Automatic medieval charters structure detection : A Bi-LSTM linear segmentation approach

Sergio Torres Aguilar, Pierre Chastang, Xavier Tannier

Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Automatic medieval charters structure detection : A Bi-LSTM linear segmentation approach

Sergio Torres Aguilar[1], **Pierre Chastang**[2], **Xavier Tannier**[3]

[1]École nationale des chartes, Centre Jean Mabillon, France
[2]UVSQ-Université Paris-Saclay, France
[3]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, France

Corresponding author: Sergio Torres , sergio.torres@chartes.psl.eu

Abstract

This paper presents a model aiming to automatically detect sections in medieval Latin charters. These legal sources are some of the most important sources for medieval studies as they reflect economic and social dynamics as well as legal and institutional writing practices. An automatic linear segmentation can greatly facilitate charter indexation and speed up the recovering of evidence to support historical hypothesis by the means of granular inquiries on these *raw*, rarely structured sources. Our model is based on a Bi-LSTM approach using a final CRF-layer and was trained using a large, annotated collection of medieval charters (4,700 documents) coming from Lombard monasteries: the CDLM corpus (11th-12th centuries). The evaluation shows a high performance in most sections on the test-set and on an external evaluation corpus consisting of the Montecassino abbey charters (10th-12th centuries). We describe the architecture of the model, the main problems related to the treatment of medieval Latin and formulaic discourse, and we discuss some implications of the results in terms of record-keeping practices in High Middle Ages.

Keywords

medieval charters, automatic structure detection, linear text segmentation, Latin NLP, medieval digital studies

I INTRODUCTION

The wording of most medieval charters was framed by well-defined documentary models. Charters are essentially property deeds or privileges and being legally binding documents, they must match the structure of stereotyped models and formularies to constitute a valid document gathering specific details of exchanges. Just like other formularies, those used by charters are normally designed to classify information using a scheme presenting modules or sections. The study of diplomatics knows these sections as “diplomatic parts of discourse”. The studies about charters and their configuration have been key in understanding the evolution of writing and legal traditions in medieval Europe. In that sense, structure detection can help provide an indexed structure to this kind of corpora, allowing to deploy information retrieval systems; it can greatly facilitate a larger-scale implementation of diplomatics and historical research methods.

This work aims to 1) use the only digital medieval corpus annotated with parts of diplomatic discourse, to create a supervised model that can automatically recognize these parts; 2) quickly provide a query structure for medieval charters able to facilitate the retrieval of specific informa-

tion from massive datasets; 3) enable a massive comparison of charters at the level of complex units with complete meaning such as phrasemes, formulae or clauses.

II RELATED WORKS

In the last years, many medieval corpora have become available, especially from massive digitalization of 19th and 20th century erudite and critical editions. Among the most important are the CBMA[1] and HOME-Alcar[2] from French charters, the DEEDS, from Anglo-Saxons charters[3], the Diplomata Belgica (DiBe), from Belgian charters[4] and the CDLM, from Lombardian charters[5]. All these projects provide different kinds of structured data for thousands of charter collections and cartularies dating from he ninth to the 13th centuries – but the CDLM is the only one to provide an annotation of the parts of diplomatic discourse, which is a time-consuming task when done manually. In the field of the digital edition, the CEI (Charters Encoded Initiative) has offered an XML-TEI extension to annotate charter editions (Burkard et al. [2008]) based on diplomatic definitions from the famous manual of Diplomatics International Commission (Ortí [1997]) since regular TEI is not adapted to describe these specialized structural documents. However, this is a work in progress and corpora are not fully available yet. Current literature shows only one work in the field of automatic structure extraction of medieval charters, which uses a hidden Markov model to detect sections in a collection of 57 Czech royal charters from the 14th century (Galuščáková and Neužilová [2018]). Best results show high precision, but very poor recall, which is partially explained by the small size of the corpus.

More broadly, our research is related to works about linear text segmentation, text structure detection, and sentence level classification which are popular fields in natural language processing (Achananuparp et al. [2008]). Some recent advances on medical records predicting and capturing sentences structure based on distributional similarity can be considered close to our approach (Jagannatha and Yu [2016]). The RNN approaches and more specifically the LSTM networks seem to provide best results in similar tasks: Koshorek et al. [2018] use a hierarchical LSTM model to predict the table of contents in a huge English Wikipedia dataset; a custom Segment Pooling LSTM is used by Barrow et al. [2020] to build a model for joint segments boundary detection and segment labeling tasks using Wikipedia sections headers as dataset labels; another work by Varma [2018] proposes a language-agnostic deep-learning approach (Bi-LSTM) to predict the paragraph labels in a text. The unsupervised approaches by the means of lexical clustering and semantic relatedness (Glavaš et al. [2016]), are popular in this field, due to paucity in sentence-level annotated datasets, but they are inefficient and require long execution times.

III CORPORA DESCRIPTION

3.1 The medieval Lombard corpus

The CDLM (*Codice diplomatico della Lombardia Medievale*) is a corpus made public by the University of Pavia in 2006 in the form of an XML edition containing about 5,300 edited charters (Ansani [2006]). Documents come from many monastic and ecclesiastical institutions as well as from the Bergamo civil archives; they range from the ninth to 13th century and specially from mid-11th and 12th century (78% of the corpus). Like many other charter collections, the

[1] http://www.cbma-project.eu/
[2] https://zenodo.org/record/5600884
[3] https://deeds.library.utoronto.ca/
[4] https://www.diplomata-belgica.be/
[5] http://www.lombardiabeniculturali.it/cdlm/

CDLM is composed mostly of land exchange private charters. Abbeys and monasteries were large landowners since they were the main recipients of land donations before the 13th century, and since they launched an extensive movement of land domination as of the Gregorian Reform (mid-11th century onwards). Most charters are preserved by abbeys and monasteries because they want to keep a full historical record of their properties acquired by donation or purchase. Many other charters come from public institutions: letters and bulls from Apostolic chancelleries and bishops' offices, diplomas and privileges from royal and noble chancelleries testifying to a very active exchange network. The classification of documents proposed by the CDML editors is quite precise because of its extensive typology of legal actions, but in general 5 main document types emerge : charters (almost 80%), notices (charter summaries), diplomas, letters and bulls; as well as 5 main legal actions: donations, purchases, land rents, judgements and allowances.

Besides, the CDLM includes charters from notarial tradition, which means, they are mostly produced by professional scribes. The notary institution founded on Late Roman traditions was well-established in Northern Italy, when it had almost disappeared from most of Western Europe after the eighth century (Bautier [1989]). Consequently, since the 10th century, Lombard charters had used a stricter and more formal diplomatic discourse integrating a large variety of legal validity clauses and using authentication signs such as stamps and subscriptions. In fact, the drafting of a charter is a complex process : a charter must be validated, revalidated and evaluated by the stakeholders, the notary, the witness and even the authorities before gaining legal value. Therefore, some parts of discourse identified in Lombard charters are not found in other parts of Europe until the mid-13th century, or are only found in documents coming from royal and apostolic chancelleries. However, even charters produced by semi-professional, non-professional scribes or *tabellion* officers – that are quite common in ecclesiastical charter collections between the 10th and the 13th centuries – generally follow many widely recognized notarial practices from early medieval formularies; hence the discursive models generally displaying more similarities than specific differences.

3.2 The Montecassino cartulary

The Montecassino cartulary (also known as *Registrum Petri Diaconi*) is a volume composed in 1131-1133, gathering copies of documents relating to the famous Montecassino abbey's properties and rights. The volume contains 717 acts of a large variety including public (bulls, royal and prince privileges, precepts) and private acts (donations, sales and farming contracts) reflecting the activity of a large landowner in the Lazio and Abruzzo regions. These charters range from the mid-10th to the early 12th century (Chastang et al. [2009]), thus coinciding in time with the CDLM corpus. As annotating diplomatic parts is a difficult task to perform manually, a section of 200 charters ranging from the 10th to the 12th century was chosen to build a sub-corpus, serving here as a validation corpus to evaluate our model's robustness.

3.3 LTS and diplomatics charters

Linear text segmentation (LTS) is the segmentation of a text into contiguous sections. Each section is defined by a shared semantic and lexical structure, all sections also being interdependent. LTS helps provide a basic structure to a text before tasks such as information retrieval and topic classification are performed. LTS is a traditionally challenging task and can vary according to the subject and origin of the text because almost each area has developed specific writing models to convey information. Charters, just like other documents designed to claim a right or keep a memorial record of a juridical fact, place great importance on following a characteristic writ-

ing form, giving it validity even in the absence of proper validation signs (stamps or seals, sign manuals, consents). This writing form or model typically consists of a sequence of utterances – the parts of the discourse – designed to gather different details of a transaction. (See an example of diplomatics sections in the charter in figure 1). In general, the writing practice suggests using some model or other according to the type of document and the nature of the legal action. The scribe reproduces a model, but each copy carries multiples differences depending on multiple factors: the quality of the participants, the regional operating traditions that can suggest many variations from the original model (Fichtenau [1957]), the request for particular details of the exchange, or even the scribe's mistakes or own personal taste.

