Eastern Law in Western Words: Analyzing Roman Legal Terminology in Medieval Charters

Tamás Kovács[1], Angelos Nikolaou[2], Johannes Laroche[3],
Georg Vogeler[4]

[1] Department of Digital Humanities, University of Graz, Austria - tamas.kovacs@uni-graz.at
[2] Department of Digital Humanities , University of Graz, Austria - angelos.nicolaou@uni-graz.at
[3] Department of Digital Humanities , University of Graz, Austria – johannes.laroche@uni-graz.at
[4] Department of Digital Humanities , University of Graz, Austria – georg.vogeler@uni-graz.at

ABSTRACT (ENGLISH)

Monasterium.net provides a dataset of over 600,000 European charters. To use this dataset for diplomatics questions on a European scale instead of focusing on individual, regional or collection-specific studies, automatic methods are necessary. In the ERC project “From digital to distant diplomatics”, we combine traditional approaches of diplomatics with machine learning methods from the fields of natural language processing (NLP). We propose a method for analyzing the reception of Roman law in medieval Central Europe through examining diplomatic sources. Corpus-wide token encoding and subsequence generation via strategic omission of n positions are implemented in this research using Leave-N-Out grams (LNO-grams). The system’s array-based encoding enhances this process by generating integer-encoded patterns, where each row represents a different skip combination. This unified representation allows the Intersection over Union (IoU) metric to consider all possible matching patterns, capturing both local variations and long-range dependencies. We demonstrate our new method with a case study of Roman law’s reception in medieval Central Europe, showcasing the evolution and regional differences in legal language.

Keywords: Roman Law Reception, Digital Diplomatics, Legal Terminology, Text Analysis

ABSTRACT (ITALIANO)

Il diritto orientale in parole occidentali: Analizzare la terminologia giuridica romana nelle carte medievali

Monasterium.net fornisce un dataset di oltre 600.000 carte europee. Per utilizzare questo set di dati per questioni diplomatiche su scala europea, invece di concentrarsi su studi individuali, regionali o specifici per le collezioni, sono necessari metodi automatici. Nel progetto ERC “From digital to distant diplomatics”, combiniamo gli approcci tradizionali della diplomazia con metodi di apprendimento automatico provenienti dal campo dell'elaborazione del linguaggio naturale (NLP). Proponiamo un metodo per analizzare la ricezione del diritto romano nell'Europa centrale medievale attraverso l'esame delle fonti diplomatiche. La codifica dei token a livello di corpus e la generazione di sottosequenze attraverso l'omissione strategica di n posizioni sono implementate in questa ricerca utilizzando i Leave-N-Out grams (LNO-grams). La codifica basata su array del sistema migliora questo processo generando modelli codificati con numeri interi, dove ogni riga rappresenta una diversa combinazione di salti. Questa rappresentazione unificata permette alla metrica Intersection over Union (IoU) di considerare tutti i possibili pattern di corrispondenza, catturando sia le variazioni locali che le dipendenze a lungo raggio. Dimostriamo il nostro nuovo metodo con un caso di studio sulla ricezione del diritto romano nell'Europa centrale medievale, mostrando l'evoluzione e le differenze regionali nel linguaggio giuridico.

Parole chiave: Ricezione del diritto romano, Diplomatica digitale, Terminologia giuridica, Analisi del testo

1. INTRODUCTION

This study presents a new computational approach to analyzing how Roman law was received in medieval Central Europe, using diplomatic sources. Traditional historical scholarship has extensively explored the influence of Justinian’s Corpus Iuris Civilis on medieval legal practice, yet our understanding of its practical implementation across different temporal and geographical contexts remains limited. The proposed research addresses this gap by applying sophisticated computational analysis to medieval charters from Central European archives, focusing on Austrian documents from the period 1200-1400 CE. Central Europe’s key Roman law reception period falls within this research timeframe, providing ample documented evidence for meaningful computational study.

Medieval charters relied heavily on formulaic language, using pre-established templates for legal actions that represented conventionalized pairings of form and function. However, these formulas present a moving target for modern interpretation, exhibiting significant variation across time and regions. This variability stems from several factors, including the evolution of legal practices, regional dialects, and individual scribal preferences. One legal concept could be expressed differently in various historical documents. For example, Papal documents use “appellatione remota”, German Episcopal charters use “appellatione cessante”, and French ecclesiastical courts used “sine appellationis obstaculo.” Such variations complicate the identification and analysis of formulaic language in an extensive corpus of documents.

