Building the Penitentiary Document Corpus (PeDoCo) for NLP: Balancing Data Complexity and Uniform Data Structure

Hanna-Mari Kupari, Timo Korkiakangas and Veronika Laippala

The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB)
March 5–7, 2025

Tartu, Estonia

Contents:

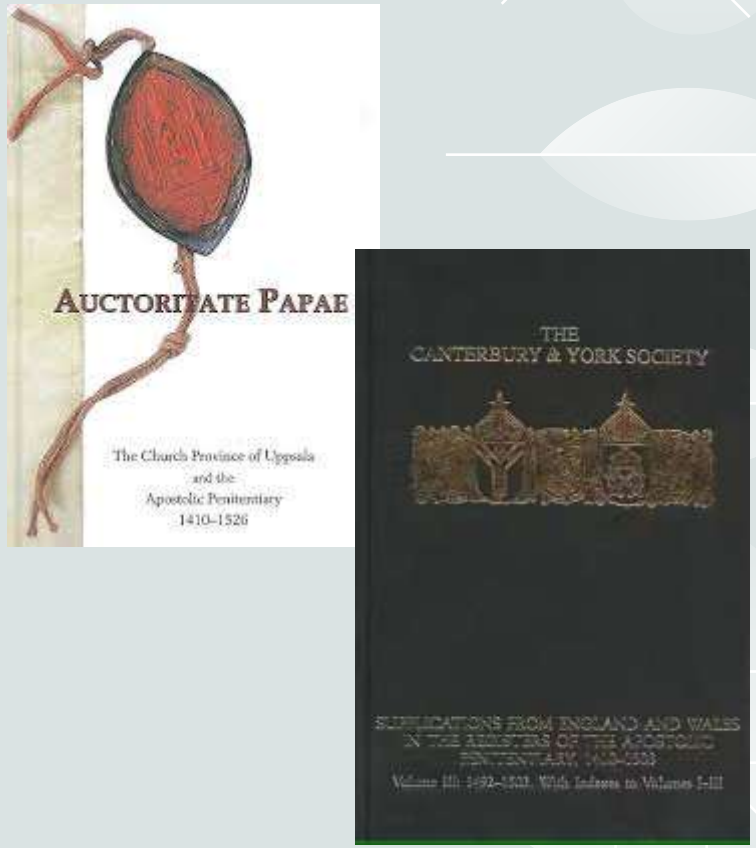
Slides as pdf to download

Introduction

THE PENITENTIARY DOCUMENTS
OF THE VATICAN

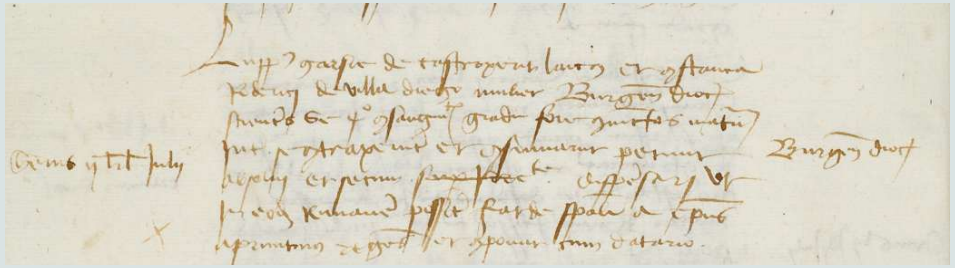
The Apostolic Penitentiary

- Cases of diverse petitions from the Apostolic See of the Vatican
- Late Middle Ages, c. 1390-1530
- A “tribunal of mercy”, responsible for issues relating to the forgiveness of sins in the Church
- Copybooks survive and edited according to modern principles for historians using quantitative methods

