphological expansion and word embeddings to SemioGraphs". *Studi e Saggi Linguistici* 58 (1): 121-155. https://doi.org/10.4454/ssl.v58i1.276.

[9] Perreaux, Nicolas. 2021. "Possibilities, Challenges and Limits of a European Charters Corpus (Cartae Europae Medii Aevi – CEMA)". arXiv:2105.00932.

[10] Reynolds, Susan. 1994. *Fiefs and Vassals. The Medieval Evidence Reinterpreted*. Oxford University Press.

[11] Schiel, Juliane, Ludolf Kuchenbuch, Isabelle Schürch, Nicolas Perreaux, and Tim Geelhaar. 2023. "Historical Semantics: A Vade Mecum". *Österreichische Zeitschrift für Geschichtswissenschaften (OeZG)* 34 (2): 18-47. https://doi.org/10.25365/oezg-2023-34-2-2.

[12] Schonhardt, Michael, Tim Geelhaar, Tobias Hodel, and Jan Odstrčilík. 2025. *Automated Text Recognition: Theory, Platforms, Best Practices*. Bielefeld University Press.

[13] Sinclair, John. 2005. "Corpus and Text – Basic Principles". In *Developing Linguistic Corpora: a Guide to Good Practice,* edited by Martin Wynne, 1-16. Oxbow Books.