1 < Protocol >

DTCRON INSCRIPTIO	[1]Anno ab incar(nacione) domini nostri Iesu Christi milesimo centesimo octavodecimo, die dominico mensis octubris, indicione undecima. [2]Ecclesiæ Sancti Iohannis site foris civitatis Brixiæ,
INTITULATIO	[3]nos Cocalius et Brixianus germani, de ia(m)dicto loco Caucalio, qui professi sumus lege viv(er)e Longobarda, offertores et donatores in ia(m)dicta ecclesia Sancti Iohannis de fora p(resentes) presentibus diximus :

2 < Text >

EXORDIUM	[4]dum in statu sanitatis humane vitæ cursus peragitur et pleno animo mentis ratio vegetatur, sic debet homo se(m)per cogitare atque dispon(er)e quæ sibi profutura sint, ut eum, cum Dominus de hoc seculo vocari iusserit, non de neglegencia iudicet sed de bono disposito ordinetur ut pius
NARRATIO	[5]Manifestum est nobis qui supra Cocalio et Brixiano, germanis, eo quod non habemus filium nec filiam, quod volumus omnia nostra bona ordinare atque dispon(er)e taliter qualiter hic subter statuerimus et nostra decreverit volu(n)tas pro anime nostræ nostrorumque pa|rentum mercede.
DISPOSITIO	[6]Ideoque volumus et firmiter statuimus atque per hanc cartulam offersionis nostre confirmamus ut a presenti die deveniant omnes res iuris nostri in predicta ecclesia Sancti Iohannis tam mobiles quam i(m)mobiles atque semoventes tam eas quas nunc habemus quam eas quas in antea adquirere potuerimus, inintegrum, in Cocalio et in omni loco ubicumque aliquid de nostro iure inveniri potest, inintegrum, ita ut a presenti die
Sanctio	pars ipsius ecclesie Sancti Iohannis faciat exinde, iure proprietario nomine, quicquid voluerit, [7]sine omni nostra atque heredum nostrorum contradicione, pro anime nostræ nostrorumque parentum mercede.

3 < Eschatocol >

DTTOP	[8]Actum est iuxta ia(m)dictam ecclesiam Sancti Iohannis de fora.
SMR	[9]Signa ++ manuum predicti Cocalii et Brixiani, qui hanc cartulam offersionis pro anime suæ suorumque parentum mer|cede fieri rogaver(unt) ut supra.
SMT	[10]Signa +++++ manuum | Lanfranci de Cologne et Iohannis Curtisi atque item Iohannis de Cocalio et qui Nanus vocatur testium.
SUBSCRIPTIO	[11]+ Petrus sacerdos eiusdem ecclesie Sancti Iohannis vice ipsius ecclesiæ hanc cartulam offersionis accepit.
COMPLETIO	[12](SN) Ego Guido notarius rogatus huius offersionis cartulæ scriptor post traditam co(m)plevi.

Figure 1: Diplomatics sections in a *donatio pro anima* (donation for the soul) charter. Brescia, Lombardy, 1118. Coccaglio and Bresciano, brothers, having no progeny, donate all their possessions to the church of St. John (an English translation can be found in Notes). From the 10th century, these almsgiving exchanges of material for spiritual goods became quite common in charters collections.

Traditional literature proposes a two-level hierarchy – a broader level and a finer one – to describe the parts of diplomatic charters. In the first one, we can distinguish a tripartite model comprising the *Protocol*, the *Text* and the *Eschatocol* and corresponding to the initial, central and final sections of the charters respectively. The juridical action itself is located in the *Text*, the other two sections being formal frameworks where formulation is not necessarily related to this action, but contains the majority of the traditional formal elements required to validate the charter. As a result, both display many formulae and named entities (persons, places and dates) since they act, like in many other discourse practices, as completing formulae producing an

individual document on the basis of a conventional structure. Conversely, in the *Text*, all details and conditions about the transaction are exposed in a freer manner.

The second level of the hierarchy operates inside these three macro-modules. Inside-parts belonging to the initial and final frames are mostly composed of or introduced by formulae containing : invocations (*invocatio*), dates (*data cronica*, *data topica*), stamps (*promulgatio*, *corroboratio*), signs (*subscriptio*, *completio*), religious quotes and complex named entities since they must clearly identify and localize participants of the transaction (*inscriptio*, *intitulatio*, *smt*, *smr*). On the contrary, the central frame (the *Text*) can present a freer and richer writing form, since it deals with specific details of the exchange as well as with descriptions of the lands and goods, or terms and conditions of the contract (*dispositio*); it can also contain antecedents, aims, justifications (*exordium*, *narratio*), penalty clauses or clauses destined to ensure its execution (*sanctio*, *clausulae*), etc.

Diplomatics has invested many efforts in classifying these parts properly for two main reasons : on the one hand, the study of semantic relationships between sequences of objects and statements forming a model makes it possible to characterize a writing style according to writing traditions and typologies. On the other hand, as documents are the product of social and intellectual practices, changes in their wording can help to elucidate complex phenomena such as the circulation of ideas, the evolution of legal vocabulary and the configuration of social usages of writing objects. In this sense, the retrieval and proper classification of the internal structure of a charter are major steps in studying the information contained therein.

3.4 Parts of diplomatic discourse in the CDLM

The CDML considers 26 different section tags in the corpus. Not all tags are formally recognized as parts of discourse by diplomatics. This is however the case for five tags introduced to distinguish the signatories' respective roles : SMT (*signa manuum testium*), SMR (*signa manuum rogantium*), SMC (*signa manuum consentientum*), SME (*signa manuum estimatorum*) and SMF (*signa manuum fideiussorum*) corresponding to: persons who testify and command the exchange for the first two; persons who allow, estimate and guarantee the exchange for the last three. The annotation of these signatures in separate tags aims to translate the form of notary charters into a digital template. The signs of participants are stamped at different moments since a notary charter may follow a long process of drafting and validation.

Two other tags are related to notarial writing: the *Completio*, to indicate the authentication subscription done by the notary (the name comes from the formula "post traditam complevi et dedi" affixed at the end of a charter to confer it a recognized legal value) and the *Tenor-Additum* containing the possible *extra sigillum* notes to the written record – corrections or indications about the circumstances, which are not strictly speaking part of the translated juridical act, but related to its drafted form.

In a similar way, the CDLM editors introduced *Clausulae* and *Formulae* in order to distinguish clauses and formulae at the end of the *Dispositio*. *Formulae* are normally short, stereotyped sentences strongly connected with diplomatic forms and juridical language used to express the clauses of the exchange, while *Clausulae* outline the particular dispositions of an exchange, so they are common formulaic expressions (see examples in table 5). The line between these two categories is often very thin and their annotation is ambiguous in many texts. Furthermore, formulae and clauses are widely variable sub-parts that are not always distinguished in diplomatics. As they all are annotated under a single tag, the model might not be able to fit efficiently on these. We have not omitted them in our training set, but we only considered them in cases

where they act as final clauses locking and guaranteeing the business, to avoid an overlap with the *Dispositio* since they are normally considered as final sentences of this section.

Part of dipl. discourse	Freq	% of corpus	Avg. length	Median length	Charter section
DTCRON	4917	95.5	18.3	17	A
Dispositio	4596	98.8	366	295	B
DTTOP	4272	92.8	10.7	6	A
Completio	4085	89.2	19.7	14	C
Formulae	3429	75.0	65.9	63	B
Clausulae	3374	73.8	95.7	87	B
SMT	3117	58.9	18.4	16	C
SMR	2711	59.0	19.4	17	C
Subscriptio	2667	24.8	11.6	9	C
Invocatio	1462	32.0	5.0	4	A
Exordium	817	17.9	44.3	32	A
Rogatio	636	13.8	12.1	9	C
SMC	621	12.7	21.2	20	C
Narratio	372	8.0	268.7	170	A
SME	276	6.0	21.3	20	C
Intitulatio	235	5.1	7.9	7	A
Sanctio	200	4.3	39.5	30.5	B
Iussio	176	3.8	9.1	5	B
Inscriptio	133	2.9	21.3	18	A
SMF	129	2.8	14.0	12	C
Estimatio	115	2.5	113.2	112	C
Corroboratio	94	2.0	22.9	21	C
Promulgatio	88	1.9	20.7	14	C
Recognitio	40	0.9	9.4	9	C

Table 1: Frequency of parts of discourse annotation in the CDLM corpus. "% of the corpus" indicates percentages of charters containing the concerned section (some parts can be used more than once in the same charter. It is the case of the DTCRON and Subscriptio). "Charter section" indicates 3 major sections: initial (A), middle (B) and final (C). The "length" is expressed in number of words.

Among the remaining 20 tags, three groups emerge by frequency of use. The first group is in line with the most widespread charter model in Western Europe, and we find these tagged parts in at least one third of CDML corpus. It consist of the *Dispositio*, the heart of the act that is present in 99% of the corpus; the dates : *DTCRON* (time date) and *DTTOP* (place date), since at least 93% of the acts are dated and localized; the *Invocatio* (divine invocation), normally the first sentence in the text; and the main subscriptions : *SMT*, *SMR* and *Subscriptio* which correspond to the main participants in the act.

The second group includes less used parts belonging to a more formal model of charters. It includes the *Exordium* and *Narratio*, which introduce legal or religious justifications and antecedents, or circumstances of the action respectively, in the beginning of the charter *Text*. The *Sanctio*, *Iussio* and *Formulae* are very formal final clauses used to "lock" the written act and state that the required formalities were performed; *SMC* and *SME* are less usual signatories, and the *Intitulatio* and *Inscriptio* are common parts in charters from other European regions before the 13th century, because they introduce the identification of the author and recipient of the charter – they are however less used in notary models operating in Lombardy.

The last group consists of six scarcely used parts corresponding to final validation signs and clauses, as is the case for *Recognitio, Corroboratio, Estimatio, Rogatio, SMF* and *Promulgatio*, which is conversely very common in other charters collections. All these clauses state that

the above-describe juridical action has followed all the accreditation steps and that the charter has a legal value, but they are not always available in the document, and they rarely appear all together in the same document.