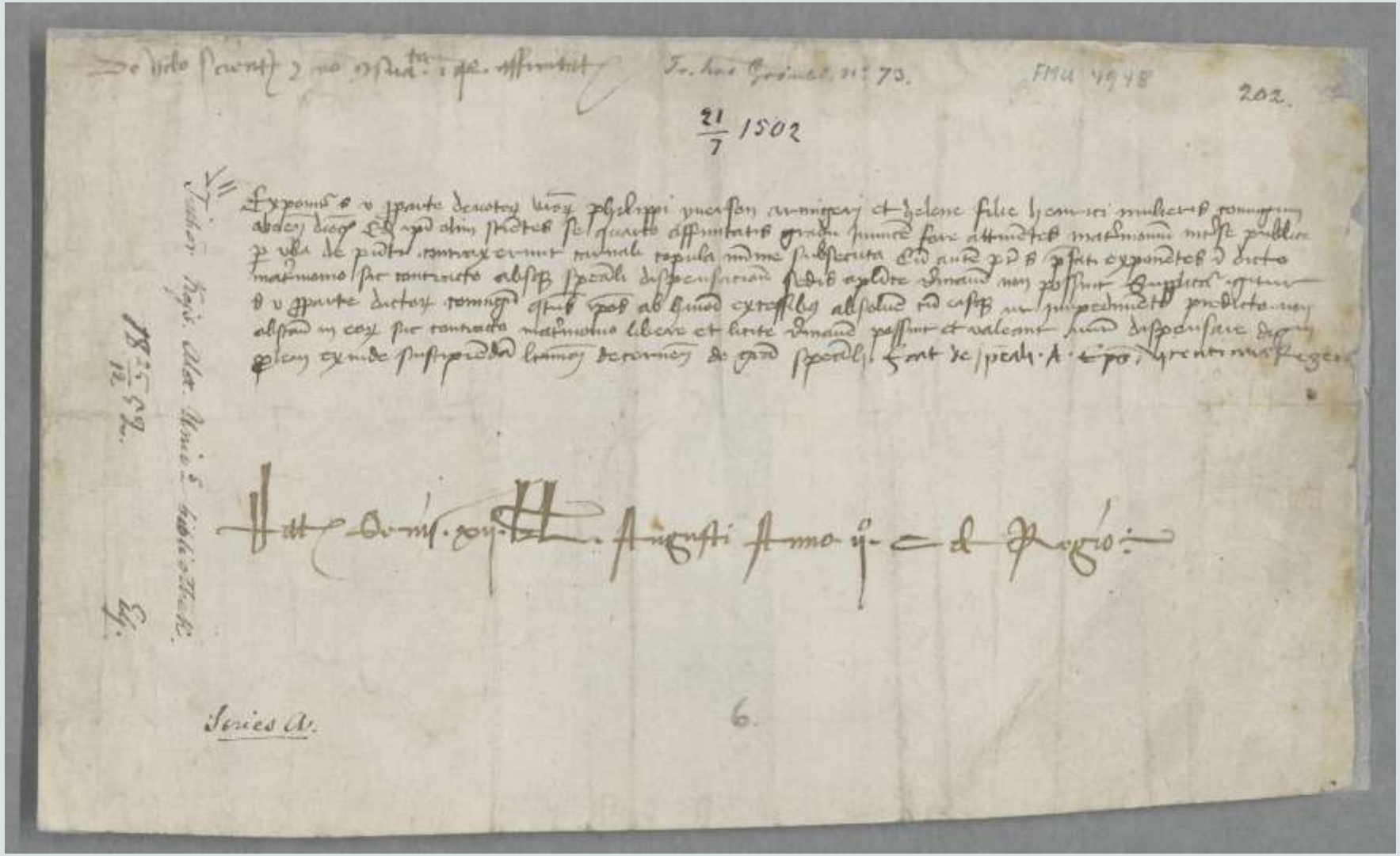


The steps of the penitentiary documents turning into a machine-readable resource

- Language use context in the Middle Ages
- Process to handle your case with the office
- Document to have the official pardon
- Internal records to keep track of approved cases
- Modern day editions to improve readability
- Digital resource to increase our knowledge of Late Medieval Latin language use

