Spoken Latin behind Written Texts: Formulaicity and Salience in Medieval Documentary Texts

Timo Korkiakangas, Department of Philosophy, Classics, History of Art, and Ideas, University of Oslo

Email and postal addresses: timo.korkiakangas[ät]ifikk.uio.no, Georg Morgenstiernes hus, Post Box 1020, Blindern, 0315 Oslo, Norway

Abstract

This study exploits treebanking to investigate how spoken language infiltrated into legal Latin in early medieval Italy. Documents are always formulaic, but they also always contain a 'free' part where the case in question is described. This paper uses this difference to measure how ten linguistic features, representative of the evolution that took place between Classical and Late Latin, are distributed between the formulaic and free parts. Some variants are attested equally often in both parts of the documents, while perceptually or conceptually salient variants appear to be preserved in their conservative form mainly in the formulaic parts. Conceptual salience will be defined as the cognitive prominence of a (syntactic) construction.

Diese Studie nutzt das Treebanking, um zu untersuchen, wie die gesprochene Sprache in das Urkundenlatein des frühmittelalterlichen Italiens eingedrungen ist. Urkunden sind immer formelhaft, enthalten aber immer auch einen "freien" Teil, in dem der fallspezifische Inhalt beschrieben wird. Dieser Artikel macht sich diesen Unterschied zunutze, um zu ermessen, wie zehn sprachliche Züge, die für die Entwicklung zwischen klassischem und spätem Latein repräsentativ sind, zwischen den formelhaften und freien Teilen verteilt werden. Einige Varianten sind in beiden Teilen der Urkunden gleich oft bezeugt, während die perzeptuell oder konzeptionell saliente Varianten in klassischer Form vor allem in den formelhaften Teilen erhalten zu sein scheinen. Die konzeptionelle Salienz wird als die kognitive Bedeutung einer (syntaktischen) Konstruktion definiert.

Cette étude exploite le treebanking pour étudier comment le langage parlé s'est infiltré dans le latin légal de l'Italie du Haut Moyen Âge. Des documents sont toujours formels, mais ils contiennent aussi toujours une partie «libre» où le cas en question est décrit. Cet article utilise cette différence pour mesurer comment dix caractéristiques linguistiques, représentatives de l'évolution qui avait lieu entre le latin classique et le latin tardif, sont réparties entre les parties formelles et libres. Certaines variantes sont attestées également souvent dans les deux parties des documents, tandis que les variantes perceptuellement ou conceptuellement saillantes semblent être conservées sous leur forme classique principalement dans les parties formelles. La saillance conceptuelle sera définie comme la proéminence cognitive d'une construction (syntaxique).

Keywords: treebank, salience, documentary Latin, Early Middle Ages, scribe, diplomatics, formulaicity, language acquisition

1. Introduction and objectives[1]

In this paper, I use syntactically annotated data made available in a dependency treebank to explore how spoken-language features infiltrate into written texts in historical text corpora with conservative text genres. This is important because with historical language data the extent to which the written surface reflects the reality of the spoken language is usually unknown. At the same time, the very same texts are often the only material available for tracking spoken-language developments. This study discusses the subject by quantitatively examining the mechanisms that determined why certain spoken-language features crept into the documentary Latin of early medieval Italy – and why others did not. In this way, the study seeks to establish criteria for using written historical texts for the study of spoken language. The aim is thus methodological, but the analyses of specific constructions chosen to illustrate the approach also provide new linguistic findings.

The methodology involves a corpus study to isolate linguistic features that are sensitive (or insensitive) to the formulaicity of the documents. All spoken and written communication heavily relies on prefabricated fixed or semi-fixed expressions, i.e. formulae (MacKenzie & Kayman 2018). In early medieval Italian documents, formulae mean standard phrases and clauses that recur in multiple documents of the same type. These phrases and clauses, which guaranteed the juridical validity of the contents, can be identified by comparing documents with each other. In traditional diplomatic terminology, formulaic phrases are called, for example, invocation, inscription, and corroboration (Guyotjeannin & al. 1993: 72–85). Yet, documents also contain one or more non-formulaic slots where the distinctive characteristics of the case, such as the extensions of the sold property, are described in detail. Sabatini (1965) noticed that the language utilised in these slots differs considerably from that of the centuries-old formulae.

Documentary Latin, a concept utilised, for example, by Sabatini, represents an ideal data set for this particular study because it allows for a contrasting of the linguistically conservative formulae with the linguistically innovative non-formulaic passages, which are assumed to draw on the early medieval spoken idiom. This kind of examination of two evolutionary stages of Latin, i.e. conservative features derived from Classical Latin and innovative features developed by the 8th/9th century AD, is an indirect way of addressing diachronic variation within a relatively synchronic treebank (see §2).

The theoretical framework adopted here highlights the role of salience, a factor that will be proposed to determine the distributions of conservative and innovative forms and constructions. In this endeavour, certain findings of second language acquisition studies prove to be helpful because it can be plausibly argued that the scribes, native speakers of an early Romance vernacular, learnt the written documentary Latin as a second language (L2).

2. Data

This study is possible due to the Late Latin Charter Treebank (LLCT), a syntactically annotated corpus of original private documents, i.e. charters, from Central Italy (519 texts, c. 226,000 words, Korkiakangas & Passarotti 2011).[2] The treebank method allows for the study of all the grammatical domains from lexicon to syntax, surpassing the possibilities provided by simple morphologically tagged text corpora. Consisting mainly of contracts about transferring landed property, the LLCT documents were written in historical Tuscia, a region which comprises most of modern Tuscany, between 714 and 869. The time span is considered too short to enable a normal diachronic approach, hence the decision to contrast conservative classical and innovative Late Latin features.

[1] I am deeply grateful to the reviewers for their precious comments. I also thank Dr. Hilla Halla-aho for commenting on a previous version of the article.

[2] LLCT is currently being enlarged with ca. 200,000 newly annotated documentary Latin words (LLCT2).

LLCT is based on three copyright-free diplomatic editions and is available for download (see Appendix). As for morphological and syntactic annotation, LLCT is based on the Ancient Greek and Latin Dependency Treebank (AGLDT) style codified in the *Guidelines for the Syntactic Annotation of Latin Treebanks* (Bamman et al. 2007). Since documentary Latin is a non-standard variety that often contains ambiguous morphological and syntactic features, Korkiakangas & Passarotti (2011) introduce a set of additions and modifications to the *Guidelines*, designed for Classical Latin.

In early medieval Italy, scribes did not copy documents from model document collections, which were used later in the Middle Ages, but reproduced the wordings of the documents from memory (Schiaparelli 1933: 3). This, together with the fact that classical standards were obviously not strictly required from documentary texts, led to considerable linguistic variation, fruitful for variationist and diachronic studies of the linguistic situation at a time when the transition from Latin to an Italo-Romance vernacular must have been well advanced to all appearances in spoken language. Indeed, documentary Latin is a variety of non-standard Latin that has several features proven to originate from the spoken language of the time, either through direct reflection or indirectly by way of misinterpretations of classical legal Latin (Korkiakangas 2016: 240). Obviously, the term 'non-standard' calls for a definition of 'standard'. No standardisation of language in the modern sense of the word was practised, i.e. the grammar and spelling were not canonised by the authorities. Nevertheless, the essentially classical orthography and morphology seem to have still served as a valued model for the best-written LLCT texts. Thus, there was rather a clear idea of a standard, in terms of a substantial consensus about 'correct' or 'accepted' morphology and spelling (Korkiakangas 2016: 36).

