Formulaic language in diplomatics: Investigating formulas as charter type discriminators

Florian Atzenhofer-Baumgartner, Daniel Luger, Tamás Kovács, Johannes Laroche, Anguelos Nicolaou, Franziska Decker, Nicolas Renet, Sandy Aoun, Niklas Tscherne, Georg Vogeler; University of Graz; Centre for Information Modelling

We investigate the role of formulaic language in diplomatics, with a focus on its function in historical charters and specifically in the context of arbitration. Utilizing data from the Monasterium.net database and employing statistical models, we identify formulaic expressions that are effective in soft-classifying texts related to arbitration. Our work underscores the potential of using formulas and their feature importance for a more nuanced understanding and classification of texts.

Introduction

Large-scale research into formulaic language spans various fields of research. It is particularly evident in digital diplomatics[1], which is concerned with analyzing collections of historical legal documents (e.g., charters). While most research on formulaicity is based on the idea that it involves repetitive language patterns within or across texts, diplomatists put another emphasis on the value of a formula mostly independent of its frequency. They emphasize its role in upholding the legal integrity of documents, where the presence, absence, or incorrect use of formulaic expressions can significantly impact a text's authenticity and validity.

This work offers a glimpse into formulaicity by accomplishing two goals. Firstly, it compares the diplomatic perspective on formulaic language with some views from linguistics in order to highlight key similarities. Secondly, it addresses whether domain-specific formulas (and their variations) can reliably indicate a text's type or category.[2] Given recent trends in diplomatic research[3], we focus on analyzing the concept of arbitration in charters, employing semi-automatic methods to detect and correlate respective formulas across historical languages and categorize their host media. We achieve this by drawing on document metadata, employing statistical models of various complexities, and interpreting the results with expert knowledge. As such, we also tackle the questions: How do formulas differ in their indicative or discriminative power? Does this assessment give us more qualitative insights into charter formulas?

Formulaicity: a working definition

Linguists have accrued an almost overwhelming amount of definitions of formulaicity and its manifestations. What practically all of them have in common is that formulaic language is realized in multiple words, is standardized in usage, and conveys a

---

[1] Antonella Ambrosio, Sébastien Barret, and Georg Vogeler, eds., *Digital diplomatics: the computer as a tool for the diplomatist*. (Böhlau, 2014).
[2] Michael Stubbs and Isabel Barth, “Using Recurrent Phrases as Text-Type Discriminators: A Quantitative Method and Some Findings.” *Functions of Language* 10 (1) (2003): 61–104.
[3] “Arbiter, arbitrator, or compositor amicabilis”, H-Soz-Kult, 06.06.2023, https://www.hsozkult.de/event/id/event-136743.

contextually novel meaning established through convention. We extend this understanding by Brommer's[4] and Bubenhofer's[5] definition who consider the concept of contextual, statistical, and parametric significance. We also take into account Hunston's point in analyzing language patterns with relation to their environments[6], allowing an open interpretation of how an implied vicinity as a correlate manifests. For this work, we choose the 'genre' of legal language and the text type of a charter as points of reference, presupposing considerable differences to other categories.

As indicated, diplomatists share with linguists a predictive model of language that makes frequency and distribution of language pivotal. However, to them, formulaicity primarily stems from its role in legal historical communication. In this regard, it pertains to instances of communication that adhere to templates based on legal contexts and processes. These templates and its derivations often emerge because of repeated usage and as mandated by legal traditions: scribes had to follow rules. Understanding this and the repercussions is formative – it means that chanceries and notaries adopted and exhibited certain norms of how to craft documents. In this context, one can differentiate formulaicity in diplomatics based on what can be called formulaic units, i.e., forms and form parts (e.g., core sections of a legal act), form part abstraction (e.g., a form part grouping), smaller formulaic segments (e.g., phrases), or formulaic templates (e.g., a chancery formula book).[7]

In sum, formulaicity in diplomatic research can be related to either the unique nature of individual utterances and their performance-driven[8] characteristics or the distinctiveness of recurring patterns and their (in)flexibility. In the context of this work, we primarily focus on short performative segments that are likely to follow templates. Analyzing a formula's communicative function, its (changing) form as well as its spread