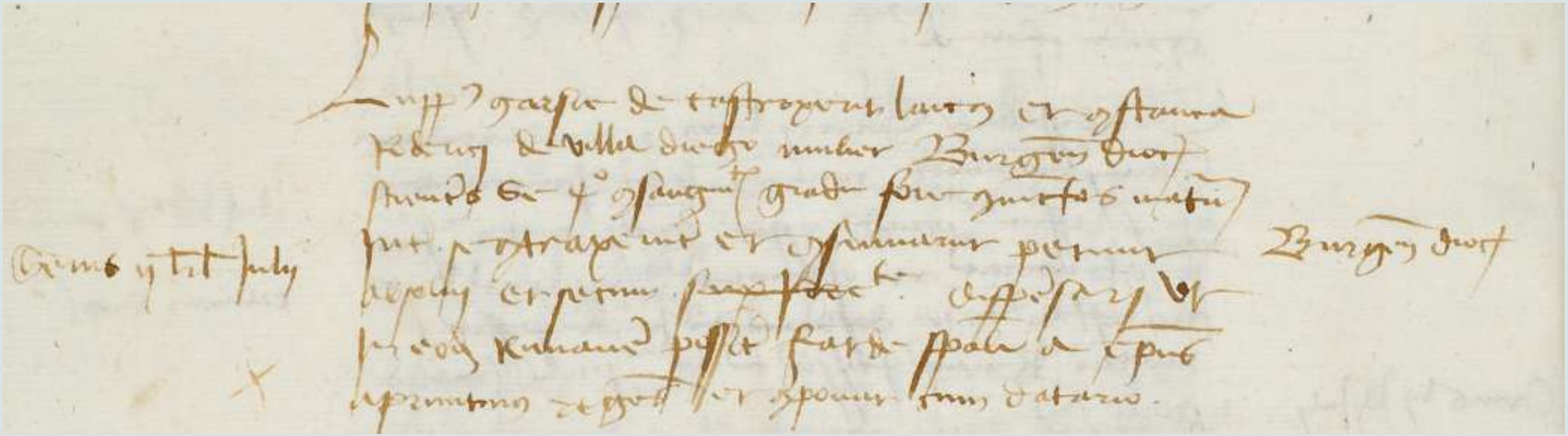


Step 1: An original petition from a Finnish local archive



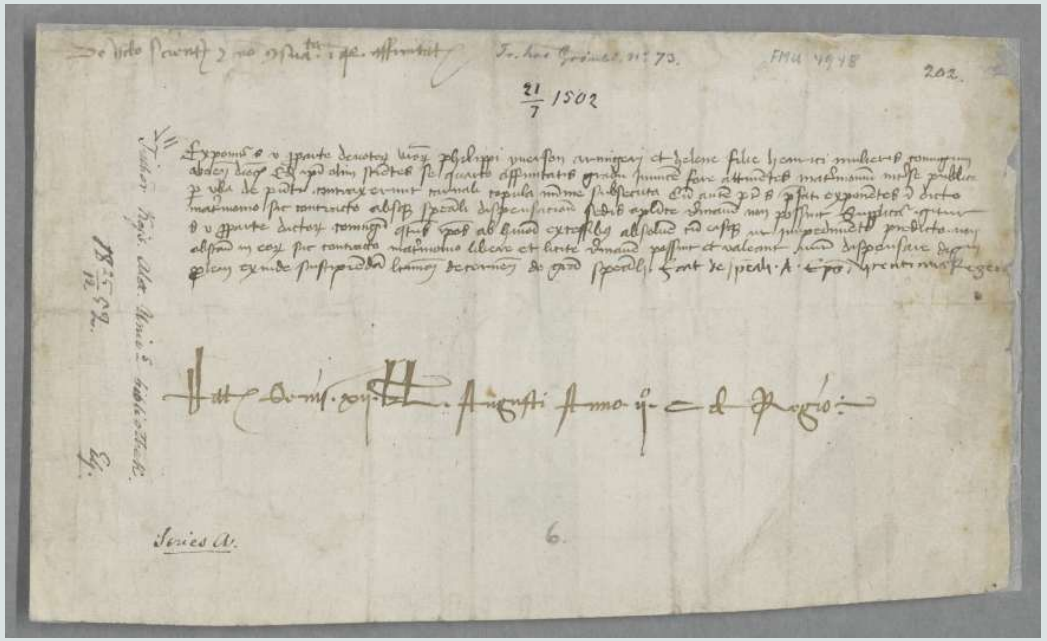
https://df.narc.fi/document/4948#picture