3. Formulaicity

Formulaic expressions guaranteed the juridical validity of early medieval Italian documents. Most formulae date from the imperial Roman chancery tradition. Sabatini (1965) emphasised that each document also contains non-formulaic case-specific parts which record non-universal features, such as descriptions of the transferred property or ownership central to the current legal act. This information primarily lies in the so-called disposition, i.e. declarative part, but may be scattered within and between the very formulae. Given that the scribes could not rely on prefabricated phrases when composing the case-specific descriptions, there appears a distinct difference in language between these so-called 'free' parts and the formulaic parts which were anchored to the age-old legal tradition, alien to the everyday language. The free parts understandably have recourse to the spoken idiom, while the formulaic parts reflect the spoken language only by way of hypercorrections.

Sabatini relied upon the free/formulaic dichotomy to support his theory on the formation of the Italian plural forms (Sabatini 1965: 978–987). My hypothesis is that this distinction is indicative of linguistic change and variation on other levels of linguistic representation as well. So far, the lack of any annotated corpus of documentary Latin has made large-scale quantitative research impossible. To overcome this, I have provided LLCT with annotation that separates the free parts, most notably the disposition, from the formulaic parts, i.e. the rest of the document (Korkiakangas & Lassila 2013). The following quotes from a sales contract (*CDL* 26, Lucca, March 720) illustrate a typical free sentence from the disposition (1) and a typical formulaic sentence called *sponsio* (2). The quotes show that the distinction between free and formulaic is not clear-cut: both sentences contain both free and formulaic elements. The free/non-formulaic words are underlined.

(1) *Consta me Aufrid v(ir) d(evotus) hanc die vendedisset et vendedi, tradedisset et tradedi vobis Aunuald, Teutpald, Leonaci, Petronaci, Teutp(er)t, Dommuli, Vuilifrid, Nandulo, Geminiano clerico, Teuderaci ortu meum quem avire videor ante s(an)c(t)o Selvestre, qui latere tene prope curte vel orte s(an)c(t)i Selvestri, rectu casa Domnici vel de filio Iovanni.*

'It is manifest that I, Aufrid, *vir devotus*, in the present day sell and hand over to you Aunuald, Teutpald, etc., my orchard which I am known to have in front of the church of Saint Sylvester and which has its border close to the court and garden of Saint Sylvester, by the house of Domnicus and his son Iohannes.'

(2) *Et*, *sicot non crido*, *ut si ego aut eridis meus vos molestaverimus aut da qualivet homine vobis defensare non potuero*, *spondeo vobis cunponere dupla condicionem*.

'And, which I do not believe, if I or my heirs molest you or if I cannot defend you from whoever man, I promise to recompense double the price.'

Although free parts usually contain some formulaic elements and vice versa, the formulaicity status is assigned in LLCT to each sentence because otherwise syntactic features, which operate on a sentence level, become difficult to examine (Korkiakangas 2016: 25–29).

4. Theoretical background and research setting

I have selected ten varying linguistic features that are traditionally assumed to reflect a language change in Late Latin. The features will be described in §5, and §6 shows which of these features are sensitive and which are insensitive to formulaicity. My working hypothesis is that if a certain innovative feature is in a statistically significant way more frequent in the free parts of documents, it is more likely to reflect the current state of the spoken language. Conversely, a conservative feature is expected to occur more frequently in the formulaic parts. If, however, an innovative feature is clearly more frequent in the formulaic parts or a conservative feature in the free parts, they cannot actually be innovative and conservative, respectively, at the time of LLCT, and have to be subjected to further study. This did not, however, happen with the features examined.

To find out which types of features are sensitive to formulaicity, the ten features were chosen in such a way that they cover the grammatical landscape of the language broadly. To be able to examine all kinds of features from lexicon to syntax within a single framework, I adopt here a cognitive view of grammar in terms of a syntax-lexicon continuum (Table 1). The syntax-lexicon continuum is a uniform model of grammatical representation which locates constructions on a continuum according to their generality. Atomic means that an item cannot be further divided into meaningful parts unlike in complex constructions, whereas a schematic construction subsumes specific constructions, like *adjective* subsumes *green*.[3]

Table 1. The syntax-lexicon continuum (Croft & Cruse 2004: 255, Broccias 2012: 738).

Rank	Grammar domain/ Construction type	Traditional name	Example
5	Complex and (mostly) schematic	Syntax	noun verb noun (= transitive construction)
4	Complex and (mostly) specific	Idiom	pull one's leg
3	Complex but bound	Morphology	noun-s
2	Atomic and schematic	Word class	pronoun, adjective
1	Atomic and specific	Word/lexicon	*this*, *green*

The selected linguistic features concern the following stages or construction types, as they are called in the construction grammar tradition, of the syntax-lexicon continuum: atomic and specific (lexicon), complex with bound morphemes (morphology), and complex and schematic (syntax). This ordering of grammatical domains partly overlaps with the classification of morphemes in free and bound lexical and free and bound functional morphemes

[3] For the theoretical motivation and a detailed definition of the terminology, see Croft & Cruse 2004: 247–256.

(Croft & Cruse 2004: 254–256). Lexical morphemes carry meaning by themselves (e.g. *dog*), whereas functional morphemes (e.g. *of*) specify the relationship between other morphemes. Free morphemes are free-standing words (e.g. *dog*, *of*) while bound morphemes occur only as part of other words (e.g. *form* in *transform*, *-s* in *dogs*). It needs to be emphasised that the syntax-lexicon continuum is a simplification of a complicated linguistic reality, like all organisational schemes intended to capture the whole of grammar.

Importantly for the present study, the ranking inherent in the syntax-lexicon continuum (see the numbers in Table 1) can be considered to reflect the cognitive effort involved in recognising the linguistic domains in question, at least under certain conditions. I assume that, roughly speaking, a higher ranking means greater mental effort required by a language learner to adopt the features that pertain to that domain, due to the higher complexity of these features. This higher complexity is assumed to result from the higher-ranking construction types being generalisations based on a large number of exemplars and lower-level generalisations. I expect this to apply principally to language-learning situations, not to the language processing of L1 language users in general. These assumptions are in harmony with the overall picture sketched by the studies on the L2 acquisition order of grammatical categories: lexical morphemes appear to be acquired before functional morphemes and, within each of these groups, free morphemes are acquired before bound ones (Zobl & Liceras 1994: 172–175, Goldschneider & DeKeyser 2001: 28). An effectively similar picture arises from the processability-informed theories of language acquisition, which assume a complexity-based processing hierarchy: learners are supposed to first acquire the relations between lemmas, then those within words (lexical morphology), within phrases, between phrases, and, finally, between clauses (Pienemann 1999: 7–9). I exploit the complexity ranking combined with the morpheme classification (free/bound) to explain the conceptual salience in this study (see §6).

The examined features are known to have one variant associated with Classical Latin and another associated with Late Latin or Italo-Romance, e.g. GENITIVE CASE FORM vs. PREPOSITIONAL PHRASE WITH *DE*. I call these diachronic variants conservative and innovative, respectively. Although the starting point and the endpoint of the development are known, the chronology often remains uncertain: it is not always clear to which extent the innovative variant has established itself and ousted the conservative variant in the spoken language. For example, it is known that the replacement of the genitive case form by the prepositional phrase with *de* was a gradual process which took centuries and arguably was not yet fully completed by the time of LLCT (Valentini 2017: 47ff.).