---

[4] Sarah Brommer, *Sprachliche Muster: Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte* (De Gruyter, 2018), 54.
[5] Noah Bubenhofer, *Sprachgebrauchsmuster* (De Gruyter, 2009), 43.
[6] Susan Hunston, *Corpus Approaches to Evaluation* (Routledge, 2010), 5.
[7] Florian Atzenhofer-Baumgartner, "Quantifying Formulaic Flexibility of Middle High German Legal Texts", Master's Thesis (University of Graz: 2023), 20-23.
[8] Performance in the linguistic sense.

across time and space is fundamental to understanding past legal processes and hierarchies, their influence as well as their development.

The Case of Arbitration

Arbitration, rooted in Roman law, has long been used to resolve disputes. Originally, arbitrators in Rome acted as both mediators and judges, selected by the disputing parties. This less formal, cost-effective, and flexible alternative to traditional court proceedings became popular during the medieval reception of Roman law. Its widespread use is documented in legal manuals, canonical literature, and records of practical cases. Arbitrators form a tribunal and issue arbitration awards, which are legally binding and enforceable in court.

Research into medieval arbitration focuses on understanding its procedures and principles. Historical documents, particularly charters, reveal the roles and structures involved in arbitration which adds to understanding the nuances of the process. Examining the rules and criteria for arbitrators, the types of disputes they addressed, and the content of their decisions is crucial. Additionally, studying regional and legal system variations in arbitration language and terminology offers insights into its diverse cultural and (legal) applications.

A key to this analysis is the examination of formulaic language in arbitration charters, which standardized legal language and ensured clarity in decisions. As a first step, the retrieval and identification of items related to arbitration is necessary. To differentiate these from others, we define arbitration charters as those that cover or refer to activities and processes of arbitration at any level of legal hierarchy.

Data and Methodology

Our data stems from the Monasterium.net database provided with the project From Digital to Distant Diplomatics (DiDip)[9]. We first compile a list of documents from various European archives and collections, for which both the text as well as the abstract are available. An arbitrary date limit up to 1800 is set, with the average year being at ca. 1330. Also, an ensemble of language detection models is utilized to label the data and systematically narrow down its scope.[10] This results in a list of ca. 41,000 items, with ca. 25,000 in Latin and the rest in historical German. Abstracts are mostly in (modern) German (ca. 33,000), English, Latin, and Hungarian. The mean length of the text is 393 words, and the abstract's is 33. Apart from lowercasing, no preprocessing is applied.

Our methodology introduces a novel approach to the quantitative analysis of charters by conceptualizing their type identification as a soft binary classification task. Working without a definitive ground truth, our strategy focuses on categorizing data based on the presence of pre-defined terms of a closed group that indicate relevance to arbitration (arbiter, arbitrator, compromissum, spruchman, hindergang, hintergang, Schiedsgericht, Schiedsrichter, Schiedsspruch, Schiedsmann, [Vermittlung, Beilegung, Abkunft]). This pivot towards using individual words and sequences as first discriminators is both a response to the labor-intensive nature of 'traditional methods' as well as a strategic decision aimed at resource efficiency and reproducibility.[11]

Our study is grounded in the assumption that charters referencing arbitration activities are, by default, classified as arbitration charters. This classification is especially 'soft' due to its reliance on broader generalizations rather than definitive classes, focusing on identifying features that differentiate document types. Results are thus understood rather as indicators of a likelihood to a (high) likelihood of a label instead of the label

9 https://didip.hypotheses.org/.
10 https://huggingface.co/ERCDiDip/langdetect for tenor detection, fastText for abstracts.
11 Compare with Anguelos Nicolaou, Daniel Luger, Franziska Decker, Nicolas Renet, Vincent Christlein, and Georg Vogeler. "Efficient Annotation of Medieval Charters." In Document Analysis and Recognition – ICDAR 2023 Workshops, ed. by Mickael Coustaty and Alicia Fornés (Springer, 2023), 284–95.

directly. This approach entails an assessment of non-indicative formulas, supplemented by a qualitative analysis to differentiate their embeddings. It allows for a more detailed examination of classifier features since individual words (and phrases) pose discriminative values.[12]

Results and Discussion

Quantitative insights

After preparing the dataset, the initial partially overlapping word lists are used to identify occurrence counts in charter texts and abstracts. The first set counts 732 items about arbitration, the second 617. After distilling top-15 skip-gram-based features into formulaic tri-grams, confirming, and then merging them based on co-occurrence in the importance matrix, 2,977 charters are identified as related to arbitration.