Finally, in table 1 we present a classification of a, b and c, corresponding to initial, central and final sections of charters. Less-used parts are mostly integrated in section c, where notary charters display different validation signs; most large parts belong to the central section (b), containing descriptions of lands and goods as well as terms and conditions of exchanges. The shorter and more formulaic parts are, in general, found in the *Protocol* (a).

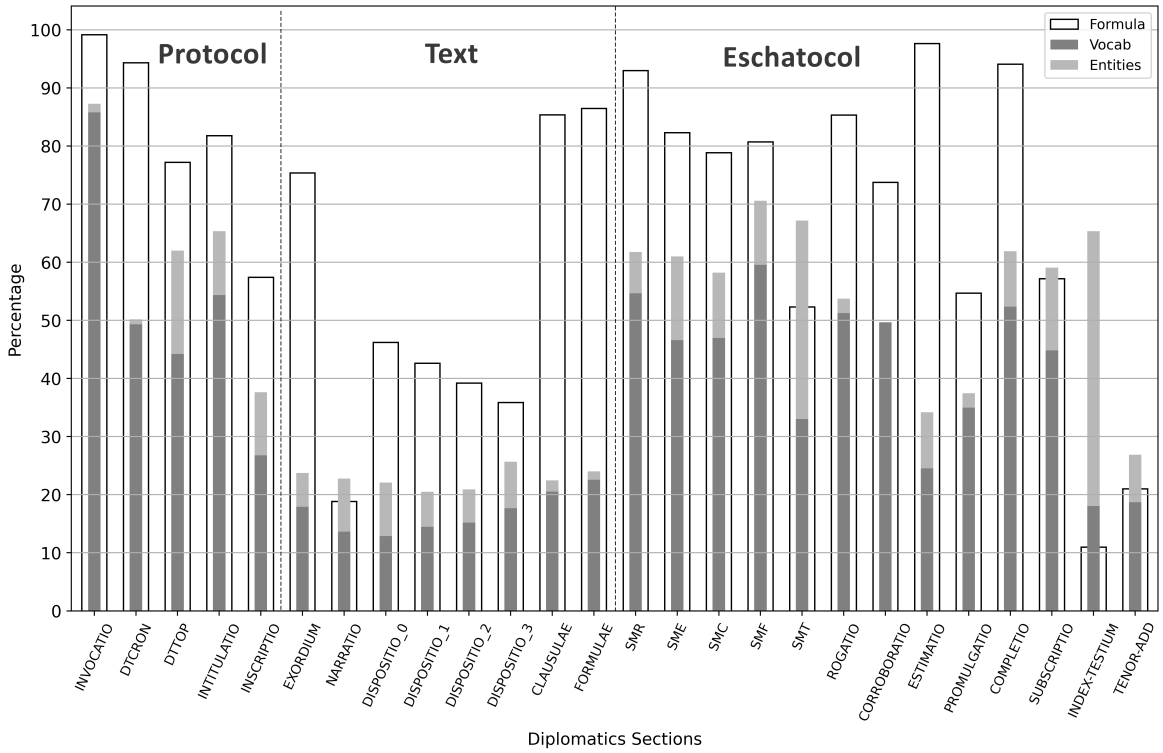


Figure 2: Statistics on formulaic usages in diplomatic sections. White bars (Formula) represent the percentage of formulaic content; gray (Vocab), the percentage of text covered by the 15 most commons words and light-gray (Entities), the percentage of named entities. For example, in the case of *Intitulatio*, 82% of the usages are formulae, 54% of all content is expressed using a 15-term vocabulary, and 16% are named entities – which seems appropriate for a section presenting a person preceded by relevant titles and devoted formulae.

3.5 Clustering of sections by formula

Since the formula is a key piece in diplomatic sections, its detection, and the calculation of formulaic content ratio within each section are useful information to further evaluate and understand modeling issues, as well as to acquire a global statistical vision of each section from massive datasets. Clustering sections according to a pairwise similarity score is an efficient strategy to get formulaic statistics. Literature shows that using cosinus or sentence embeddings to define clusters centroids is a good approach, but rather we chose to use the Dice Coefficient designed to quantify shared information between two data sequences. The advantage of the Dice score is that it easily measures how similar two sequences (X, Y) are in spite of terms transposition. Even if *Formulae* are stereotyped sequences by definition, they normally present a high number of variations in character and word-level features : spelling variations; order and lexical transpositions; presence of synonyms, periphrases and named entities; grammatical ac-

cidents on flexion, tense, number, etc. So, Dice must be applied on lemma sequences instead of words to avoid the grammatical accidents and main character variations. Named entities must also be removed eventually as they are not part of either the formula or the language dictionary.

$$Dice(X, Y) = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \quad (1)$$

Many tests indicate that a coefficient of 0.5 or more is the minimal argument to form clusters. A simple snippet of code transforms our raw sections into grouped sections according to their shared formulae. For example, in the *Intitulatio*, a very formulaic section, three sets can be identified, grouping these 5 examples:

- a_1 = ['Otto', 'Dei', 'favente', 'pietate', 'imperator', 'augustus']
- a_2 = ['Henricus', 'divina', 'favente', 'clementia', 'Romanorum', 'imperator', 'augustus']
- b_1 = ['Nicholaus', 'episcopus', 'servus', 'servorum', 'Dei']
- b_2 = ['Celestinus', 'episcopus', 'servus', 'servorum', 'Dei']
- c_1= ['Adalgerius', 'cancellarius', 'et', 'missus', 'gloriosissimi', 'et', 'piissimi', 'regis', 'Henrici']

The first (a) and second (b) sets including formulaic *Intitulationes* were widely used by kings and popes respectively and the third (c), a less formulaic but lexically restricted used in this case by Adalgerius, chancellor and emissary of Emperor Henry III (1016-1056). Thus, the *sui generis* formula, the scarce used formulations, but above all the sections with no formulae will form sets with few or only one member. As shown in figure 2, after grouping sentence sections, the proportion of text grouped in sets of three or more members is higher than 70% in most cases, indicating a high density of formulaic uses. But as indicated by gray bars (frequency of terms and named entities), in many cases the formulaic nature has more to do with the use of a restricted vocabulary than with the use of a fixed and invariable sequence. We must take into account that many sections are mono-formulaic, like the *Invocatio* or *Subscriptio*, but that in others such as the *Dispositio*, the formula is a sub-sequence in the texts with boundaries that may be more or less ambiguous. Three sections are paradigmatic : the *Invocatio*, *Clausulae* and *Index-Testium*. Almost all the *Invocationes* are formulaic (99.6%) and 86% of all their content can be expressed using 15 terms or less, which shows the highly stereotyped nature of this section. Conversely, even if 87% of content in *Clausulae* belongs to a formulaic forms, vocabulary seems to be much broader because *Clausulae* are here used to label a large variety of final exchange clauses. Finally, the *Index-Testium* emerges as the less formulaic section both in formula ratio and in vocabulary. In fact, this section can include short formulae (see Table 5), but since its function is to present the list of witnesses, 50% of all content is recognized as named entities, thus hindering formula detection.

IV TRAINING THE BI-LSTM MODEL

4.1 Data preparation

This gold-standard corpus consists of 4 570 documents ($\sim$ 2,5M of tokens) and was split into three sections with a 0.8 to 0.2 ratio: a training set (3,664 documents), a validation set (184 documents) and a reserved test set (722) with documents that were not part of the training. We consider each charter as one training unit with a max length of 1,200 words (and a median of 243) and a max word length of 12 characters (and a median of 5).

4.2 Corpus pre-processing

The extraction of lexical features can be a challenging task in medieval Latin, which is a subversion of Latin for which few automatic language processing tools exist. Tokenization is a two-step task. First, diphthongs (*ae, oe, vv, ee*, etc.) and enclitic suffixes must be converted, as they are extensively used in Latin (*ne, ve, que*, for ex: *populusque, nihilne*), and flagged in research as problematic automatic issues; then, a stemming algorithm must be implemented to provide tokens and roots of words.

Parts-of-speech tagging (PoS) is provided by the Omnia projet lemmatizer – which is based on a dictionary of 75,000 hand-validated lemmata (Bon [2011]). This is a robust tool which aims to tackle problems linked to false lemmas (non-existent words) and intense spelling variability in medieval Latin. The tool uses a TreeTagger approach (Schmid [2013]) to generate PoS annotation.

Other features coming from named entity models and chunkers have become available in past years for medieval Latin, but as is shown in the evaluation section, their integration into the model provides a small improvement in results, making the training much more complex. (see 4.10)

Besides, some diplomatic parts such as the *Dispositio* or *Narratio* can be very large (150 to 350 words) and display a combination of formulae and freer redaction, while most parts of discourse have between 5 and 20 words and only use one or two stereotyped formulae. This problem of category imbalance can generate an important bias to the model that can label some parts as *Dipositio* or *Narratio* to minimize its error ratio. To control that, we have artificially divided these sections into several sentences (max 5) as follows: *Dispositio-0*, first sentence; *Dispositio-1*, second sentence, etc.