Defining conservative and innovative variants is not only problematic with respect to diachrony. The need to define variant pairs often leads to forced dichotomies which cannot take into account the various nuances connected with register or other preconditions, amply examined in various studies (e.g. Valentini 2017 for genitive/*de* PP). With some of the features, such as the dative case form, the conservative variant was not replaced by any single innovative construction, but rather by a plethora of (partly) new means of expressing recipiency. In these cases, no complementary distribution between conservative and innovative variant can be established, and only the conservative variant lends itself to quantification. Although this variant cannot be meaningfully compared to any other, its relative frequency in the total word count can be calculated.

5. Linguistic features

The examined features are presented in the order they appear in Table 2 (§6).

1) NON-CLASSICAL LEMMAS: 81 innovative lemmas, e.g. *barba* ‘uncle’, *cambium* ‘exchange’, *fossatum* ‘ditch’, *petia* ‘piece’ (see the full list in Appendix). This first feature is not part of the analysis proper, but is introduced as a ‘calibration variable’. Lexicon cannot contribute genuinely to the present analysis because vocabulary is not optional, like grammatical variants, but is determined by the propositional content of the phrase. For example, several non-classical agricultural words, such as *tessero* ‘to mark boundaries with signs’ or *cavallarius* ‘horsekeeper’, are relevant

in free parts, where the highly varying case-specific traits of the transferred landed property are described. Instead, formulaic parts establish the universal legal circumstances of the act and cultivate classical technical terminology, such as *indictio* 'dating cycle' or *confirmatio* 'confirmation'. Thus, formulaic and free parts seem to consist of different vocabulary, which is assumed to be visible in the formulaicity distribution of conservative and innovative lemmas. This is not the case with morphological and syntactic features, where the distribution of conservative and innovative variants is not predetermined by the propositional content, but is expected to depend on the different prestige attributions between formulaic and free parts. The innovativeness of the lemmas has been verified by cross-checking them in Lewis & Short's *Latin Dictionary*, available through the Perseus Word Study Tool (http://www.perseus.tufts.edu/hopper/morph?la=la), and in Du Cange's *Glossarium Mediae et Infimae Latinitatis* (http://ducange.enc.sorbonne.fr/).

2) FUTURE PERFECT FORM: e.g. *apparuerit* 'it/he/she will have appeared'. The classical future perfect survives in Romance only sporadically and mainly in idiomatic expressions (Lausberg 1962: 205–206; cf. Weber 1924: 60–62, who calls the form 'conditional'). In documentary Latin, the form is utilised to anticipate a future execution of what was agreed on by the contracting parties.

3) DATIVE PLURAL IN *-BUS*: e.g. *potestatibus* 'to dominions'. Along with the general decline of the Latin case system, the dative was increasingly replaced by other means of expressing recipiency, such as the prepositional phrase with preposition *ad* 'to' (Adams 2013: 278–294). Since the *ad* PP is utilised in LLCT for a plethora of adnominal relationships of ambiguous interpretation, I could not calculate the relative frequency of the dative form and *ad* PP. Thus, I instead counted the percentage of the dative form in the total number of words. The Romance personal pronominal system retains the indirect object forms as clitics (Salvi 2011: 322–324). Therefore, pronouns are excluded from this investigation. Only the 3rd, 4th, and 5th declensions with the phonologically/graphically substantial (see §6.2) ending *-bus* are examined.

4) ADNOMINAL POSSESSION: genitive case form vs. prepositional phrase with *de*, e.g. *regis* vs. *de rege* 'of king'. The *de* PP competed with the original possessive strategy, the genitive case, and, in the Romance languages, became the main means of expressing adnominal possession (Valentini 2017, Adams 2013: 267–274). The majority of the possessive constructions are still expressed by the case form in both formulaic and free parts in LLCT.

5) PHRASAL COMPLEMENTATION: accusative and infinitive (ACI) vs. complementiser clause, e.g. *Sichiprandum*.ACC *scribere*.INF *rogavi* vs. *rogavi ut*.COMP *Sichiprandus scriberet* 'I asked S. to write'. In Latin, the innovative complement clause, introduced by a complementiser, had long rivalled the accusative and infinitive construction, the principal means of complementation in Classical Latin. The Romance languages have generalised the complementiser pattern (Zamboni 2000: 119–120, Ledgeway 2012: 244ff.). Considered a classical prestige feature, ACI is the prevailing complementation strategy in LLCT, likely because it was part of some documentary formulae.

6) ABSOLUTE CONSTRUCTIONS: e.g. *iuvante Deo* 'God willing'. It is agreed that the Latin absolute constructions were typical of (classical) literary texts and hardly occurred in spoken language (Väänänen 1981: 166–168). They continued to be utilised as stylistic prestige features in LLCT, apart from being part of certain formulae, such as *regnante* + the name of the king 'under the reign of N'. Here only constructions with participles are examined.

7) SECOND-PERSON SINGULAR: form with classical *-s* vs. without *-s*, e.g. *teneas* vs. *tenea* 'you should hold'. The classical 2nd-person singular ending of all the active indicative and subjunctive tense forms, except for the perfect, was *-s*. All the word-final consonants were either lost or weakened by the Early Middle Ages in the spoken language of Italy (Väänänen 1981: 67–69, Adams 2013: 132ff.), and the second-person singular of the subjunctive came to end in /a/ while, in the indicative, the outcome was /i/ (Maiden 1996).

8) DATIVE SINGULAR: e.g. *potestati* 'to dominion'. Like the dative plural, the dative singular form was being replaced by other constructions. See above 3.

9) SUBJECT CASE ENCODING: nominative vs. accusative subject, e.g. *ista portio*.NOM *sit in potestate tua* vs. *ista portionem*.ACC *sit in potestate tua* 'let this parcel be in your possession'. In Late Latin, the accusative is known to have extended partially to the subject function as a symptom of a major reorganisation of grammatical relations, i.e. the so-called semantic alignment, where the Agent-like arguments tended to be encoded with the nominative and the Patient-like arguments with the accusative (Ledgeway 2012: 328–335, Korkiakangas 2016: 57–74). Here, only those non-pronominal 3rd-declension subjects are counted where the morphological contrast between the nominative (*portio*) and the accusative (*portione*(*m*)) remained intact in Late Latin for phonological reasons (Korkiakangas 2016: 111). The few person and place names were also excluded because their case endings are often ambiguous.

10) VERB/OBJECT ORDER: e.g. *casam donavit* (OV) vs. *donavit casam* (VO) 'he/she/it donated a house'. The most typical verb/direct object order in Classical Latin was OV, although much variation existed. As time passed, the originally mostly pragmatically conditioned Latin order became increasingly syntactically motivated and gradually turned into the VO order, predominant in the Romance languages. OV still remained frequent for a long time and was obviously considered a stylistic means (Ledgeway 2012: 225–235, Zamboni 2000: 101–102). Here, only the clauses with one non-coordinated finite verb and non-pronominal direct object are examined because they are prototypical and unambiguous (for the motivation of this choice, see Korkiakangas 2016: 196). Main and subordinate clauses are treated equally. Pronominal objects are discarded because they have peculiar syntactic characteristics of their own, such as the relative pronoun's typical clause-initial position.

6. Results and their interpretation

Table 2 presents the examined features in two groups according to whether they appear to be sensitive to the formulaic/free distinction or not. For each feature that only has a conservative variant, its share in the total of words of LLCT is presented (parts per thousand values, ‰). For the features which allow the identification of both conservative and innovative variants, the distribution of the conservative and innovative variant is presented (percentages). The formulaic parts contain 169,520 words and the free parts 56,314 words. The statistical significance is the *p* value of the Chi-Squared test (95% confidence interval). When calculating the Chi-Squared test for those variables that present only the conservative variant, the frequencies are compared to thousands of words. N indicates the population size, i.e. the number of occurrences.