Step 2: A register copy of the same document

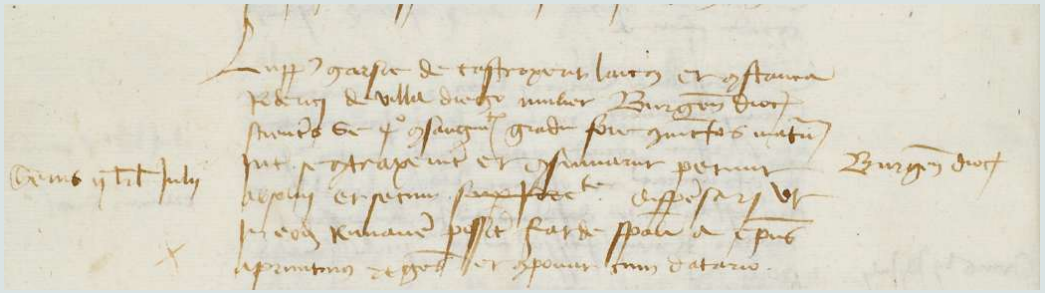


With permission from The Apostolic Penitentiary
APA Reg Matrim. et Divers. 8, 63 verso

Only a short summary is in the register copies



Exponitur sanctitati vestre pro parte devotorum vestrorum Philippi Yverson armigeri et Helene filie Henrici mulieris coniugum Aboensis diocesis,…



Philippus Yverson armiger et Helena Henrici mulier Abuensis diocesis quarto affinitatis gradu coniuncti petunt similem gratiam sibi fieri. Fiat de speciali. Antonius episcopus Apruntinus regens.

Background & Aims

A HIGH-QUALITY LANGUAGE RESOURCE

Background

- The need to have more diverse sources:
- Many of the current sources are:
  - Official documents or literary texts
- Penitentiary documents offer a window into all social strata
- To understand history
- Better understanding of language use

Aims

- Several digital resources:
  1. A database in TEI XML
  2. Three new tools (parsers) to automatically add morpho-syntactic annotation
  3. A sample treebank of the PeDoCo
- To study variation in language use by computational methods