4.3 Problem definition

We see our problem as a traditional two-step sequence labeling task. The input is a defined sequence of tokens $x = (x_1, x_2...x_{n-1}, x_n)$ and the output must be defined as a sequence of tokens labels $y = (y_1, y_2...y_{n-1}, y_n)$. We use the conventional BIO format to represent the category labels. Each label was assigned to a BIO class as follows: B-tag for Begin (B), I-tag for continuation (I) and O-tag for absence (O) of label, respectively. The first step involves the use of NLP tools to extract and transform character-level and word-level features; the second step is the classification of sequences according to the 26 categories of the corpus, among which we used : Invocatio, Narratio, Exordium, Dispositio, Inscriptio, Subscriptio, Intitulatio, SMC, SMF, SMR, SMT, Corroboratio, DTTOP, DTCRON, Completio, Promulgatio, Rogatio, Estimatio, Clausulae, Formulae, Index-Testium and Tenor-additum.

4.4 Model Architecture

Figure 3 shows the overall architecture of our proposed model. We trained three embedding vectors from our data: word representation, character-level word representation and PoS character-level embedding. Alternatively, for our best model, we used a word embedding model pre-trained on a collection of diplomatics medieval corpora (10.5M of tokens). Then, we applied the Keras TimeDistributed wrapper-layer to the character and POS-char embeddings in order to apply the same features extraction to each frame at each time step. Finally, we merged embeddings before feeding them into a Bidirectional LSTM layer, thus producing a hidden state for each word. The final CRF layer considers each LSTM output as a weighted matrix of feature vectors of each word, and predicts the final tag sequence by using a statistical approach.

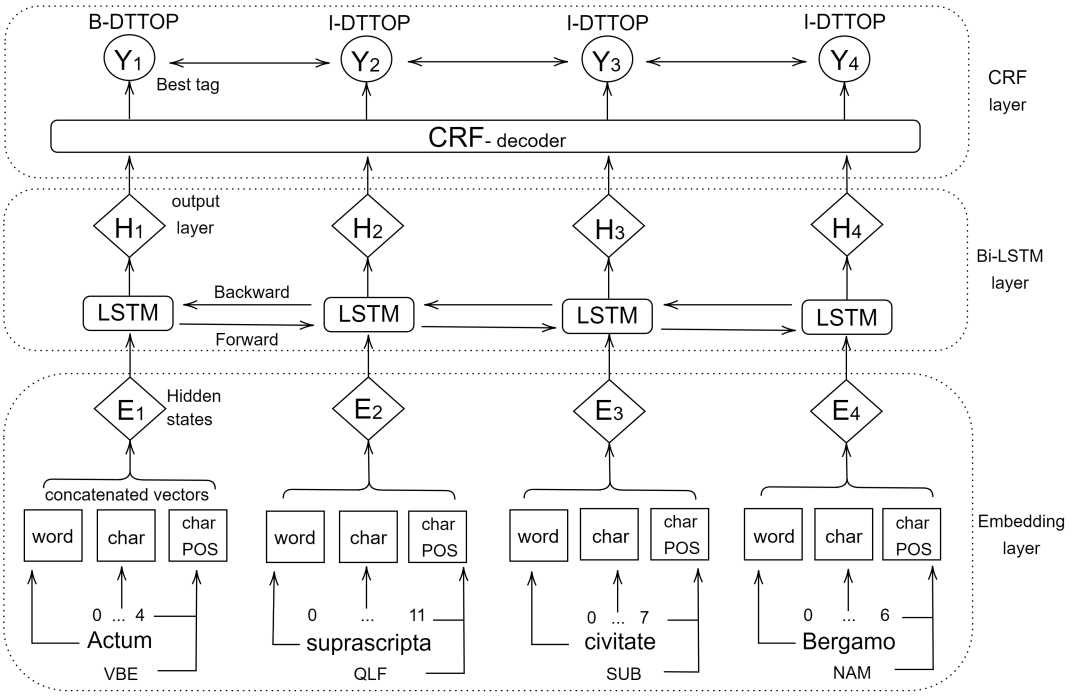


Figure 3: BiLSTM-CRF architecture using words, characters and PoS embeddings as input on the excerpt "Actum suprascripta civitate Bergamo" (written in the aforementioned city of Bergamo) tagged in the hypothesis as a DTTOP (Data Topica) category.

4.5 Word-representations

In our model, the word and character-embedding, previously transformed into one-hot encoded vectors, are concatenated before being decoded. Feeding the model with character and word information is crucial because in inflected languages such as Latin the grammatical relationships between words in the sentence are expressed using declension and suffix. Besides, spelling mistakes in medieval Latin are an important issue since scribes were not always following grammatical rules – false lemmas, *hapax*, unique word variants as well as abbreviations or erased text are very common in this kind of documents. Word vectors depending of a limited dictionary and a predefined grammar pattern are not enough. A character-level approach can alleviate this situation allowing to encode all types of textual phenomena using a simple character dictionary (103 keys in ours).

4.6 POS information

Using syntactical information as PoS and Lemma in neural networks remains an open challenge (Zhou et al. [2020]). The PoS features can greatly help in tasks working with large dependency plots as they import contextual features into each token, but PoS features are related to words; using them in a character-level approach requires distributing PoS tags among characters for each word. In this case – inspired by the work of Li et al. [2018] – we generate a new feature combining character position and PoS tags. Positions are distributed using a 4-set tags as follows: B:Beggining, M:Middle, E:end, S:single (Table 2 shows an example for a single sentence). This enriched character-level feature is later put into an embedding model. With this approach, we expect to add some auxiliary lexical information to characters as character embeddings do not capture any of these textual aspects.

4.7 Bi-LSTM Layer

Bidirectional long short-term memory (BiLSTM) models have been proven to be effective for multiple sequence labelling tasks and long dependency problems. In classical RNN networks the vanishing gradients have quickly become a major shortcoming as they do not allow to learn long dependencies. The LSTM tackles this issue using three control gates on each memory activation cell to maintain the persistence of the information by keeping the relevant content of the sentence and ignoring the irrelevant ones. The idea of this bidirectional variant is to reinforce this persistence learning with a two-way sequence analysis: one in natural reading order and the other on the opposite way, thus connecting present and past context of each token in the sentence. In that sense, the output of a BiLSTM layer is a vector formed by the concatenation of a double sequence of LSTM hidden states for each token and token features embeddings $y_t = \vec{h_t} + \overleftarrow{h_t}$. This output is finally decoded by a CRF-layer (Huang et al. [2015]).

4.8 CRF layer

The BiLSTM output assumes that each time step is independent when many tags are in fact interdependent. A way of overcoming this issue is to incorporate the output vectors as observations in a Conditional Random Fields (CRF) layer which can predict the entire label sequence in each time step. CRF is a widely validated method for classifying mutually dependent sequences because it takes contextual and multidimensional data observations into account and estimates transition probabilities between tags to predict their output (Lafferty et al. [2001]). The order of the parts of discourse is well-determinate on charters and the model must learn that for example a *Dispositio* is frequently displayed after a *Promulgatio* or a *Narratio* is never followed by a *Iussio*, but it must also learn where a category ends and another one starts, since good detection of category boundaries is crucial to determinate the topicality of a section. As a discriminate model, CRF randomly generates all possible label sequences $y = (y_1, y_2...y_{n-1}, y_n)$ given a sequence of observations $h = (h_1, h_2...h_{n-1}, h_n)$ and chooses the best combination by measuring the conditional probability of each tag in each position. Formally, the score of each sequence can be written as:

$$s(h, y) = \sum_{t=1}^{n+1}(A_{yt-1,yt} + P_{t,yt}) \tag{2}$$

where $P_{t,yt}$ is the probability of an x_t word tagged as y_t and $A_{yt-1,yt}$ the probability to see a y_t tag preceded by a y_{t-1} tag.

4.9 Training parameters

The LSTM decoder accepts the data-features vectors under the form of a multiple-class matrix and initializes random weight matrices. We do not retrain word embeddings along with the Bi-LSTM-CRF. The grid of hyper-parameters was tested on four key options: batch-size $\in$ {2, 4, 16, 32}, output embeddings dimensions $\in$ {100, 200, 400}, learning methods $\in$ {sgd, adam rmsprop} and activation functions $\in$ {relu, tanh, sigmoid}. An optimal combination was chosen with a batch size of 4, embeddings dimensions of 200, an rmsprop optimizer and a relu activation of the cell state. Furthermore, the weights convergence (using loss-validation), on adaptive and non-adaptive optimizers, occurs on a 20x epochs threshold.

All tests were performed using a 10-cores processor with a Gtx2080-ti (11GB) GPU for about 12 - 22 hours of training depending on training set size and batch sizes.

Token	Lemma	PoS	Char	POS-char	BIO-Tags
Datum	do	VBE	D	B-VBE	B-DTTOP
			a	M-VBE	
			t	M-VBE	
			u	M-VBE	
			m	E-VBE	
Laterani	Latero	SUB	L	B-SUB	I-DTTOP
			a	M-SUB	
			t	M-SUB	
			e	M-SUB	
			r	M-SUB	
			a	M-SUB	
			n	M-SUB	
			i	E-SUB	
,	,	PON	,	S-PON	I-DTTOP
III	3	NUM	3	S-NUM	B-DTCRON
idus	idus	SUB	i	B-SUB	I-DTCRON
			d	M-SUB	
			u	M-SUB	
			s	E-SUB	
iunii	iunius	QLF	i	B-QLF	I-DTCRON
			u	M-QLF	
			n	M-QLF	
			i	M-QLF	
			i	E-QLF	

Table 2: Training excerpt for the sequence : "*Datum Laterani, III idus iunii* [pontificatus nostri anno tertio decimo.]". The place (DTTOP) and date (DTCRON) when the charter was written are indicated: Lateran, three days before the Ides of June in the thirteenth year of the pontificate of Pope Innocent III (11-06-1210).