Table 2. The examined features with their relative frequencies in formulaic and free parts of documents.

Statistically significant sensitivity to formulaicity		**N**	**‰ in total of words**		**% distribution**		**Sig. level**
Domain	**Measured variant**		**Formulaic**	**Free**	**Formulaic**	**Free**	
lexicon	81 innovative lemmas	767	1.5	9.1	-	-	$p < 0.001$
morphology	future perfect form	2,315	12.3	3.5	-	-	$p < 0.001$
	dative plural form	154	0.8	0.4	-	-	$p = 0.004$
morphology/syntax	adnominal genitive form	8,027	37.2	30.6	90.3	69.2	$p < 0.001$
	adnominal *de* PP	1,441	4.0	13.6	9.7	30.8	
syntax	ACI	982	5.6	1.8	84.7	57.8	$p < 0.001$
	conjunction clauses	235	0.9	1.3	15.3	42.2	
	absolute constructions	916	4.9	1.5	-	-	$p < 0.001$
Statistically non-significant sensitivity to formulaicity		**N**	**‰ in total of words**		**% distribution**		**Sig. level**
Domain	**Measured variant**		**Formulaic**	**Free**	**Formulaic**	**Free**	
morphology	2nd person singular -s	248	1.0	1.4	89.0	87.6	n.s.
	2nd person singular not -s	32	0.1	0.2	11.0	12.4	
	dative singular form	3,878	15.8	21.4	-	-	n.s.

syntax	nominative subjects	278	1.2	1.3	74.8	65.8	n.s.
	accusative subjects	107	0.4	0.7	25.2	34.2	
	OV order	1,118	3.8	8.5	66.4	62.5	n.s.
	VO order	611	1.9	5.1	33.6	37.5	

The variables NON-CLASSICAL LEMMAS, FUTURE PERFECT FORM, DATIVE PLURAL IN *-BUS*, ADNOMINAL POSSESSION, PHRASAL COMPLEMENTATION, and ABSOLUTE CONSTRUCTIONS show a statistically significant dependence with the formulaicity variable. No statistically significant dependence is attested between formulaicity and the variables SECOND-PERSON SINGULAR, DATIVE SINGULAR, SUBJECT CASE ENCODING, and VERB/OBJECT ORDER.

6.1 Formulaicity and salience

This and the following section interpret the results of the quantitative analysis. The results seem to support the intuitive postulate adopted by earlier scholarship, namely, that the scribes did draw from different linguistic repositories when writing formulaic and free parts of documents (Sabatini 1965). This becomes evident on the basis of lexicon. NON-CLASSICAL VOCABULARY was utilised as a 'calibration variable'. It was thought that if the lexicon variable showed a clear difference between free and formulaic parts, it would be reasonable to carry out the formulaicity analysis with other, more complex domains of grammar. This assumption proves to be sound on the basis of the numbers of Table 2: the innovative lemmas occur six times more often in free parts than in formulaic parts.

Apart from lexicon, formulaicity is likely to affect the distributions of other features as well. The question is whether there is something in common to all the features that show a statistically significant difference between formulaic and free parts in Table 2. The numbers reveal that the complexity ranking based on the syntax-lexicon continuum alone is not enough to explain the behaviour of the examined features because the same linguistic domains may be sensitive or insensitive to formulaicity. The concept of SALIENCE turns out useful at this point. A widely used notion in semiotics, social psychology, and sociolinguistics, salience is a gradient property which operates on the physical world/cognition interface. In linguistics, salience can refer to the characteristics of the linguistic input/output itself or to those external-world factors that cause some parts of the input/output to become salient, such as the referent of the linguistic expression being bright-coloured or interesting to the language user (Cintrón-Valentín & Ellis 2016). In this paper, salience is understood in terms of how prominent or noticeable certain lexical items, morphemes, or syntactic constructions appear to a language learner in the linguistic input. Several quantitative and experimental L2 acquisition studies focus on the role of salience in language acquisition. For example, the eye-tracking measurements of Cintrón-Valentín & Ellis (2016) prove that the low perceptual salience of certain short inflexional morphemes essentially contributes to L2 learners' difficulty in learning them.

These findings based on modern language learning situations can be fruitfully extrapolated to early medieval Latin. I suggest that the statistically significant features of Table 2 involve a variant which is either 1) conceptually, i.e. in terms of its grammatical prominence, or 2) perceptually, i.e. in terms of phonetic or graphic substance, more SALIENT than those which do not show a statistically significant formulaicity distribution. In other words, the features with statistically significant sensitivity to formulaicity are salient forms or constructions in terms of one or the other of the mentioned salience types, or, in case they involve two or more variants, at least one of the variants is salient.

The criterion of perceptual salience, the amount of formal prominence or noticeability in terms of phonetic/graphic substance, seems to apply to the lowest ranking domains of the syntax-lexicon continuum, i.e. here words and morphemes. Most words are, as such, phonetically/graphically perceptible units that convey lexical meaning, whereas in morphology, the grammatical information is carried by morphemes of differing phonetic/graphic perceptibility. This is not the case with the highest ranking domain, i.e. syntax, where it often makes no sense to

speak about perceptual salience. In syntactic constructions, free-standing words or phrases are linked to each other by an underlying rule, and in Latin each involved word is usually encoded by a certain bound morpheme. Thus, the salience of a variant of a syntactic construction must rather be thought of as conceptual salience, which I define here as the grammatical prominence or noticeability of the syntactic rule to the language learner. I have adopted the term 'conceptual' to distinguish the just-described salience from 'semantic' salience, a vague term utilised to cover a vast variety of uses from the prominence of discourse referents to that of extra-linguistic entities (e.g. Chiarcos et al. 2011: 1–3).

To give an example, a construction is more salient conceptually if it involves free morphemes instead of bound morphemes, as in the case of COMPLEMENTISER CLAUSE vs. ACCUSATIVE AND INFINITIVE. It is true that in this case both variants can also be considered perceptually salient given that they consist of (more than one) free-standing words. Instead, the rule conditioning the SUBJECT ENCODING is considered non-salient because it involves only bound morphemes and is thus more unnoticeable. As regards perceptual salience, it may emerge, among other things, from the amount of phonetic substance, stress level, or usual serial position in a sentence (Dulay & Burt 1973: 409). In the case of written texts, graphic substance, i.e. characters, may determine salience, especially when the phoneme/grapheme relationship is weak, as it certainly was between written and oral codes in early medieval Italy.

Given that the early medieval Tuscan scribes very likely spoke a variety which might already be described as a Romance vernacular, they learnt documentary Latin in practice as a second language. Indeed, the innovative and conservative features examined here can be seen as characteristics of the scribes' native L1 and of the literary Latin L2 to be learnt, respectively.[4] The L2 studies about the acquisition order of grammatical morphemes lend this study a useful framework which also seems to be extendible to predominantly syntactic features. Goldschneider et al. (2001) have shown that the acquisition order of certain English morphemes is largely explained by perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and frequency. The authors claim that these five factors all constitute aspects of salience in a broad sense of the word (Goldschneider et al. 2001: 35).

