Information Extraction on Delegated Jurisdiction in the Iberian Peninsula of the 12th Century

Marlene Ernst & Alina Ostrowski | DiDip2022
DFG project (2022–2025)
„Delegierte Gerichtsbarkeit auf der Iberischen Halbinsel im 12. Jahrhundert. Struktur und computergestützte Analyse eines urkundlichen Massenkorpus“
Iberia Pontificia I–VI

Göttingen Academy of Sciences and Humanities / papal charters of the early and high Middle Ages

Additionally:
- Regesta Imperii
- Papsturkunden in Spanien I–III
- Papsturkunden in Portugal

➢ Overall approx. 1,000 papal charters
Digital Humanities Perspectives

- Actors / Historical Network Analysis
- bishops, abbots, in the late 12th century also dignitaries of the cathedral chapters, their relations with each other and with the papacy
- litigants and judges involved

Formulaic Language / Text Reuse Detection
Detection and Investigation of Formulaic Language in Papal Charters
“Formulaic phrases [formelhafte Wendungen] are in the broadest sense:

• One-word expressions, typologically heterogeneous combinations of several constituents, or whole sentences and/or texts,

• which need to be understood holistically,

• are at different (even not yet completed) stages of formal, semantic and functional conventionalization,

• are based on usage conventions of a linguistic community whose established cultural (also communicative) experiences and knowledge stocks they hand down, and

• which can be characterized by a strong functionalization in the communication process or in the text structure.”

Cited according to Filatkina 2018, S. 164, translated from German to English with the help of DeepL.
Formulaic Language – Working Definition

1. Formal structure of the papal charters
   - *intitulatio, inscriptio, datatio, etc.*

2. Indicated by set phrases structuring the elements of a charter.
   - *episcopus servus servorum dei, venerabili fratri, Datum per manum, etc.*

3. Recurring phrases with a constant lexical/syntactical form signifying a special concept connected to the topics discussed in the charter
   - *ad aures nostras pervenit...; Si qua vero partium duxerit appellandum...*
**Known Formulas**

**Questions:**
- What formulas were still/already in use in the 12th century?
- How does their early lexical/syntactical form or their meaning differ from the fully developed expression?

**Sources:** Formularies of the 11th and 13th century

**Research Methods:** Information Retrieval methods (string matching) and Text Re-Use Detection

---

**Unknown Formulas**

**Questions:**
- Which set expressions are used in the charters of our corpus?
- Are there formulas that are unique to the 12th century?
- How is their distribution related to individual judges delegate or to special cases?

**Research Method:** Text Re-Use Detection
Formulaic Expressions and Charter Section Detection
Using formulaic expressions to detect diplomatic sections


• 14th century Bohemian royal charters
• String similarity scores (Cosine Distance and TF/IDF)
• Good evaluation results for well defined sections, e.g. intitulatio

Fig. 1 from Galuščáková/Neužilová 2018: „An overview of the information retrieval-based method.”
Using formulaic expressions to detect diplomatic sections

In: Étienne Doublier, Daniela Schulz und Dominik Trump (ed.): Die Historischen Grundwissenschaften heute, p. 139–166.

• 10th-12th century German royal charters
• String similarity scores (Token Substring- and Subset-Matching, Jaccard Coefficient over BOW, positional arguments)
• High evaluation scores for eschatocol and protocol detection

Example of an unannotated input charter and its automatically annotated output. The highlighted formulas are characteristic for their corresponding charter section.
Historical Text Reuse Detection (TRD)
Research on Historical Text Reuse Detection

Marco Büchler (2013): Informationstechnische Aspekte des Historical Text Re-use (PhD thesis, Leipzig University) ➔ Basic publication on the subject of Historical Text Reuse

Example studies:


Types of Text Reuse

- **Syntactic Reuse**
  Verbatim/near verbatim quotes, idioms, editions etc.

- **Semantical Reuse**
  Ideas and semantic concepts in allusions, paraphrases etc.
TRACER

Status Quo & Next Steps


- searchable PDFs; accuracy rate of 98.19 %


- also e-publication; accuracy rate of 100 %


- searchable PDF of modern type; accuracy rate of 100 %

Mansilla, D. (ed.): La documentación pontificia hasta Inocencio III: (965–1216), Rome 1955

- poor quality scan; accuracy only 95.75 %
Thank you for your attention!