4.10 Pre-trained embeddings and NER

We have trained several models, two of them using pretrained word-embedding and named entities. The embedding models for medieval Latin are not available and Latin embeddings published in the past years mostly come from classical literature corpora, which do not fit our domain, period or language state very well. In order to use pre-trained embeddings, we have trained a customized 200-dimensions Word2vec (Mikolov et al. [2013]) model using a limited collection of medieval Latin charters (10.5M of tokens). This collection is not a formal corpus, but an *ad hoc* resource formed mostly of freely available digital editions of charters.[6]

On the other hand, automatic named entities hypothesis were generated using a CRF-model (Aguilar et al. [2016]) adapted to medieval texts and trained on the CBMA charters collection (10th-13th centuries). Personal names and place names are recognized in a range between 0.80-0.92 of precision according to the published evaluation for this model on four medieval European corpora.

V EVALUATION OF THE MODELS

Table 3 shows the best results obtained with a training set of 3, 664 charters. We presented the usual Precision, Recall and F1 measures as the evaluation metrics.

We designed four character-based models as baseline methods to make performance comparisons. The architecture and hyper-parameters were defined in section 4.9; the word and subword features for each model are as follows :

[6]This collection includes the aforementioned corpus: CBMA (2.2M tokens), DiBe (4.8M tokens), CDLM (2.5M tokens) and the HOME-Alcar corpus (1M tokens), amounting to a total of 32k Latin acts.

- **W+Ch** : character-based BiLSTM with the concatenation of word and character embedding as inputs.
- **W_Emb+Ch** : character-based BiLSTM with the concatenation of pre-trained word- and character-embeddings as inputs.
- **W_Emb+Ch+PoS** : character-based BiLSTM with the concatenation of pre-trained word-, character-embeddings, and PoS-character embedding as inputs.
- **W_Emb+Ch+PoS+NER** : the previous model plus embeddings of automatic named entity hypothesis of places and persons.

From these experiments we can propose 6 primary conclusions:

1. Our approach performs well in charter text segmentation – it displays great performance (over 0.85 in F1) in the recognition of most sections (17 of 23) from the four models. The difference on average performance between the first model (**W+Ch**) and the best model (**W_Emb+Ch+PoS**) is about 4 points but the latter treats sections with scarce representation (ratio from 0.71 to 0.85) and sections with freer redaction in the *Text* macro-section (0.72 to 0.91 ratio) much better (5 to 12-point difference).
2. Adding extra features as NER, POS and pre-trained embeddings helps to reinforce learning in problematic areas and can become a determining factor in the final performance. But as shown by the **W+Ch** model, optimal models can be trained for our task even if these features are not or only partially available – as is usually the case for historical languages.
3. Using pre-trained embeddings significantly increases the efficiency of the model: they provide weighted features trained on a large dataset and extra semantic information for our originally imbalanced dataset, and thus help learn scarcely represented categories and boost model generalization.
4. The impact of PoS and NER information is less remarkable. They can make for a slightly more efficient model (2 to 3 points more on average), specially PoS in the most formulaic sections, while NER helps in modeling sections with a high density of proper names.
5. In most sections, differences between performance in B(egin) and I(inside) tags are not relevant (1-3 points) confirming full sequence recognition; but on large sections B(egin) tags are usually less successfully recognized (4-8 points less), suggesting some issues in the recognition of his class transition sequences.
6. The imbalanced nature of the corpus does not seem to be an insurmountable issue. Splitting large sections by sentences seems to help control the bias as they only affect adjacent sections, which suggest deeper problems (see discussion).

Results show high performance (a ratio of over 0.85) in F1-measure mostly on the parts of discourse with a high number of examples in the training corpus and in the most formulaic sections. This corresponds to the most used parts in charters in almost all Western Europe traditions, such as the *Invocatio*, *Dispositio*, time date (*DTCRON*), date of place (*DTTOP*), *Subscriptio*, *Intitulatio*, *Inscriptio*,*Completio* and *Index-Testium* (*IT*), as well as the identification of the main subscribers (*SMC, SMR, SMT, SME*). Most of these parts are found in the *Protocol* sections of charters and they are comprised of one or two formulae, completed by named entities, using a stereotypical vocabulary and an order that is more or less defined.

The case of sections belonging to the central part of charters such as the *Dispositio*, *Exordium* and *Narratio* is special because they present a freer redaction. They are normally displayed sequentially and represent the three largest parts of the corpus. They have an overall good ratio of recognition (from 0.70 to 0.93), which represents a major step forward – it means we are able

to classify the central part of the charter and to separate, when the three are used simultaneously, the description of the juridical action from previous information such as the reasons or moral justification of the exchange. Moreover, the model is also able to provide acceptable recognition (0.75 to 0.84 overall ratio) on the *Formulae* and *Clausulae*, partially tagged on our dataset, which mostly corresponds to the final clauses used to lock the exchange and normally found at the end of the central part. Both are key sections because they complete the classification of the *Text* macro-section, which is technically the most complex section to learn.

MODEL / LABEL		W + Ch			W-Emb + Ch			W-Emb + Ch + PoS			W-Emb+Ch+PoS+Ner			Support
		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	
CLAUSULAE	B	0.77	0.67	0.72	0.78	0.71	0.74	0.88	0.68	0.77	0.89	0.74	**0.81**	126
	I	0.87	0.75	0.81	0.80	0.82	0.81	0.90	0.73	0.81	0.90	0.78	**0.84**	12 110
COMPLETIO	B	0.98	0.96	0.97	0.98	0.97	**0.97**	0.97	0.97	0.97	0.97	0.96	0.96	801
	I	0.99	0.97	0.98	0.98	0.99	0.98	0.99	0.98	**0.99**	0.99	0.97	0.98	10 322
CORROBORATIO	B	0.99	0.92	**0.96**	0.96	0.92	0.94	0.96	0.93	0.95	0.96	0.92	0.94	24
	I	0.99	0.96	0.98	0.99	0.97	**0.98**	0.97	0.98	0.97	0.97	0.98	0.97	455
DISPOSITIO	B	0.87	0.88	0.87	0.92	0.91	0.92	0.92	0.91	0.91	0.91	0.92	**0.92**	891
	I	0.98	0.94	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	**0.98**	308 915
DTCRON	B	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.98	**0.99**	0.99	0.99	0.99	947
	I	0.99	0.99	0.99	0.99	1	0.99	0.99	1	**0.99**	0.99	1	0.99	14 416
DTTOP	B	0.94	0.97	0.96	0.99	0.97	0.98	0.99	0.98	**0.99**	0.99	0.97	0.98	848
	I	0.97	0.98	0.97	0.98	0.97	**0.97**	0.97	0.96	0.97	0.98	0.97	0.97	4 197
ESTIMATIO	B	0.79	0.96	0.87	0.96	0.92	0.94	0.89	1	0.94	0.96	0.96	**0.96**	24
	I	0.89	0.98	0.93	0.95	0.97	0.96	0.92	1	**0.96**	0.94	0.98	0.96	2 753
EXORDIUM	B	0.95	0.94	0.94	0.93	0.96	0.94	0.93	0.97	0.95	0.94	0.96	**0.96**	174
	I	0.97	0.92	0.95	0.93	0.99	0.95	0.93	0.99	0.96	0.95	0.98	**0.97**	7 755
FORMULAE	B	0.56	0.60	0.58	0.84	0.54	0.68	0.90	0.55	0.68	0.82	0.64	**0.72**	159
	I	0.68	0.71	0.69	0.88	0.65	0.75	0.94	0.61	0.74	0.87	0.74	**0.80**	10 017
INSCRIPTIO	B	0.96	0.82	0.88	0.94	0.74	0.84	0.96	0.83	**0.89**	0.92	0.76	0.83	29
	I	0.95	0.80	0.87	0.82	0.91	0.86	0.86	0.84	0.85	0.99	0.82	**0.90**	381
INTITULATIO	B	1	0.95	0.97	0.95	0.94	0.94	1	0.95	**0.97**	0.96	0.95	0.95	56
	I	0.95	0.92	**0.93**	0.96	0.80	0.88	0.98	0.85	0.91	0.98	0.84	0.91	380
INVOCATIO	B	0.99	0.99	0.99	1	1	**1**	1	0.99	0.99	1	0.99	1	299
	I	0.99	0.99	0.99	0.99	1	0.99	1	1	**1**	0.99	0.99	0.99	1 099
Index Testium	B	0.93	0.95	0.94	0.93	0.97	0.95	0.94	0.96	**0.95**	0.93	0.96	0.94	243
	I	0.94	0.95	0.94	0.93	0.99	0.96	0.96	0.96	0.96	0.95	0.97	**0.97**	5 035
NARRATIO	B	0.47	0.72	0.57	0.53	0.58	0.56	0.61	0.61	0.61	0.64	0.64	**0.64**	74
	I	0.51	0.91	0.65	0.73	0.72	0.72	0.65	0.83	0.73	0.82	0.83	**0.83**	15 456
PROMULGATIO	B	0.78	0.67	0.72	0.73	0.79	0.76	0.70	0.90	**0.79**	0.75	0.60	0.67	21
	I	0.62	0.71	0.66	0.70	0.64	0.72	0.68	0.97	**0.80**	0.77	0.58	0.66	303
ROGATIO	B	0.76	0.81	0.79	0.84	0.76	**0.80**	0.82	0.77	0.80	0.85	0.62	0.72	122
	I	0.80	0.81	0.80	0.84	0.79	**0.82**	0.84	0.78	0.81	0.86	0.68	0.76	1 296
SMC	B	0.77	0.94	0.85	0.86	0.93	0.89	0.84	0.94	0.89	0.97	0.89	**0.93**	120
	I	0.75	0.95	0.84	0.91	0.93	0.92	0.84	0.95	0.89	0.98	0.92	**0.95**	2 453
SME	B	0.92	0.86	0.89	0.96	0.91	**0.93**	0.85	0.94	0.89	0.96	0.85	0.90	57
	I	0.93	0.85	0.89	0.91	0.95	0.93	0.84	0.96	0.90	0.96	0.92	**0.94**	1 140
SMF	B	0.99	0.25	0.40	0.63	0.44	0.52	0.75	0.67	**0.71**	0.60	0.67	0.63	28
	I	0.92	0.29	0.44	0.73	0.64	0.68	0.72	0.74	**0.73**	0.53	0.77	0.63	339
SMR	B	0.98	0.95	0.97	0.98	0.98	**0.98**	0.97	0.98	0.98	0.97	0.98	0.98	546
	I	0.98	0.93	0.95	0.99	0.98	**0.99**	0.96	0.98	0.97	0.97	0.98	0.98	9 877
SMT	B	0.96	0.98	0.97	0.96	0.98	0.97	0.97	0.97	**0.97**	0.97	0.96	0.97	620
	I	0.98	0.98	0.98	0.99	0.99	**0.99**	0.99	0.98	0.99	0.99	0.99	0.99	10 494
SUBSCRIPTIO	B	0.92	0.94	0.93	0.93	0.93	0.93	0.90	0.93	0.91	0.90	0.97	**0.93**	559
	I	0.94	0.98	**0.96**	0.93	0.96	0.95	0.93	0.97	0.95	0.91	0.98	0.94	4 864
TENOR-ADDITUM	B	0.89	0.90	0.89	0.93	0.88	0.90	0.93	0.88	0.90	0.92	0.90	**0.91**	141
	I	0.88	0.98	0.93	0.89	0.97	0.93	0.92	0.96	**0.95**	0.90	0.97	0.93	10 847
Macro-Average	B	0.87	0.84	0.85	0.89	0.86	0.87	0.90	0.88	**0.89**	0.90	0.86	0.88	6909
	I	0.89	0.88	0.88	0.90	0.90	0.90	0.90	0.91	0.91	0.92	0.90	**0.91**	434 904