The early medieval documentary scribes had imbibed the basics of Latin spelling and morphology when learning to write, but that was not enough for writing documents. They also had to adopt the formulae, and this was likely done by reading existing documents. The imperfect command of certain formulaic passages indicates that many scribes had memorised the formulae rather superficially, without profound comprehension of them. However, Classical legal Latin enjoyed a high prestige as the language of law.[5] The scribes knew they were supposed to use this venerable variety when writing documents, especially the formulae, which guaranteed the legal validity. I suggest that, during and after the original memorisation process, the learners who were to become scribes recognised the perceptually or conceptually salient conservative features more readily as Classical Latin forms than the less salient ones, and remembered to reproduce those more often in formulaic parts, which were considered vital for legal validity. This is also likely to work the other way around: the salient innovative features, which were felt to belong to the spoken language, were recognised as stigmatised, i.e. having a kind of negative prestige, and consequently avoided in the formulaic parts. The perceptually or conceptually non-salient features, instead, went unnoticed by the scribes and failed to be attributed a (positive or negative) prestige.

As a consequence, non-salient features appear to be distributed evenly (i.e. due to chance) between formulaic and free parts. The drive to recognise and imitate words and expressions that were considered echoes of the dignified legal tradition and correct grammar seems to result from the scribes experiencing external or internal normative pressures, especially when writing the formulaic parts. Apparently, this aspiration to classical grammar was a common phenomenon even though Italian documentary Latin seems to have been a recognised genre sanctified by

[4] On the other hand, the picture is complicated by that written Latin was still apparently considered the regulated form of the language people spoke at the time. On the metalinguistic change between Latin and vernacular, see Wright 1991.
[5] For the challenges involved in the reconstruction of prestige patterns in historical language varieties, see Sairio & Palander-Collin (2012) and Adams (2013: 841ff.).

the long traditions and not subject to similar corrective interventions of the authorities, as it was in Carolingian Gaul (Bartoli Langeli 2006: 28ff.).

6.2 Analysis of the morphological and syntactic features

I now go through the evidence to justify the above theoretical considerations. The distributions of the variants of the here examined morphological features seem to be plausibly explained by their perceptual salience, i.e. by the amount of phonetic/graphic substance of the feature or of one involved variant in respect of the other variant. The DATIVE SINGULAR form *potestati* 'to dominion' differs only by one character from, for example, the genitive singular form *potestatis* or from the accusative singular form *potestate*(*m*).[6] Since the case system had already largely collapsed and the use of the dative and genitive was, in all likelihood, no longer supported by the spoken language, it is probable that a form like *potestati* with the dative morpheme *-i* did not stand out enough from the paradigm where the most common form must have been something like *potestate*. This accusative-based form had possibly become the nearly all-purpose or default form in late spoken Latin and competed, perhaps, only with the nominative form *potestas*. It then became the only form of the noun in the Romance languages.[7] Assuming thus that *potestat-* was the most typical stem of the paradigm, morphemes resulting in forms such as *potestati* or *potestate* cannot be considered perceptually salient.

Instead, the DATIVE PLURAL form *potestatibus* 'to dominions' differs more clearly from the other case forms of the word: the ending -(*i*)*bus* is both graphically[8] and phonologically more substantial than, for example, the respective singular ending *-i*. Indeed, the enduring prestige of *-bus* is witnessed by its hypercorrect use in subject and object function (see also Sornicola 2012: 57–58), as in (3).

(3) [--] *ut p*(*er*) *singulos annos ego*, *dum vixero*, *et s*(*upra*)*s*(*crip*)*ti nepotes mei vel* <u>*heredib*(*us*)</u> *eor*(*um*) *dare et reddere debeam*(*us*) *ad ep*(*iscopu*)*m* [--] *unum sol*(*idum*) (*CDL* 285)

'[--] so that every year I, as far as I live, and my mentioned descendants as well as their <u>heirs</u> have to give to the bishop [--] one *solidus*'

The same explanation also applies to the 2nd-person singular and the future perfect forms. The SECOND-PERSON SINGULAR morpheme *-s* in the form *teneas* 'you should hold' is only one phonologically and graphically non-substantial sound/character, and is not particularly discernible from the most frequent forms of the same paradigm *teneat* (3rd person) and *teneam* (1st person). This minor distinction does not seem to have been enough to guarantee the scribes' attention and the subsequent recognition of the form's classical prestige in a situation where all the singular persons of the subjunctive were likely to end in /a/ in the spoken language. Instead, the FUTURE PERFECT affix *-er*(*i*)-, with the phonologically persistent *r*, forms an entire syllable and results, thus, in clearly more substantial forms, such as *apparuerit* 'he/she/it will have appeared'. These forms were easily associated with prestige. The future perfect leaps out from the verbal paradigm as one of the graphically longest inflexional forms, along with the subjunctive pluperfect. All this suggests that, for a single morpheme to be perceptually salient, it is required to stand out from the horizon of expectation consisting of the most common or typical forms of the paradigm.

How about features which involve two or more variants? I argue that for the formulaicity distribution to be statistically significant, at least one of the variants has to be salient, perceptually or conceptually. If the conservative

[6] Word-final vowels were likely to be vague in Late Latin and the same applies to the *-s* as well. The final *-m* had ceased to be pronounced very early on (Adams 2013: 62, 128–147).
[7] For the two-case system, see Zamboni 2000: 110–115, Korkiakangas 2016: 74–79.
[8] However, the two last letters were sometimes abbreviated in handwriting by a loop resulting roughly in -(*i*)*b*[9].

variant is perceptually salient, the same mechanism applies as with the morphological features above. Instead, if only the innovative variant is salient, it is avoided because it is recognised as stigmatised. Note that motivations of this kind can be perceived only as statistical tendencies because several simultaneous conflicting motivations are involved as well. Some scribes were more aware of classical grammar than others, and specific linguistic features were given highly idiosyncratic prestige attributions. Therefore, the distributions of the innovative and conservative features are nowhere near fully determined by formulaicity, as can be seen in Table 2. What is important is the statistically significant difference in relative frequency between the formulaic and free parts.

In the case of ADNOMINAL POSSESSION, i.e. genitive form replaced by prepositional phrase with *de*, an average genitive case form (except for the infrequent genitive plurals) is rather non-salient perceptually due to its minor phonetic/graphic substance, whereas the *de* PP, which consists of the preposition and its complement, is easier to notice both perceptually, because it is two words, and conceptually, because the words are two free morphemes (instead of a bound one in the genitive) (e.g. Zobl & Liceras 1994: 172). So, although a learner may not have recognised the non-substantial genitive form as a classical prestige form and, indeed, may not even have been aware of the two variants being in complementary distribution with each other in terms of prestige, he may have been able to induce a rule to avoid the innovative *de* PP because it does not occur in prestigious texts or it has been defamed by the school master. In this way, scribes could learn to shun the salient innovative variants when writing. However, especially in the free parts of documents, where the scribes had to resort to their own linguistic instinct rather than to ready-made formulae, they sometimes let the PP creep in. Of course, several LLCT scribes did not succeed well in attributing stigma to the PP given that the PP is found in 9.7% of cases in the formulaic parts.

The case of PHRASAL COMPLEMENTATION is essentially alike, although here both the variants, the infinitival construction and the complementiser clause, are syntactic constructions that involve several words. I maintain that the innovative variant, the complementiser clause, is also here the more salient one. This is because the structure of the complementiser clause only consists of free morphemes, consequently rendering the structure less abstract. The accusative and infinitive, instead, is based on the syntactic interplay of the bound morphemes in the subject noun and infinitive. In any event, the attribution of prestige status is enabled. ABSOLUTE CONSTRUCTIONS are salient only conceptually, although they may involve words which are perceptually salient *per se*, but this is not relevant for its recognition as a prestigious syntactic construction. On the other hand, it has to be remembered that the range of the absolute constructions attested in LLCT is lexically limited.