TEI XML structure

A DATABASE

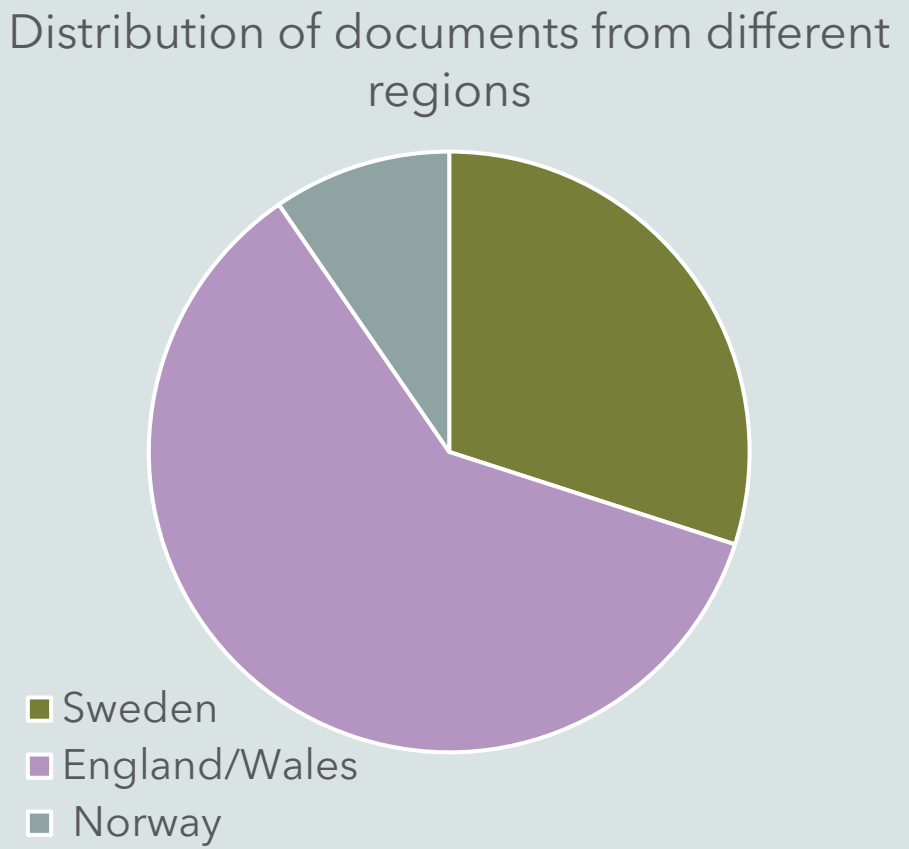
Finding a correct place for everything

A seemingly trivial task of just sorting
A lot of close reading to understand the necessary classification

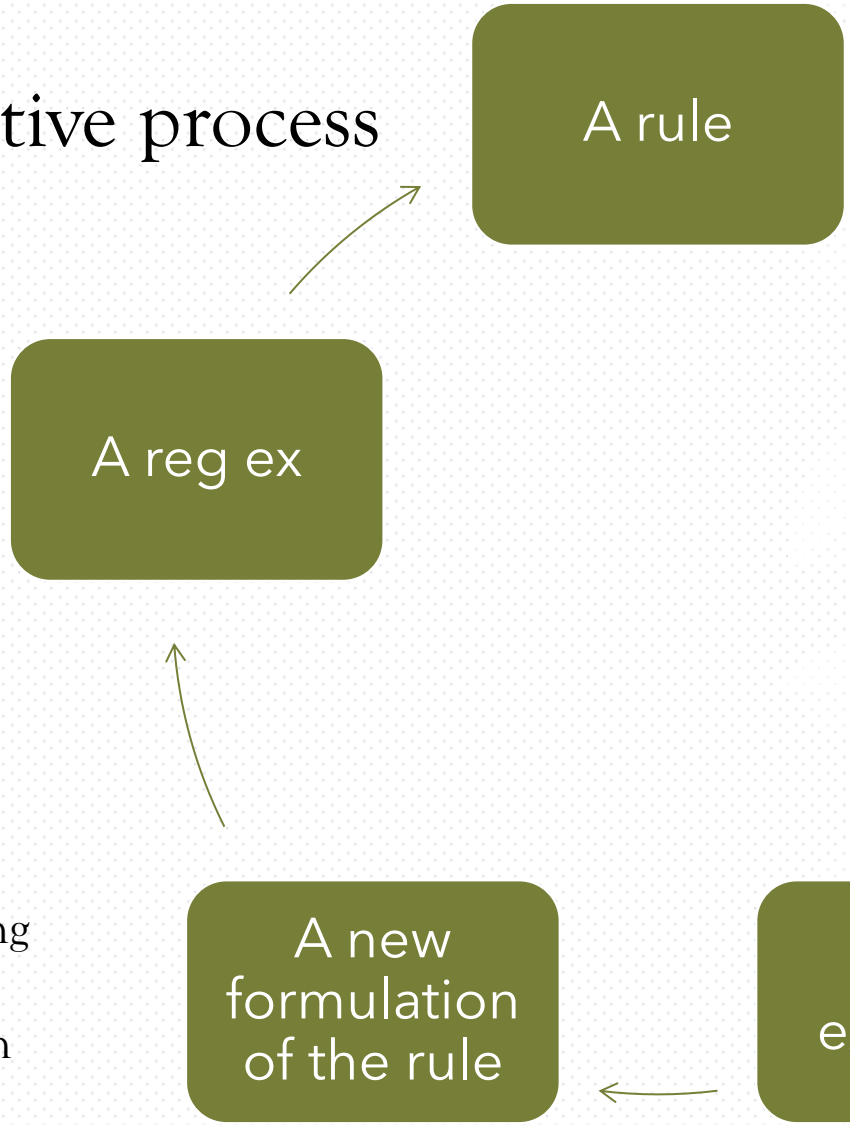
Pictures: K-ruoka and Wikipedia

Step 3: From pdf documents to machine readable database

- The editions of 3 regions in 3 different formats
- An iterative process that is made completely open
- All code in https://github.com/HannaKoo/PeDoCo