To assess variation in results, two subsamples are created for each set, with a ratio of 1:1 and 1:5, respectively. The list of curated tri-grams is tested accordingly. To generate (and evaluate) these candidate formulas and especially considering their discriminatory characteristics, we apply vanilla extreme gradient boosting on the charters' skip-grams (n=3, k=2) based on document-term matrices as vectorizers and closely examine their metrics[13], i.e., gain, cover, and weight.

For the given task, several classificational methods are explored, including random forests and multinomial naive bayes. Given its proven reliability[14] as well as its use in low-resource settings, we employ a linear-kernel support vector machine (SVM), without further optimization for cost-efficiency. A few embedding options are tested, including term frequency/inverse document-frequency (TF/IDF) as well as document-

---

[12] Gerard Salton, Anita Wong, and Chung-Shu Yang. “A Vector Space Model for Automatic Indexing.” Communications of the ACM 18 (11) (1975): 618.
[13] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Association for Computing Machinery, 2016), 785–94.
[14] Sida Wang and Christopher D. Manning. “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume* 2 (Association for Computational Linguistics, 2012), 90–94.

term matrices (n=1,2,3; n=2,3/k=1,2). For generating formula candidates, class ratios are considered.

As expected, we find that uni-grams do poorly for the less balanced datasets, as reflected in <= 50 % in recall. Their balanced counterparts pose macro-averages of around 87 %. Bi-grams mostly do equally poorly. For uni-grams, the initial term list performs better than the formulas. For bi-grams, the formula list exhibits macro-averages of 93 %, and clearly shows notions of domain-relevant formulaicity, such as the segments ‘una parte’, ‘andern tail’, ‘bono pacis’.

As anticipated, this trend increases with tri-grams, reaching a 97 % macro-average for recall, and indicating significant features (e.g., ‘dem andern tail’), while the term-based classification suffers from over-generalization, as reflected in negative feature scores for very general phrases, e.g., ‘mit allen den’.

This rise in performance continues when employing simple minimal skip-grams (n=2, k=1). However, for all subsets, longer skip-grams decrease the overall results considerably, indicating potentially robust parameters on the level of grams.[15] The best constellation is found in embedding the curated tri-grams in tri-grams with a count vectorizer, reaching a high point of 99 %; a similarly high score is achieved with this set using skip-grams, while here the initial term-list as classes is worse than a coin-toss.

For all naive n-gram runs, recall and precision for *non*-arbitration charters is consistently around and above 90 %. This indicates that, overall, they appear to be more easily discriminated than classified: across all runs, determining non-arbitration is mostly ‘easier’ than a charter being about arbitration.

Benchmarking the classification is repeated by embedding abstracts as well, which leads to high precision and recall values of above 90 %, and reaching highs of 98 % across all sets. This performance underscores the fundamental role that abstract

[15] Naturally, beyond a certain threshold, longer sequences that are transformed embed worse than shorter ones with regards to model performance. However, shorter sequences are also less accessible and open to direct interpretation without considering their co-occurrences.

metadata can play in enhancing the accuracy of classification, and as a proxy to other languages.

All trends are generally robust considering the non-sampled sets as well. The class ratio of the tri-gram-based set is 1:12. Applying the best SVM setting on it still results in macro-scores of >= 97 %.

Qualitative insights

The new formulas are mostly Latin tri-grams that domain-experts allocate to three categories: (1) general charter formulas, (2) formulas from the field of medieval law, (3) formulas from the field of arbitration. By interpreting the list that is sorted by feature importance, we see that three out of five top features are related to arbitration, delimiting themselves from more general legal documents and language.