Let us now look at the two syntactic features that show no statistically significant sensitivity to formulaicity, i.e. SUBJECT CASE ENCODING and verb/object order. The statistical non-significance indicates that the use of the syntactic cases nominative and accusative as the case forms of the subject is not dependent on formulaicity. The morphological difference between the examined 3rd-declension nominative and accusative forms, such as *portio* and *portione*(*m*), might be perceptual *per se*, although attributing the salience status to either of the two forms would be difficult.[9] What is relevant, however, is that the underlying principle of the subject encoding (alignment of the arguments of the verb according to semantic or syntactic criteria) is particularly unnoticeable and abstract and, consequently, not conceptually salient. Intuitively, the rule that conservatively assigns the nominative case to all the subjects of finite verbs, rather than only to semantically active ones (the Late Latin way), is not to be learnt as easily as, for example, the rule 'remember to put the genitive and not the *de* PP'. Additionally, the rule of subject case encoding involves only bound morphemes, which keeps the variants perceptually non-salient.

As with subject encoding, learning a VERB/OBJECT ORDER which differs from that of one's native tongue also calls for a profound understanding of syntactic functions, a matter hardly promoted by school teaching. Indeed, syntactic issues seem to have passed largely unheeded in the Latin grammatical tradition. The linearisation of the verb and the object complement is abstract and involves bound morphemes to encode the constituent, so I consider it

[9] The salient one is perhaps *portio* because it differs by its stem from the rest of the paradigm and, moreover, *portione* was probably the all-purpose form of the day.

conceptually non-salient.[10] On the other hand, OV was still clearly the prevalent order in LLCT. Therefore, it can be asked to what extent the Romance-type VO had spread in the spoken language. Perhaps the crucial stabilisation of the VO order took place only after the period examined here. Indeed, in many Late Latin texts, the word order still essentially follows the same pragmatic constraints as in earlier Latin (Spevak 2010). It cannot be excluded, however, that the insistently favoured classical clause-final position had kept the OV order perceptually salient, at least to some scribes and with certain substantial verb forms (Ledgeway 2012: 229ff., Dulay & Burt 1973). All this said, the verb/object order is perhaps not as felicitous an indicator as would have been wished.

7. Conclusion

This study has examined the role of formulaicity in the distribution of conservative and innovative linguistic features in documentary Latin. The scribes had memorised the formulaic parts, which were considered the juridical heart of the document. Thus, the scribes reproduced many conservative, classical forms and constructions predominantly in these parts and, correspondingly, avoided using innovative spoken-language features in them. However, this sensitivity to formulaicity seems to be limited to features that the scribes recognised as prestigious or non-prestigious. The results of this study support a view that the recognition of prestige, i.e. a feature being Classical Latin, required a certain type of prominence from the linguistic variants. Based on the analysis of ten linguistic features, I argued that this prominence can be assimilated to perceptual and conceptual salience, the former being salience in terms of phonetic/graphic substance and the latter in terms of noticeability of the underlying grammatical (syntactic) rule.

I have argued that the formulaicity distribution of the features seems to be related to the cognitive prominence of those features (syntax-lexicon continuum and free/bound morphemes), so that the morphological features are explained by perceptual salience and the syntactic features by conceptual salience. According to this view, the domains of grammar that rank the highest on the syntax-lexicon continuum involve abstract and, thus, unnoticeable syntactic rules. The scribes, native speakers of a Romance-type variety, did not often recognise these due to lack of syntactically-informed education, hence the non-salient syntactic constructions' statistically non-significant distributions between formulaic and free parts.

By clarifying the role of salience, the study has identified one mechanism that makes it possible to examine spoken language-related features in conservative written genres. The salience approach can presumably be applied to assessing the status of language use in other historical treebanks as well, provided that they have been written by non-native speakers. All the same, it will be desirable to verify the validity of the here-delineated salience approach on a larger and more varied repertoire of linguistic features in a further study. In this way, it could be evaluated whether the *per se* efficient concept of perceptual and conceptual salience can still be reduced to other, even simpler linguistic motivations: whether there is a systematic association between perceptually and conceptually salient features and, for example, phonetically/phonologically motivated and semantically motivated language change, respectively.

The achieved results manifest the usability of treebanks for historical linguistics. Especially, philological annotation, such as that concerning the free/formulaic parts in LLCT, makes it possible to subject relatively well-known data sets to detailed quantitative analysis that would be unimaginable without treebanking. At the same time, this study exemplifies how essentially synchronic data can be used for diachronic research, due to the differing linguistic origin of the formulaic and free parts.

[10] Note that, before the stabilisation of the syntactically motivated (S)VO order, the Late Latin word order was also likely to be affected by the semantic realignment of grammatical relations, mentioned with the subject case encoding (Ledgeway 2012: 335–336, Korkiakangas 2016: 212–216).

References

Adams, James Noel. 2013. *Social variation and the Latin language*. Cambridge: Cambridge University Press.

Bamman, David, Marco Passarotti, Gregory Crane & Savine Raynaud. 2007. *Guidelines for the Syntactic Annotation of Latin Treebanks* (v. 1.3), http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf. (3 June, 2017.)

Bartoli Langeli, Attilio. 2006. *Notai: scrivere documenti nell'Italia medievale*. Roma: Viella.

Broccias, Cristiano. 2012. The syntax-lexicon continuum. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford Handbook of the History of English*, 735–747. Oxford: Oxford University Press.

Chiarcos, Christian, Berry Claus & Michael Grabski. 2011. Introduction: Salience in linguistics and beyond. In Christian Chiarcos, Berry Claus & Michael Grabski (eds.), *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, 1–28. Berlin: Gruyter.

Cintrón-Valentín, Myrna C. & Nick C. Ellis. 2016. Salience in Second Language Acquisition: Physical Form, Learner Attention, and Instructional Focus. *Frontiers in Psychology* 7. 1284.

Croft, William & Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Dulay, Heidi C. & Marina K. Burt. 1973. Should we teach children syntax? *Language Learning* 23. 245–258.

Goldschneider, Jennifer M. & Robert M. DeKeyser. 2001. Explaining the 'Natural Order of L2 Morpheme Acquisition' in English: A Meta-analysis of Multiple Determinants. *Language Learning* 51. 1–50.

Guyotjeannin, Olivier, Jacques Pycke & Benoît-Michel Tock. 1993. *Diplomatique médiévale*. Paris: Brepols.

Korkiakangas, Timo. 2016. *Subject Case in the Latin of Tuscan Charters of the 8th and 9th Centuries*. Helsinki: Societas Scientiarum Fennica.

Korkiakangas, Timo & Matti Lassila. 2013. Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material. In Francesco Mambrini, Marco Passarotti & Caroline Sporleder (eds.), *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities*, 61–72. Sofia: Bulgarian Academy of Sciences.

Korkiakangas, Timo & Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal of Language Technology and Computational Linguistics* 26. 103–114.

Lausberg, Heinrich. 1962. *Romanische Sprachwissenschaft.* Zweiter Teil: *Formenlehre*. Berlin: Gruyter.

Ledgeway, Adam. 2012. *From Latin to Romance: Morphosyntactic typology and change*. Oxford: Oxford University Press.

MacKenzie, Ian & Martin A. Kayman (eds.). 2018. *Formulaicity and Creativity in Language and Literature*. London: Routledge.

Maiden, Martin. 1996. On the Romance Inflectional Endings *-i* and *-e*. *Romance Philology* 50. 147–182.

Pienemann, Manfred. 1999. *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Sabatini, Francesco. 1965. Esigenze di realismo e dislocazione morfologica in testi preromanzi. *Rivista di Cultura Classica e Medievale* 7. 972–998.