Step 4: A repetitive process



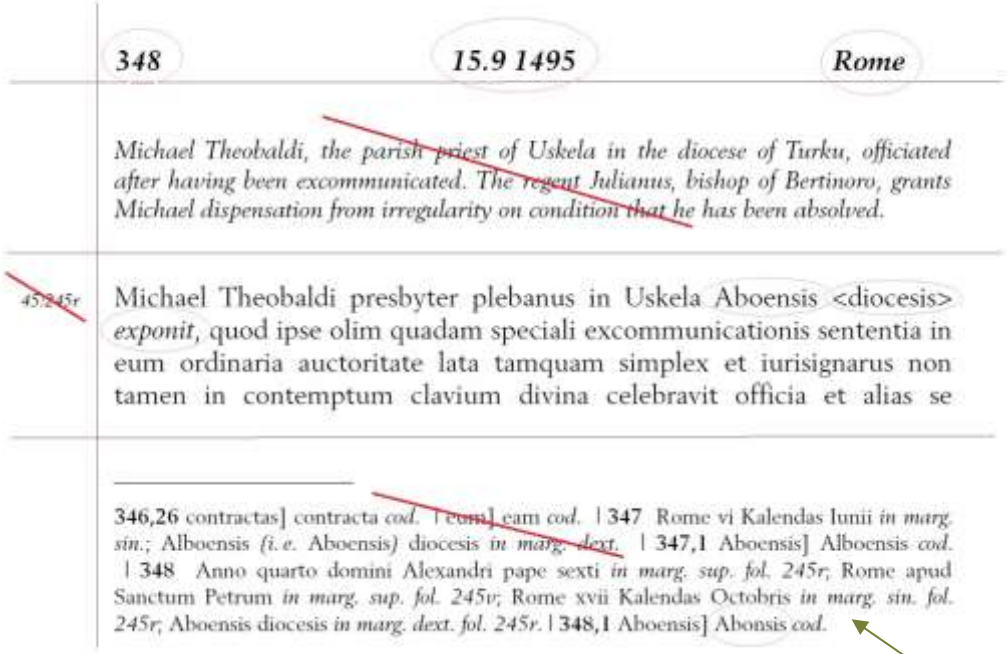
All dates should be YYYY-MM-DD

Look for number (number) space number (number) period followed by 4 numbers

Realize that you can't match all documents with reg ex, e.g. a two-day interval

Make a policy of getting one machine readable date and take variation into account

Manually check that sums match up and make 1-new reg ex'es



348 **15.9 1495** **Rome**

Michael Theobaldi, the parish priest of Uskela in the diocese of Turku, officiated after having been excommunicated. The regent Julianus, bishop of Bertinoro, grants Michael dispensation from irregularity on condition that he has been absolved.