- pro bono pacis (2)
- arbitratores seu amicabiles (3)
- appellatione remota fine (2)
- alto et basso (3)
- dem andern tail (2)
- et amicabiles compositores (3)
- ex una parte (2)
- ex parte vna (2)
- omnibus et singulis (1)
- cum inter nos (1)
- quod si non (1)
- baider tail red (1)
- ex altera super (2)
- ab utraque parte (2)
- inter nos et (1)
- pacis et concordie (2)
- red vnd widerred (2)

It appears that medieval arbitration charters are closely modeled after contemporary court-related documents, as visible in the high proportion of category 2. While this is interesting, it is not entirely surprising. More importantly, three formulas are identified that are clear indicators of arbitration. While two have been known so far, we retrieve a new candidate (‘alto et basso’) that is a more noteworthy finding. Complementing our finding with its occurrences in other charter databases, we figure that this formula is being used over centuries and across many regions of Europe.

Conclusion

This study has demonstrated the benefits of using formulas over single terms for soft-classifying text embeddings. Moreover, it is evident that the feature importance exhibited by the training processes of a model as well as the manual evaluation of formulas can benefit from considering language segments as text-type discriminators.

We utilized soft classification to identify novel and potentially significant Latin formulas as markers for arbitration in charters. It sets the foundation for future research in this area that could incorporate more diverse and multimodal approaches, including the analysis of visual elements, e.g., seals and manuscript images, to further differentiate them. Also, future enhancement and innovations in methodology can be easily derived, including other variables, e.g., space, time, institutions, language, to provide further insights into the distribution and characteristics of charter types.

Method-wise, a major takeaway is the considerable performance in the face of an enormous potential for improving the given workflow. This includes refining text preprocessing techniques to improve results while preserving the integrity of formulaic structures, e.g., by employing advanced lemmatization or masking/encoding methods. While these improvements can lead to greater robustness, we advocate for cautious application to avoid undue trust in the model predictions.

We also find that the exploration into the discriminatory power of individual words and sequences suggests a promising direction for future research. It could contribute to the refinement of classification models by incorporating the discriminative scores of these formulas more explicitly, especially so in more general applications of information retrieval. Most importantly, it aids in the creation of ground truth datasets.

References

Ambrosio, Antonella, Sébastien Barret, and Georg Vogeler, eds. 2014. *Digital diplomatics: the computer as a tool for the diplomatist*. Böhlau.

Atzenhofer-Baumgartner, Florian. 2023. “Quantifying Formulaic Flexibility of Middle High German Legal Texts”, Master’s Thesis. University of Graz. https://doi.org/10.5281/ZENODO.8141830.

Brommer, Sarah. 2018. *Sprachliche Muster: Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte*. De Gruyter. https://doi.org/10.1515/9783110573664.

Bubenhofer, Noah. 2009. *Sprachgebrauchsmuster*. De Gruyter. https://doi.org/10.1515/9783110215854.

Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–94. Association for Computing Machinery, 2016. https://doi.org/10.1145/2939672.2939785.

Hunston, Susan. 2010. *Corpus Approaches to Evaluation*. Routledge. https://doi.org/10.4324/9780203841686.

Nicolaou, Anguelos, Daniel Luger, Franziska Decker, Nicolas Renet, Vincent Christlein, and Georg Vogeler. “Efficient Annotation of Medieval Charters.” In Document Analysis and Recognition – ICDAR 2023 Workshops, edited by Mickael Coustaty and Alicia Fornés, 14193:284–95. Springer, 2023. https://doi.org/10.1007/978-3-031-41498-5_20.

Salton, Gerard., Anita. Wong, and Chung-Shu, Yang. 1975. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18 (11): 613–20. https://doi.org/10.1145/361219.361220.

Stubbs, Michael, and Isabel Barth. 2003. “Using Recurrent Phrases as Text-Type Discriminators: A Quantitative Method and Some Findings.” *Functions of Language* 10 (1): 61–104. https://doi.org/10.1075/fol.10.1.04stu.

Wang, Sida, and Christopher D. Manning. 2012. “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, 90–94. Association for Computational Linguistics.