Sairio, Anni & Minna Palander-Collin. 2012. The Reconstruction of Prestige Patterns in Language History. In Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, 626–638. Chichester: Blackwell.

Salvi, Giampaolo. 2011. Morphosyntactic persistence. In Adam Ledgeway, Martin Maiden & John C. Smith (eds.), *The Cambridge History of the Romance Languages.* Volume 1: *Structures*, 318–381. Cambridge: Cambridge University Press.

Schiaparelli, Luigi. 1933. Note diplomatiche sulle carte longobarde II: Tracce di antichi formulari nelle carte longobarde. *Archivio storico italiano* 19. 3–34.

Sornicola, Rosanna. 2012. Bilinguismo e diglossia dei territori bizantini e longobardi del Mezzogiorno: le testimonianze dei documenti del IX e X secolo. *Quaderni dell'Accademia Pontaniana* 59. 1–102.

Spevak, Olga. 2010. *Constituent Order in Classical Latin Prose*. Amsterdam: John Benjamins.

Väänänen, Veikko. 1981. *Introduction au latin vulgaire*. Paris: Éditions Klincksieck.

Valentini, Cecilia. 2017. *L'evoluzione della codifica del genitivo dal tipo sintetico al tipo analitico nelle carte del* Codice diplomatico longobardo. Firenze: Università degli Studi di Firenze dissertation.

Weber, Shirley Howard. 1924. *Anthimus, De Observatio[ne] Ciborum: Text, Commentary, and Glossary, with a Study of the Latinity*. Leiden: Late E.J. Brill.

Wright, Roger. 1991. The conceptual distinction between Latin and Romance: invention or evolution. In Roger Wright (ed.), *Latin and the Romance Languages in the Early Middle Ages*, 103–113. University Park: The Pennsylvania State University Press.

Zamboni, Alberto. 2000. *Alle origini dell'italiano: dinamiche e tipologie della transizione dal latino*. Roma: Carocci.

Zobl, Helmut & Juana Liceras. 1994. Functional Categories and Acquisition Orders. *Language Learning* 44. 169–180.

Appendix

The LLCT treebank and the queries relative to the ten examined features are available in the online Appendix.

Appendix: data and queries

The current version of the Late Latin Charter Treebank (LLCT) is available in .pml.xml format at the author's Zenodo repository: https://doi.org/10.5281/zenodo.1197357. LLCT consists of documents from the following three copyright-free editions: *Codice diplomatico longobardo* (*CDL*) 1–2 (Luigi Schiaparelli, 1929–1933), *Codice diplomatico toscano*, part 2, vol. 1 (Filippo Brunetti, 1833), *Memorie e documenti per servire all'istoria del Ducato di Lucca* (*MED*), part 5, vol. 2 (Domenico Barsocchini, 1837).

The queries are written in Prague Markup Language Tree Query (PML-TQ) language for the homonymous extension in TrEd Treebank Editor. LLCT is meant to be queried in TrEd with a customised version of the ALDT schema, which includes the LLCT-specific XML attributes, such as *seg* and *status*.

The formulaic/free distinction is encoded in LLCT with the *seg* (= segmentation) tag. The same attribute also indicates whether the tagged word is part of an autograph subscription (value *subs*). Since subscriptions are highly formulaic, they are here analysed together with formulaic words. The queries are written so that the output has to be copied to and manipulated in a spreadsheet software. The below table presents the total word counts of the formulaic/free/subscription parts of LLCT to which the numbers of occurrence are compared. Sentence count cannot be used as a reference point because the sentence boundaries of documentary Latin are opaque and the ways to define them differ between editions.

Table. The sizes of the textual segments in LLCT.

Segmentation	Number of words
formulaic	156,032
free	56,314
autograph subscription	13,488

The *status* attribute indicates whether the inflexional ending of the word in question contains an expansion of an abbreviation (*expan*), a restoration of damaged letters (*damage*), or neither of them (*normal*). Only *normal* words can be utilised for queries that concern morphology. For a detailed account, see Korkiakangas & Passarotti (2011).

Queries

1. Innovative lemmas

```
# 81 innovative lemmas with segmentation status
aldt-sentence
[ descendant aldt-word $a :=
    [ lemma in {"aldia1", "aldiaricus1", "aldio1", "aldionalis1", "arimannus1", "arra1", "banda1", "barba3",
"barbanus1", "batto1", "bluto1", "bullitanus1", "calderarius1", "cambiator1", "cambium1", "caminata1",
"canavarius1", "cavallarius1", "cavallicatura1", "cergiolitum1", "caesa1", "concambiatio1", "concambio1",
"concambium1", "debluto1", "fiuwadia1", "focacia1", "fossata1", "fossatum1", "fumarius1", "gahagium1",
"gasindus1", "gastaldus1", "grunda1", "launechild1", "mallo1", "marepas1", "marscalcus1", "monto1",
"morgingabum1", "mustariolum1", "ornile1", "patrinius1", "paupertacula1", "petia1", "petiola1", "petiolum1",
"petium1", "rasula1", "recta1", "rectum1", "scafilus1", "scherpha1", "scufia1", "sculdahis1", "spanga1", "sporus1",
"staffilus1", "stanclio1", "stantarium1", "strata1", "summarra1", "sundrialis1", "sundrium1", "tessero1", "tia1",
"tingatio1", "torta1", "ubiscarius1", "usitile1", "wadia1", "wadio1", "waldemannus1", "waldus1", "vassus1",
"weregeldum1", "viccia1", "viganatio1", "viganio1", "vitellata1", "zapa1"} ] ];
  >> give $a.lemma,$a.seg
```

2. Future perfect form

```
# future perfect form
aldt-sentence
[ descendant aldt-word $a :=
    [ status = "normal", voice = "active", pos = "verb", tense = "future_perfect" ] ];
 >> give$a.form,$a.seg
```

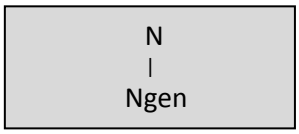
3. Dative plural form

```
# dative plural form (3rd, 4th, and 5th declension)
aldt-sentence
[  aldt-word
    [ descendant aldt-word $a :=
        [ status = "normal", number = "plural", case = "dative", declension in {"3", "4", "5"}, (pos = "noun" or pos =
"adjective") ] ] ];
 >> give $a.form,$a.seg
```

4. Genitive vs. *de* PP

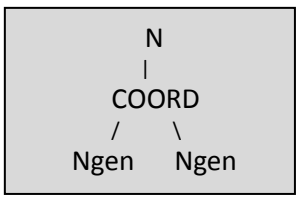
The genitives and the *de* PPs are among the hardest-to-reach items in the present study. Ten queries are needed to cover all the relevant dependency structures, and after that manual disambiguation of the output is required. The figures illustrate the queried dependency structures.