2. STATE OF ART

This inherent variability poses a challenge for traditional computational analysis methods. Methods based on frequency, counting identical word sequences, cannot easily identify formulas written differently. Similarly, inconsistent spelling, grammar, and word order impede the ability of N-gram approaches to identify formulaic expressions. N-grams, which analyze sequences of *n* consecutive words, are sensitive to even minor variations in the text (Manning et al., 2008) Medieval scribes sometimes add case-specific details to formulas, like “*causa que vertitur inter monasterium sancti Benedicti et homines ville de Capraria*” (the case between St. Benedict’s monastery and Capraria villagers). Standard n-gram methods cannot recognize the resulting patterns (Korkiakangas, 2024; Korkiakangas & Passarotti, 2011). This is important because these insertions were common practice (Koolen & Hoekstra, 2022), reflecting a need for formula adaptation in specific legal scenarios.

Although skip-grams offer more flexible matching (Mikolov et al., 2013), their limited local context restricts their capacity to model the long-range dependencies vital to understanding legal formulas. Analyzing word sequences with gaps, skip-grams handle variability to a degree, but are limited by their context window size. Intricate legal formulas, involving complex conditional statements and their consequences, can surpass standard skip-gram window sizes. Long formulas, such as “*si non omnes his exequendis potueritis interesse, duo vestrum ea nichilominus exequantur,*” may extend beyond typical window sizes. The method may overlook key relationships between elements such as “*si non omnes*” and “*duo vestrum.*”

3. METHOD

The methodological innovation at the core of this research is FLAME (Formulaic Language Analysis in Medieval Expressions, https://github.com/kreeedit/FLAME (in progress)), a computational system specifically designed for analyzing formulaic language in medieval legal documents. FLAME’s novel approach, LNO-grams, identifies legal formulas regardless of variations in expression, spelling, or word order. The value of this method is apparent in medieval texts, given the significant scribal variations in legal terminology. The system’s adaptable design enables it to recognize comparable legal ideas despite variations in wording or structure.

FLAME uses a complex two-stage procedure. First, it implements a corpus-wide token encoding scheme that translates words into integer sequences for efficient pattern matching while retaining contextual sensitivity. This approach makes it possible to recognize similar legal terms regardless of context or spelling variations. The encoding method accounts for both individual words and their contextual relationships, thus enabling more subtle pattern recognition than keyword methods. Secondly, it uses a Leave-N-Out pattern generation system which, for a text sequence of length s, creates subsequences by removing n strategic positions. This method maintains the fundamental framework of legal formulas while allowing for the diverse adaptations made by medieval scribes. Despite different wording and structures, the system can still pinpoint similar legal concepts.

The system uses array-based encoding to improve this process; it creates integer-encoded patterns, with each row showing a unique skip combination. The unified representation lets the IoU metric assess all potential matching patterns immediately, encompassing both local and long-range aspects. Although FLAME presently employs IoU, its architecture is adaptable to more refined similarity evaluations, encompassing the weighting of diverse match categories as complete matches, partial matches, word order discrepancies, and lexical replacements. The adaptable nature of similarity assessment enables researchers to tailor their analyses to specific research questions or document features.

The visualization aspect of FLAME has shown to be especially helpful in historical analysis. The system creates interactive HTML that compares similar texts and visualizes similarity scores with heatmaps across the corpus. Researchers can use these tools to explore matching document sections and discover “bridge words” connecting similar legal terms. Researchers can use the visualization system’s temporal analysis features to study how legal terminology evolves across institutions and over time. Distant reading reveals stable and variable parts of charter formulas, aiding the study of their changes. Researchers can interactively explore data within the visualizations, ranging from large-scale institutional patterns down to fine-grained textual differences.

This research uses computational and traditional historical analysis in creative ways. The research analyzes charters available on monasterium.net from Central European archives. The digital availability of these documents through monasterium.net has made it possible to conduct large-scale analysis of medieval legal documentation, though the computational processing of these materials presents its own challenges. Synchronic patterns in legal terminology adoption across institutions are identified through our use of time series clustering analysis. Viewing specific legal terms as temporal data points, this method reveals previously unseen patterns in how Roman law was received.

Several key steps are involved in the systematic processing of the charter corpus. Documents are initially classified by their source institution, date of creation, and document type. This categorization facilitates the study of adoption trends across various medieval institutions. Second, spelling variations and abbreviations in the medieval Latin texts are standardized through preprocessing. Finally, in step three, FLAME detects and trace patterns of legal terms used in the standardized texts, and the results are analyzed using both computational and traditional historical methods to ensure that identified patterns are historically meaningful.

This analysis might identify different chancery groups based on their particular legal terminology patterns (Härtel, 2015). Early adoption of Roman legal concepts and terminology might be seen in Episcopal chanceries (Willoweit, 2000). This early embrace may be due to their links with Italian legal education and deep understanding of Justinian’s principles (Huschner, 2003). Royal courts could adopt a more measured, methodical approach to implementing new legal terms, combining Roman law with existing practices (Bates, 1995). This measured approach may indicate a calculated strategy to modernize the legal system while preserving established practices. Roman law’s adoption by municipalities could have been delayed and selective, prioritizing practical application. Cities with strong trade links might have adopted Roman legal terms earlier than isolated communities, showing regional differences in adoption patterns.