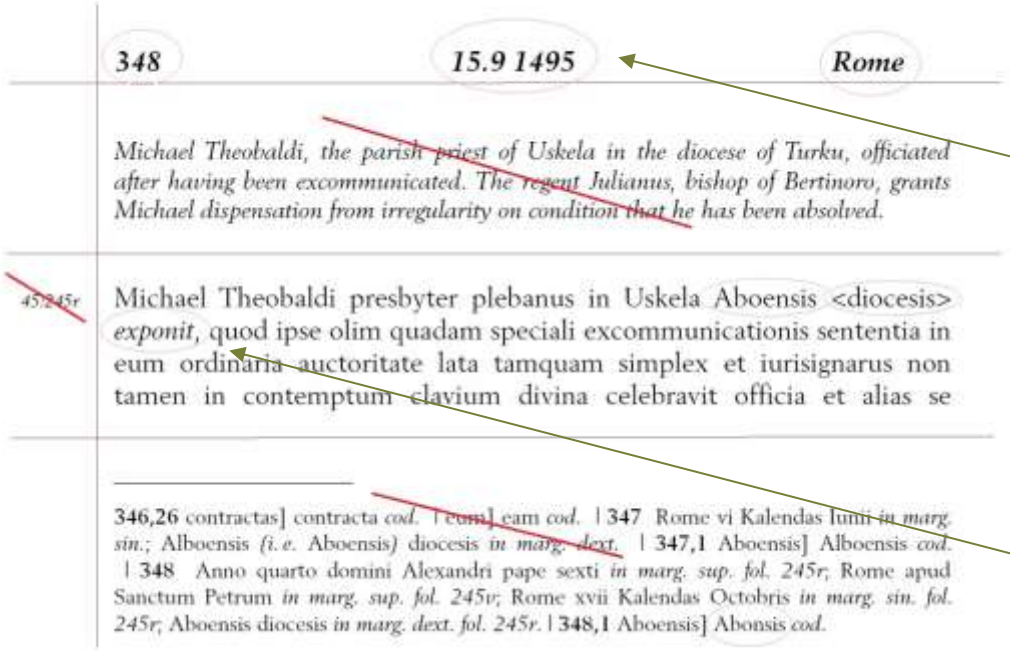
45:245r Michael Theobaldi presbyter plebanus in Uskela Aboensis <diocesis> *exponit*, quod ipse olim quadam speciali excommunicationis sententia in eum ordinaria auctoritate lata tamquam simplex et iurisignarus non tamen in contemptum clavium divina celebravit officia et alias se

346,26 contractas] contracta *cod.* | eum] eam *cod.* | **347** Rome vi Kalendas Iunii *in marg. sin.*; Alboensis *(i.e.* Aboensis*)* diocesis *in marg. dext.* | **347,1** Aboensis] Alboensis *cod.* | **348** Anno quarto domini Alexandri pape sexti *in marg. sup. fol. 245v*; Rome apud Sanctum Petrum *in marg. sup. fol. 245v*; Rome xvii Kalendas Octobris *in marg. sin. fol. 245r*; Aboensis diocesis *in marg. dext. fol. 245r.* | **348,1** Aboensis] Abonsis *cod.*

Screenshot from *Auctoritate Papae* by Risberg and Salonen (2008)

The XML framework

- A human and machine-readable text-based annotation to differentiate between different parts of text
- A basic structure of <tag> text </tag>
- Tags can contain information on words, editorial choices and background information
- Removing from the database linguistically irrelevant data



348 **15.9 1495** **Rome**

Michael Theobaldi, the parish priest of Uskela in the diocese of Turku, officiated after having been excommunicated. The regent Julianus, bishop of Bertinoro, grants Michael dispensation from irregularity on condition that he has been absolved.

45245r Michael Theobaldi presbyter plebanus in Uskela Aboensis <diocesis> *exponit*, quod ipse olim quadam speciali excommunicationis sententia in eum ordinaria auctoritate lata tamquam simplex et iurisignarus non tamen in contemptum clavium divina celebravit officia et alias se

346,26 contractas] contracta *cod.* | eam] eam *cod.* | **347** Rome vi Kalendas Iunii *in marg. sin.*; Alboensis *(i.e.* Aboensis*)* diocesis *in marg. dext.* | **347,1** Aboensis] Alboensis *cod.* | **348** Anno quarto domini Alexandri pape sexti *in marg. sup. fol. 245r*; Rome apud Sanctum Petrum *in marg. sup. fol. 245v*; Rome xvii Kalendas Octobris *in marg. sin. fol. 245r*; Aboensis diocesis *in marg. dext. fol. 245r.* | **348,1** Aboensis] Abonsis *cod.*

Example

```
<text n="300" source="AP" onum="348" type="diversis"
bundle="n" several_witnesses="n"><front>

<docDate><date when="1495-09-15"/></docDate>
<placeName type="place-issue">Rome</placeName>

</front><body><p>. . . <choice><corr>Aboensis</corr>

<sic>Abonsis</sic></choice> <supplied reason=
"omitted-unintentional">diocesis</supplied>

<supplied reason="omitted-intentional">exponit
</supplied>. . . </p>

</body></text>
```

TEI XML

- Since the formulation of different tags in free
- A framework has been set up to propose some shared guidelines
- Text Encoding Initiative for diverse types of data
- The TEI Consortium is a nonprofit membership organization
- https://tei-c.org/
- The editions of three editorial regions of the PeDoCo are structured in this format

Table 1
Document and word counts of the PeDoCo.