Table 3: Evaluation results on CDLM test set using four models. W (words), Ch (characters), PoS (Parts-of-speech tags), Emb (pre-trained word-embeddings), Ner (named entities recognition), Support (number of observations), Pr (Precision), Rc (Recall), F1 (F1-measure)

On the other hand, the sections showing a lower ratio of recognition (0.69 to 0.81) such as the *Rogatio*, *Promulgatio* and *SMF* are those with a small number of examples on the training set. As mentioned earlier, they are short sentences of validation used in chancellery acts and public notary charters, underused in notarial Lombard traditions. In fact, this is the case for two other sections : *Corroboratio* and *Estimatio*, but these last two are much more efficiently recognized, mainly because their formulation is stable throughout the corpus (see figure 2).

Finally, a good classification on most of parts on the finer-level charter hierarchy confirms that we are also able to provide a good second classification on the broader hierarchy – that is, a model that can discriminate the main three sections on charters: *Protocol*, *Text* and *Eschatocol*. This in itself is a key step for an automatic classification of charters.

	Partial Match			Exact Match		
	Precision	Recall	F1	Precision	Recall	F1
COMPLETIO	0.969	0.780	0.864	0.843	0.818	0.830
DISPOSITIO	0.898	0.975	0.935	0.831	0.896	0.862
DTCRON	0.850	0.944	0.894	0.789	0.833	0.810
DTTOP	0.920	0.958	0.938	0.791	0.863	0.826
EXORDIUM	0.785	0.846	0.814	0.533	0.727	0.615
INSCRIPTIO	0.785	0.611	0.687	0.583	0.411	0.482
INTITULATIO	0.931	0.836	0.881	0.656	0.600	0.626
INVOCATIO	0.878	0.966	0.920	0.806	0.892	0.847
NARRATIO	0.692	0.750	0.720	0.416	0.416	0.416
PROMULGATIO	0.868	0.942	0.904	0.794	0.843	0.818
ROGATIO	0.666	0.800	0.727	0.571	0.666	0.615
SMC	0.885	0.861	0.873	0.781	0.735	0.757
SMF	0.833	0.625	0.714	0.666	0.500	0.571
SMT	0.897	0.897	0.897	0.801	0.777	0.789
SUBSCRIPTIO	0.927	0.888	0.907	0.743	0.679	0.709

Table 4: Results of model evaluation on Montecassino 200-items set using Precision, Recall and F1 metrics.

5.1 Evaluation on Montecassino charters

As we can see in table 4, the results obtained on the test set from the CDML are in large part replicated when the model is applied to the Montecassino cartulary, but using a more restricted set of tags as we further explain.

From the results, we can see that the results on the *Invocatio*, *Dispositio*, dates (*DTTOP, DTCRON*), the main signatories (*SMC, SMT*) and the *Completio* are highly successful in Partial and Exact with over 85%, reaching 90% in the case of *Dispositio* and *Subscriptio*. Thus, in these heavily used parts, the performance is only a few points lower than with the CDLM test.

We can also see similar performance in the detection of less-used parts as *Narratio*, *Exordium* and *Subscriptio* which show relatively low results in F1-measure (between 0.66 - 0.80). Conversely, the *Promulgatio*, *Intitulatio* and *Inscriptio*, which are underused formulaic parts in the Lombard corpus, are successfully recognized in the Montecassino charters, where they are much commonly found in the *Protocol*.

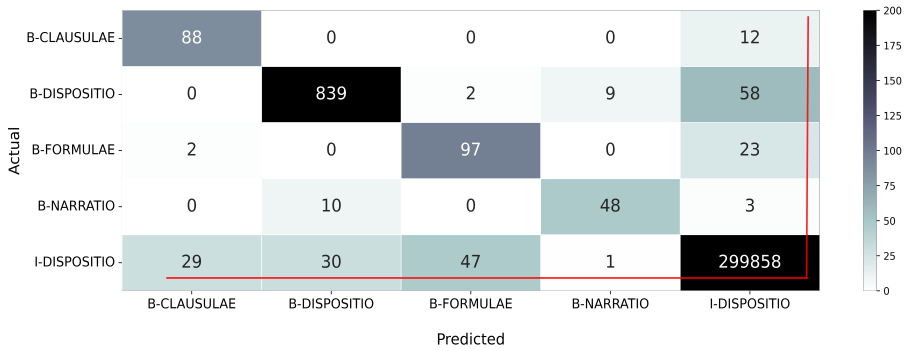
Finally, other parts belonging to notarial models are hard to find in the Montecassino corpus whereas the *Corroboratio* and *Rogatio* are well-identified, but only detected in public charters. In fact, the *Recognitio* and *Estimatio*, which are barely used in the CDLM, were not found in our Montecassino evaluation set.

VI DISCUSSION

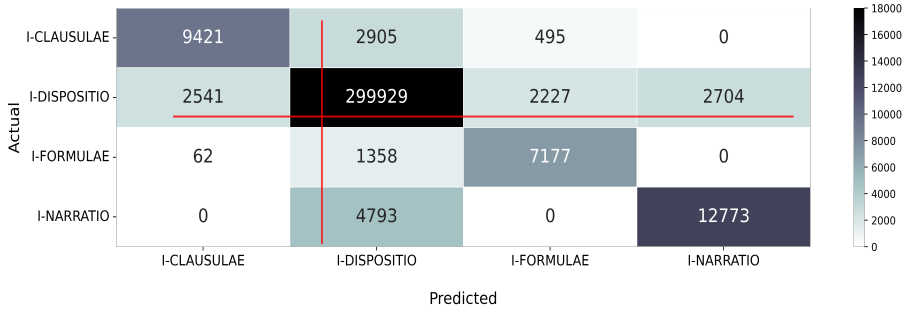
This work shows that a robust tool to classify sections on charters can be modeled using a neural approach. Three issues must be highlighted as they can provide an explanation for the optimal performance of the models on the one hand and for the main recognition shortcomings on the other :

(i) Due to the nature of its discourse, our training corpus is imbalanced both in the number of its sections and in their size, which led several categories to be over-represented in the training. Four major sections account for 74% of the corpus : *Dispositio*, *Narratio*, *Exordium* and *Formulae*; and seven others (*Promulgatio*, *Inscriptio*, *SME*, *SMF*, *Estimatio*, *Corroboratio*, *Rogatio*) are found in 5% or less of the corpus documents. Results show that in general, the most used sections are better modeled than less usual or specific sections. However, results also show that some very formulaic sections (for ex. *Estimatio*, *Intitulatio*) even if short and scarcely represented can obtain a high ratio of recognition because they are mostly composed of typical word sequences, common in normative texts, that are an easy fit for the model. This is typically one of the reasons why our model performs well on the data: the boundary and topicality detection of the sections is facilitated by their formulaic nature. In our documents, as figure 2 showed, short sections generally rely on formulae, and large sections (even if formulae are involved) present a much freer style, making them a harder fit because detection of theirs class transitions is more complicated (see ii). As observed on the confusion matrix (Figure 4), most model errors are false positives on the *I-Dispositio* tag because the mass of data from this main category makes the discrimination boundary overlook smaller categories; this is typically a problem of imbalanced classification. Splitting the *Dispositio* has proven to be an efficient and easy strategy to control this bias, helping the model to successfully predict labels for all the minority categories. Nevertheless, if we take a closer look, the distortion introduced by the *Dispositio* remains strong on the adjacent sections (*Narratio*, *Clausulae*, and *Formulae*) probably because these sections, being a part of to the same macro-section (the *Text*), share a discursive tenor which is progressively moving away from formulae. This is the main reason why these widespread sections do not exceed the 0.85 threshold on all tests. The model shows weaknesses where the trained eye does too, as it is sometimes difficult to determinate the exact transition sequence between these sections. That being said, we should not overlook the existence of several manual misannotations, especially in the *Formulae* and *Clausulae*, in which differences can sometimes be neglected by a human annotator.