Analyzing these clusters may reveal if Roman law terminology spread institutionally or geographically in medieval Europe, thus challenging current assumptions about legal innovation dissemination. A temporal analysis of Roman legal term adoption across institutions could pinpoint peaks, thereby indicating whether widespread legal reforms, influential scholars, or other factors drove the process. Discovered patterns may or may not align with documented legal reforms and new schools, possibly revealing the legal transformation coordination. The timeline of these patterns might also reveal links to major historical shifts, including the creation of new universities and the exchange of legal experts between different regions. This analysis will help determine the most significant factors —institutional, geographical, or otherwise—behind the adoption of Roman legal terminology.

FLAME demonstrates significant potential in analyzing medieval legal texts, but limitations exist, suggesting avenues for further development. Currently, the system’s structural analysis ignores legal meaning, which may lead to functionally distinct formulas being grouped together due to structural similarities. This limitation is resolvable through the integration of semantic analysis tools and domain-specific knowledge bases. Its computational cost rises with corpus size, and it demands careful parameter adjustments depending on the document type. Ongoing development of more efficient processing algorithms and more sophisticated similarity metrics is needed to overcome these technical challenges. Future work will concentrate on improving FLAME’s analytical skills by adding semantic analysis tools and specialized legal knowledge, thus enhancing its ability to differentiate between structurally similar but functionally distinct legal formulas.

However, the proposed method offers substantial contributions to the fields of digital humanities method and medieval legal history. First, it showcases how computational methods effectively analyze large medieval legal document collections; this offers a repeatable model for similar research across other times and places. The research’s method is adaptable to analyzing other historical documents, especially those using formulaic language or standardized expressions. Second, it reveals new understandings of how Roman law was adopted in medieval

Central Europe, showing a more intricate, institutionally led process than previously thought. This research challenges traditional views of Roman law dissemination, suggesting fresh approaches to understanding the interplay between institutions and legal transformations. A third benefit is a new computational tool for analyzing formulaic language in historical documents; its uses extend beyond legal texts. The adaptable pattern-matching and similarity assessment methods of the FLAME system could prove useful in analyzing other formulaic historical documents.

ACKNOWLEDGEMENTS

All of the contributing authors are part of the ERC project “From Digital to Distant Diplomatics”, located at the Department of Digital Humanities of the University of Graz. Tamás Kovács has developed the concept and the code of the presented approach. The code development benefitted from support by Angelos Nikolaou. Johannes Laroche contributed to the work with data evaluation and diplomatics research, whereas Georg Vogeler supported the research with comments and his supervision. Our team would like to thank the European Research Council for their generous funding, as well as the International Centre for Archival Research (ICARus) and the involved archives for providing their data on Monasterium.net.

REFERENCES

Bates, D. (1995). Le rôle des évêques dans l’élaboration des actes ducaux et royaux entre 1066 et 1087. *Les Évêques Normands Du XIs. : Colloque de Cerisy-La-Salle, 30 Septembre-3 Octobre 1993*, 103–115.

Härtel, R. (2015). Urkundenlandschaften zwischen Donau, Rhein und Adria. In *Andreas Schwarcz/Katharina Kaska (Hgg.), Urkunden – Schriften – Lebensordnungen. Neue Beiträge zur Mediävistik. Vorträge der Jahrestagung des Instituts für Österreichische Geschichtsforschung aus Anlass des 100. Geburtstags von Heinrich Fichtenau (1912–2000) (Wien, 13.–15. Dezember 2012)* (Vol. 63, pp. 193–211). Beck.

Huschner, W. (2003). *Transalpine Kommunikation im Mittelalter: Diplomatische, kulturelle und politische Wechselwirkungen zwischen Italien und dem nordalpinen Reich (9.-11. Jahrhundert)* (Vol. 52). Hahnsche Buchhandlg.

Koolen, M., & Hoekstra, F. G. (2022). Detecting Formulaic Language Use in Historical Administrative Corpora. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 127–151).

Korkiakangas, T. (2024). *A linguist’s viewpoint: Formulaic language as a challenge for historical linguistics*. https://doi.org/10.5281/ZENODO.10461847

Korkiakangas, T., & Passarotti, M. (2011). Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, *26*(2), 105–116. https://doi.org/10.21248/jlcl.26.2011.150

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Willoweit, D. (2000). Grundherrschaft und Territorienbildung. Landherren und Landesherren in deutschsprachigen Urkunden des 13. Jahrhundert. In *Strukturen und Wandlungen der ländlichen Herrschaftsformen vom 10. Zum 13. Jahrhundert. Deutschland und Italien im Vergleich, hg. V. Gerhard DILCHER and Cinzio VIOLANTE* (Vol. 14, pp. 215–233). Duncker & Humblot.