	England/Wales	Sweden	Norway	Total
Documents in Latin	913	453	145	1,511
Words in Latin	129,879	61,174	20,345	211,398

Future outlooks

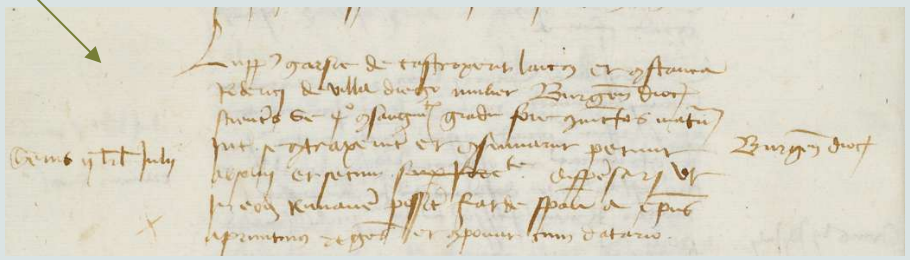
THE PEDOCO TREEBANK: PENITENTIARY DOCUMENTS INTO A LARGE LANGUAGE RESOURCE

Step 4: A small test set

- Creating a new treebank
- Using the Universal Dependencies (UD) framework
- First a test set (1,200 words) is used to evaluate the performance of parsers, like Stanza or Trankit, to grammatically analyse the text
- The best parsers (Kupari et al., 2024) are then used to make the analysis of the 200,000 PeDoCo corpus

Conclusions

AIMING TOWARDS HIGH QUALITY COMPUTATIONAL MODELS



Distant reading

```
<text>
<h2 num="88" bundle="n" several_wittnesses="y">88
<date when="1460-07-21">21.7 1460</date> Siena</h2>
<h3 num="88" version="a">88a</h3>
<version>
Exponitur sanctitati vestre pro parte devotorum
vestrorum
Philippi Yverson armigeri et Helene filie Henrici
mulieris coniugum
Aboensis diocesis, - - - </version>

<h3 num="88" version="b">88b</h3>
<version>
Philippus Yverson armiger et Helena Henrici mulier
Abuensis
diocesis quarto affinitatis gradu coniuncti petunt
similem gratiam sibi
fieri. Fiat de speciali. Antonius episcopus
Apruntinus regens.
</version>
</text>
```

The benefits of the computational and treebank approach

- From original application
- To register copies
- A modern-day edition
- A structured text to organize into machine readable format
- Parsing tools (Kupari et al. 2024) to add linguistic annotation – a new treebank
- A better understanding of medieval Latin language use that can be motivated with computational evidence from large datasets

References

Edition:

Auctoritate Papae. The Church Province of Uppsala and the Apostolic Penitentiary 1410–1526. Sara Risberg (edition), Kirsi Salonen (introduction). Stockholm: National Archives of Sweden 2008.

Papers:

Kupari, Hanna-Mari, Timo Korkiakangas, and Veronika Laippala. 2025. "Building the Penitentiary Document Corpus (PeDoCo) for NLP: Balancing Data Complexity and Uniform Data Structure". *Digital Humanities in the Nordic and Baltic Countries Publications* 7 (2). https://doi.org/10.5617/dhnbpub.12301.

Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. 2024. Improving Latin Dependency Parsing by Combining Treebanks and Predictions. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 216–228, Miami, USA. Association for Computational Linguistics.

Thanks!

Editors

The Penitentiary archive

Institutum Romanum Finlandiae

Keep in touch!

LinkedIn
hmknie@utu.fi

https://github.com/HannaKoo