(ii) Secondly, as suggested in (i), it is quite clear that the different levels of lexical contingencies coming from a formulaic or freer wording directly affect the predictive performance of the model. The dates and invocation, which are very formulaic parts, present a performance close to 1; other parts like the *Intitulatio*, *Inscriptio* and *Subscriptio* are also formulaic but introduce specific information and therefore display a drop in performance by 5 to 10 points in F1 measure. That performance is of about 0.75 on large parts as the *Narratio*, *Formulae* and *Clausulae*, introduced by formulae but continuing with a freer redaction. Formula recognition is highly precise even on small data categories as *Corroboratio* and *Estimatio*. Thus, the progressive abandon of rigid formulae, i.e. of conventional lexical sequences, is a lower-performance factor as the model must face a higher level of syntactic and lexical variability. This can also be observed on the difference between recall and precision: it is minimal on formulaic sections and more pronounced on freer one as the model starts to lose recall – which means its generalization performance decreases.



(a) B(begin) labels



(b) I(inside) labels

Figure 4: Confusion matrix for B and I labels of *Text* sections

The formulaic discourse suggests the use of well-defined and topically centered sequences, but formulae normally have more to do more with the use of a restricted vocabulary than with variable combinations. A closer inspection of some examples (see table 5) would be more eloquent here. We found almost invariable formulae e.g. in the *Invocatio*, *Inscriptio* and some *Clausulae*; slightly variable ones in *DTTOP* and *DTCRON* that use limited vocabulary to indicate years, days and places, in combination with time and places entities; and highly variable ones as in the *Eschatocol* sections, where we can find the use of conventional and precise vocabulary combined to named entities. Indeed, the legal action transferred in the document must be evaluated, approved and confirmed by witnesses, officers and notaries, and each one of these actions will appear in a dedicated section of the document involving the names of these actors and a list of conventional and coordinated expressions. Formulae are not missing in freer redaction sections such as the *Exordium*, *Narratio* or even in the *Dispositio*; in fact, these sections can start with formula or religious quotes, but as they quickly present particular details of the legal action, their lexical universe is much broader (the 15 most used terms cover less than 15% of the content, see figure 2) making their modeling more complicated, especially in terms of section boundary detection.

Indeed, training on texts with an important level of formulaic or stereotyped content can lead to a very precise model on the test set, but this should not be taken as a guarantee of a good generalization performance. Legal formalism and the collections of rhetorical models had circulate among institutions since the High Middle Ages, but the style traditions in each region, order, chancelleries or even abbeys present notable differences, increasing the tension between individual expression and normative discourse. Our evaluation of the Montecassino charters proves that the model can be generalized to a corpus of external documents as it fits many common lexical series. However, these particular charters are geographically and chronologically close

to the Lombardian charters, which could be evidence of partial evaluation. Future experiments on French and Spanish charters, that are part of more distant traditions, should bring to light a larger scope for the massive application of the model on other collections.

(iii) Thirdly, as we have already argued, adding automatic hypothesis vectors using PoS, named entities and pre-trained embeddings makes for a 4-point gain on the macro-average performance, and for a 10 to 15-point increase in the performance on the complex or small data categories. These transfer-learning operations are indeed essential, but the NLP tools to extract them are still rare. In fact, our work aims to partially fill this major gap currently existing for this kind of sources, that are essential in medieval studies. In the case of PoS and Lemma for medieval Latin, we only know one lemmatizer, manually implemented and rich in features, but that has never been updated and is inaccurate on many unknown or miss-formed lemma. Recent Latin lemmatizers show a good performance, but as they rely on classical Latin literature they cannot be fully linguistically relevant to medieval charters. The named entity models are more recent, but NER is still an open challenge regarding the wide variety of diplomatic traditions. The model has been evaluated on external corpora showing an acceptable performance in an overall 0.82 to 0.93 ratio. As for embeddings, to our knowledge no public large embeddings yet exist for medieval Latin, which is in part explained by how little this sub-version of Latin is studied and by the fact that available corpora are not only scarce, but disperse or undisclosed. The state of art of these is still far from that of Latin-derived languages. Automatic hypothesis coming from an updated PoS tool; a larger NER model and embeddings or transformers trained or fine-tuned on corpora of over 50M words will undoubtedly be greatly appreciated in future modelizations.

In summary, our best model is able to successfully recognize major and minor sections in medieval charters and we can expect a good performance in many other medieval charters, since diplomatic models are extensively used across European writing legal traditions. The main problem seems to lie not so much in the overfitting due to the imbalanced nature of the corpus, but rather in the existence of two levels of patterns and relationships between the features and the target: one close to the formula and easy to match; and another one moving away from formulae and less common, therefore much more difficult to model. The fitting on the latter can be greatly improved by splitting major classes and extracting lexical and semantic features, thanks to NLP tools.

VII CONCLUSION

We have presented a Bi-LSTM-CRF model for automatic LTS in medieval charters. The performance shows a ratio of accuracy of over 0.85 in F1 measure on 17 of 23 sections and a ratio of 0.70 to 0.84 on the remaining 6 sections. Finally, we have shown that our model is robust on an external set of charters, which confirms it can be generalized to charters from other periods and origins. Our discussion tries to confirm that the main issues are related to the existence of a double discursive pattern according to the section type and function. One pattern is very close to the formula and displays a small vocabulary, the other one uses a larger vocabulary and moves away from the formula to a varying degree. This issue is further amplified by the tension between normative expression and innovative or individual expression during the writing of charters; both are common questions in diplomatics.

We have also demonstrated the positive impact of custom NLP resources on modeling for medieval Latin. These resources and the corpus that we have annotated to evaluate our models are new contributions in themselves.

While this work concerns the development of a neural LTS model, several research areas can benefit from its application. In the field of indexation and information retrieval, a vast database of medieval charters collections could be easily organized both by metadata and by content. As each section displays specific details of a charter, granular metadata can be easily detected: name and role of the participants, type of juridical action and content, dates, formula tradition, etc. This information combined with named entity models can facilitate cross-selections, for example selecting documents produced in a particular chancellery; signed by a certain notary or family; concerning a specific place in view of reconstructing the movement of land donations, etc.; thus enhancing research tools about textual tradition. In a broader perspective, research about inter-textuality, text circulation or reuse, juridical text composition and concepts representations could use this tool to split texts more easily by constitutive units and to classify formulae and phrasemes. Finally, in the recent area of NLP for historical languages, the automatic hypothesis can be integrated as an extra feature for other learning tasks such as topic classification, text summarizing or the handwriting recognition in medieval diplomatic texts.

VIII FUTURE WORK

As mentioned above, the labelled data comes from notarial Italian tradition. The model seems to be highly generalizable but a training on more varied data must be encouraged. Current advances in character and handwriting recognition techniques have made hundreds of new cartularies and charters collections available in recent years. A bootstrapping approach to obtain silver-standard annotations of these other charter collections could help boost performance on various typologies.

Moreover, two overlapping sub-sections (*Formulae* and *Clausulae*, partially used in our training) must be reannotated in order to build a finer-grained model and a better represented corpus. The automatic detection of these parts, which are a major point of interest for medieval studies, will be the highlight in LTS charters models.

Finally, more recent techniques such as contextual embeddings and fine-tuning on pre-trained models based on self-attention mechanisms seem reach state-of-the-art performances for most language processing tasks without the need to deploy complex features engineering. Future experiments will determine whether they may also be suitable for linear text segmentation on historical texts.

IX MODEL REPOSITORIES

The model source code and corpora supporting this work are available on our git repository:

https://gitlab.com/magistermilitum/diplomatics_sections_latin_charters/

An online web-based application of our model on raw text is also available in beta version at:

https://diplomatics-sections.herokuapp.com/

X NOTES

Translation of the charter in figure 1:

[1]In the year 1118 from the Incarnation of our Lord Jesus Christ, Sunday of October, eleventh indiction. [2]In the church of Saint John, located out of the city of Brescia, [3]we, Cocalius and Brixianus, brothers, from the mentioned place of Coccaglio, who have declared to live according to the Longobard laws, present as offerers and donors in the aforementioned church of *Saint John de Fora*, have told the people present: [4]While the course of human life is taking place in a good state of health and the soul is invigorated by the full understanding of the mind, thus a man should always order and arrange what seems to be profitable for him, so that when the Lord calls him out of this world, he will not be judged for his negligence, but he will be considered as a pious man for the good he has done.

[5]It is known that we, the above mentioned, Cocalius and Brixianus, brothers, since we have neither son nor daughter, we want all our goods to be arranged and disposed in the manner that we have indicated below and that our wish has resolved for the mercy of our soul and that of our parents. [6]Thus, we want and firmly establish, and by means of this donation charter we confirm that, from the present day, all our legal possessions end up going to the church of Saint-John, both movable and immovable properties, as well as the livestock; those that we currently have, or those that we could eventually acquire, entirely, both in Coccaglio and in any other place, anything we legally possess that can be found, entirely, so that the church of Saint John can, from the present day, as legal owner, do anything It wants, [7]with no opposition at all from us or our heirs, for the mercy of our soul and that of our parents.

[8]Held in the said church of Saint John de Fora. [9]Signed by the hands of the mentioned Cocalius and Brixianus, who requested this donation charter to be made for the mercy of their soul and that of their parents. [10]Signed by the hand of the witnesses Lanfrancus de Cologne and Iohannes Curtisi and also of Iohannes de Cocalio, and that one named Nanus. [11]Petrus, priest of this church of Saint John, on account of this church, accepted this donation charter. [12]I, Guido, notary, as requested, wrote down this donation charter, fulfilled after the negotiation was completed.

References

Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In *International Conference on data warehousing and knowledge discovery*, pages 305–316. Springer, 2008.

Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *HistoInformatics@ DH*, pages 67–71, 2016.

Michele Ansani. Edizione digitale di fonti diplomatiche: esperienze, modelli testuali, priorità. *Reti Medievali Rivista*, 7(2):1–1, 2006.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas W Oard, and Philip Resnik. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, 2020.

Robert Henri Bautier. L'authentification des actes privés dans la france médiévale. In *Notariado público y documento privado: de los orígenes al siglo XIV: actas de VII Congreso Internacional de Diplomática, Valencia, 1986*, pages 701–772. Conselleria de Cultura, Educació i Esport, 1989.

Bruno Bon. Omnia: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3). *Bulletin du centre d'études médiévales d'Auxerre— BUCEMA*, (15), 2011.

Benjamin Burkard, Georg Vogeler, and Stefan Gruner. Informatics for historians: Tools for medieval document xml markup, and their impact on the history-sciences. *Journal of Universal Computer Science*, 14(2):193–201, 2008.

Pierre Chastang, Laurent Feller, and Jean-Marie Martin. Autour de l'édition du registrum petri diaconi. problèmes de documentation cassinésienne: chartes, rouleaux, registre. *Mélanges de l'école française de Rome*, 121(1): 99–135, 2009.

Heinrich Fichtenau. Arenga: Spätantike und mittelalter im spiegel von urkundenformeln. *Mitteilungen des Instituts für Österreichische Geschichtsforschung/Ergänzungsband*, 1957.

Petra Galuščáková and Lucie Neužilová. Low resource methods for medieval document sections analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130. Association for Computational Linguistics, 2016.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Abhyuday N Jagannatha and Hong Yu. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*, volume 2016, page 856. NIH Public Access, 2016.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. *arXiv preprint arXiv:1803.09337*, 2018.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 2001.

Yanzeng Li, Tingwen Liu, Diying Li, Quangang Li, Jinqiao Shi, and Yanqiu Wang. Character-based bilstm-crf incorporating pos and dictionaries for chinese opinion target extraction. In *Asian Conference on Machine Learning*, pages 518–533, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Milagros Cárcel Ortí. *Vocabulaire international de la diplomatique*, volume 28. Universitat de València, 1997.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154, 2013.

Vasudeva Varma. Attention-based neural text segmentation. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772, page 180. Springer, 2018.

Houquan Zhou, Yu Zhang, Zhenghua Li, and Min Zhang. Is pos tagging necessary or even helpful for neural dependency parsing? In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 179–191. Springer, 2020.

A CHARTERS SECTIONS EXAMPLES

Latin	English
Invocatio	
1. In nomine domini nostri Iesu Christi. 2. In nomine sanctae et individuae Trinitatis.	1. In the name of our Lord Jesus Christ 2. In the name of the Holy and indivisible Trinity
DTCRON	
1. Anno dominice incarnationis milleximo centeximo septuageximo quartodecimo die mensis decembris, indicione tertia	1. In the year eleven seventy of the incarnation of our Lord, on the fourteenth day of the month of December, third indiction
DTTOP	
1. Actum in pallatio episcopi Parme 2. Actum suprascripta civitate Mediolani	1. Done in the bishop's palace of Parma 2. Done in the aforementioned city of Milan
Inscriptio	
1. dilectis filliis Matutino abbati et fratribus de Cerreto 2. Omnibus episcopis, abbatibus, ducibus, comitibus, actionariis vel reliquis fidelibus nostris tam presentibus quamque futuris.	1. To my beloved sons, the abbot Matutino, and the brothers from Cerreto. 2. To all the bishops, abbots, dukes, counts, agents or others of our devoted, not only the current but also the future ones.
Intitulatio	
1. Alexander episcopus servus servorum Dei 2. Chuonradus divina favente clementia Romanorum imperator augustus.	1. Alexander, Servant of the servants of God 2. Conrad, by the favor of divine clemency, august emperor of the Romans

Table 5: Some examples of charters sections.

Latin	English
Promulgatio	
1. Quapropter notum sit omnibus nostris fidelibus, presentibus scilicet ac futuris, quod... 2. Omnibus vobis ceterisque nostris fidelibus notum fieri volumus...	1. wherefore, let it be known to all of our devoted, both, current and future, that... 2. We wish to be made known by you and by all the rest of our devoted, that...
Formulae	
1. Eo modo ut ipse Putianus suique heredes et cui dederint habeant et teneant et faciant exinde quicquid voluerint, sine contradicione suprascripte abbatisse suique successatricum et partis ipsius monasterii, et cum earum defensione ab omni homine cum racione. 2. Reservando in nobis usumfructum tamdiu quam visi sumus, et post nostrum decessum usufrucctus revertatur ad proprietatem. Et post suprascriptorum decessus suprascriptus prepositus vel eius succesor et clerici suprascripte ecclesie debent facere anoale illorum.	1. And in this way in order for Putianus himself, and his heirs and those to whom they have transferred [their possessions], to have, retain and do whatever they want [with this land] without any opposition from the aforementioned abbess or her successor or anyone from the monastery, and with the defense from all men with reason. 2. Reserving to us this usufruct while we are alive; and after our death the usufruct must be returned to its owner. And after the death of the above-mentioned prior or his successor and the clergy of the aforementioned church must celebrate an annual [mass] for them.
Clausulae	
1. Si quis vero, quod futurum esse non credo, si ego ipse Ingelerius, quod apsit, aut ullus de eredibus hac proeredibus meis seu quislibet ulla opposita persona contra hanc cartula ire aut eam infrangere conaverimus.... 2. et si tale ordine defendere non potuerimus aut si contra cartam agere quesierimus, tunc in duplum suprascripta vendita vobis restituamus sicut pro tempore fuerit meliorata aut valuerit sub estimacione in consimile loco.	1. If someone, if I, myself Ingelerious (God forbid!), or any of my heirs or pro-heirs, or any other person tried to oppose or break this charter (although I do not believe that it will happen).... 2, And if we had been unable to protect to such commitment or if we had wanted to go against this charter [that we signed], then we should refund you double of the aforementioned sale as well as the value of its improvements or an asset of similar value, after estimating it, in a similar place.
Rogatio	
1. Unde due carte uno tenore scripte sunt et rogate sunt fieri, quia sic inter eos convenit. 2. Prenominatus Petrus hanc cartulam fieri rogavit ut supra.	1. Therefore, two documents were written with a single tenor and they were asked to be made because thus was agreed between them [the stakeholders] 2. The aforementioned Peter asked that this charter be made, as above [mentioned]
SMC	
1. Signum + manus suprascripti Somenze, qui interfuit et suprascripto infantulo consensit. 2. Signum + manus suprascripti Arialdi patris sui, qui ei consensit ut supra et ad confirmandum manus posuit.	1. Signed by the hand of the aforementioned Somenze, who mediated and gave consent [to make this business] to his aforementioned son. 2. Signed by the hand of the aforementioned Arialdus, his father, who gave consent as before and put his hand to confirm it.
SMF	
1. Signum + suprascripti Lazari qui fideiussor estitit ut supra. 2. et iamdictorum Aenrici et Cadole et Riste, qui fideiussores estiterunt ut supra.	1. Signed by the aforementioned Lazarus who acted as guarantor as before [mentioned] 2. and of the already mentioned Aenricus and Cadolus and Ristus, who acted as guarantors as before [mentioned].
Corroboratio	
1. Quod ut verius credatur et diligentius observetur, manu propria roborantes de anulo nostro subter insigniri iussimus. 2. Quod ut ratum et inconvulsum omni tempore permaneat, presentem inde paginam conscribi et inpressione sigilli nostri insigniri iussimus.	1. and for this to be held as true and to be observed with greater diligence, we confirm it with our own hand, order it to be undersigned by our stamp. 2. and for this [charter] to remain ratified and unalterable for all time, I thus conscribed the present page and we ordered that it be signed with the print of our seal.
Subscriptio	
1. (S) Ego Arialdus iudex interfui et rogatus subscripsi. 2. (SM) Ego Arnulphus Boccardus hanc commutationem feci et subscripsi.	1. (Signature) I, Arialdus, was present as judge and I subscribed it under request. 2. (Signa manus), I, Arnulphus Boccardus made and subcribed this exchange.
IT (Index Testium)	
1. Interfuerunt Vivencius Piscator et Bornus Fusarius testes. 2. Interfuerunt testes Cassius de Lanpuniano et Saccus de la Piscina et Baldizonus Stampa et Revegiatus Guazonus et multi alii.	1. They were present as witnesses : Vicencius Piscator et Bornus Fusarius 2. They were present as witnesses : Cassius de Lanpuniano and Saccus de la Piscina and Baldizonus Stampa and Revegiatus Guazonus and many others.

Table 5: Some examples of charters sections (continued).