4.1 Genitives of type N-Ngen



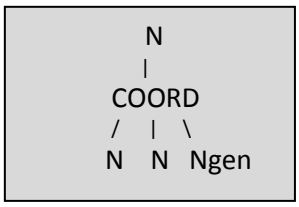

```
# N-Ngen: non-coordinated adnominal genitive (the head not headed by another nominal)
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ (pos = "noun" or pos = "pronoun"),
       0x parent aldt-word
       [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ],
       0x parent aldt-word
       [ pos = "conjunction",
          parent aldt-word
          [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ] ],
        aldt-word $c :=
       [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun") ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$b.form,$c.form,$b.lemma,$c.lemma,$c.seg
```

4.2 Genitives of type N-coord-Ngen



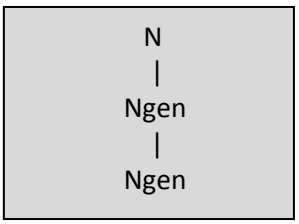
```
# N-coord-Ngen: coordinated adnominal genitive (the head not headed by another nominal)
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ (pos = "noun" or pos = "pronoun"),
        0x parent aldt-word
        [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ],
        0x parent aldt-word
        [ pos = "conjunction",
            parent aldt-word
            [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ] ],
         aldt-word $d :=
        [ pos = "conjunction",
             aldt-word $c :=
            [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun") ] ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$b.form,$c.form,$b.lemma,$c.lemma,$c.seg
```

4.3 Genitives of type N-coord-N/Ngen




```
# N-coord-N/Ngen: adnominal genitive modifying coordinated Ns
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ relation in {"COORD", "APOS", "COORD_AP"},
         aldt-word $d :=
        [ (pos = "noun" or pos = "pronoun"), (relation ~ "CO$" or relation ~ "AP$") ],
         aldt-word $c :=
        [ status = "normal", case = "genitive", !relation ~ "CO$", !relation ~ "AP$", !lemma = "is1", !lemma = "ipse1",
!lemma = "qui1", (pos = "noun" or pos = "pronoun") ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$d.form,$c.form,$d.lemma,$c.lemma,$c.id,$c.seg
```

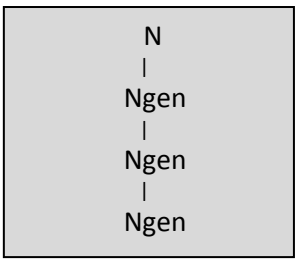
4.4 Genitives of type N-Ngen-Ngen




```
# N-Ngen-Ngen: non-coordinated adnominal genitive with the head headed by another nominal (two-genitive chain
of type rector ecclesie Stefani)
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ (pos = "noun" or pos = "pronoun"),
        0x parent aldt-word
        [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ],
        0x parent aldt-word
```

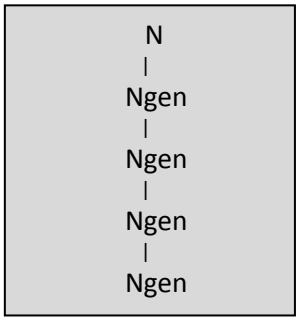
```
    [ pos = "conjunction",
        parent aldt-word
        [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ] ],
     aldt-word $c :=
    [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun"),
         aldt-word $d :=
        [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun") ] ] ] ];
>> give $a.document_id,$a.subdoc,$a.id,$c.form,$d.form,$c.lemma,$d.lemma,$d.seg
```

4.5 Genitives of type N-Ngen-Ngen-Ngen




```
# N-Ngen-Ngen-Ngen: non-coordinated adnominal genitive with the head headed by another nominal (three-genitive chain of type manus rectoris ecclesie Stefani)
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ (pos = "noun" or pos = "pronoun"),
        0x parent aldt-word
        [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ],
        0x parent aldt-word
        [ pos = "conjunction",
            parent aldt-word
            [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ] ],
         aldt-word $c :=
        [ case = "genitive", (pos = "noun" or pos = "pronoun"),
             aldt-word $d :=
            [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun"),
                 aldt-word $e :=
                [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun") ] ] ] ] ];
>> give $a.document_id,$a.subdoc,$a.id,$d.form,$e.form,$d.lemma,$e.lemma,$e.seg
```

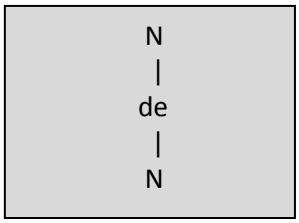
4.6 Genitives of type N-Ngen-Ngen-Ngen-Ngen




```
# N-Ngen-Ngen-Ngen-Ngen: non-coordinated adnominal genitive with the head headed by another nominal (four-genitive chain)
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ (pos = "noun" or pos = "pronoun"),
        0x parent aldt-word
```

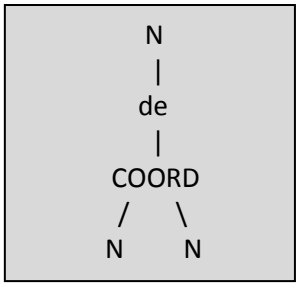
```
      [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ],
      0x parent aldt-word
      [ pos = "conjunction",
        parent aldt-word
        [ (pos = "noun" or pos = "adjective" or pos = "pronoun") ] ],
       aldt-word $c :=
      [ case = "genitive", (pos = "noun" or pos = "pronoun"),
         aldt-word $d :=
        [ case = "genitive", (pos = "noun" or pos = "pronoun"),
           aldt-word $e :=
          [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun"),
             aldt-word $f :=
            [ status = "normal", case = "genitive", (pos = "noun" or pos = "pronoun") ] ] ] ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$e.form,$f.form,$e.lemma,$f.lemma,$f.seg
```

4.7 PPs of type N-de-N




```
# PPs of type N-de-N
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ status = "normal", (pos = "noun" or pos = "pronoun" or pos = "adjective"),
       aldt-word $c :=
      [ lemma = "de1",
         aldt-word $d :=
        [ (pos = "noun" or pos = "pronoun" or pos = "adjective") ] ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$b.form,$c.form,$d.form,$b.lemma,$d.lemma,$b.seg
```

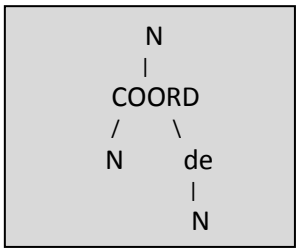
4.8 PPs of type N-de-coord-N




```
# PPs of type N-de-coord-N: de PPs with coordinated complements
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ status = "normal", (pos = "noun" or pos = "pronoun" or pos = "adjective"),
       aldt-word $c :=
      [ lemma = "de1",
         aldt-word $e :=
        [ pos = "conjunction",
           aldt-word $d :=
          [ (pos = "noun" or pos = "pronoun" or pos = "adjective") ] ] ] ] ];
```

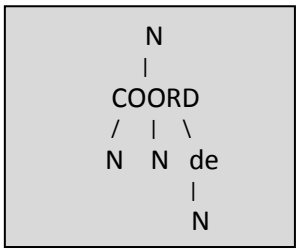
```
>> give $a.document_id,$a.subdoc,$a.id,$b.form,$c.form,$d.form,$b.lemma,$d.lemma,$b.seg
```

4.9 PPs of type N-coord-de-N




```
# N-coord-de-N: de PPs coordinated with genitives
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ status = "normal", (pos = "noun" or pos = "pronoun" or pos = "adjective"),
        aldt-word $e :=
       [ pos = "conjunction",
            aldt-word $c :=
           [ lemma = "de1",
                aldt-word $d :=
              [ (pos = "noun" or pos = "pronoun" or pos = "adjective") ] ] ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$b.form,$c.form,$d.form,$b.lemma,$d.lemma,$b.seg
```

4.10 PPs of type N-coord-N/de-N




```
# N-coord-N/de-N: de PPs modifying coordinated Ns
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ relation in {"COORD", "APOS", "COORD_AP"},
        aldt-word $e :=
       [ lemma = "de1",
            aldt-word $c :=
          [ status = "normal", !relation ~ "CO$", !relation ~ "AP$", (pos = "noun" or pos = "pronoun") ] ],
        aldt-word $d :=
       [ (pos = "noun" or pos = "pronoun"), (relation ~ "CO$" or relation ~ "AP$") ] ] ];
 >> give $a.document_id,$a.subdoc,$a.id,$d.form,$e.form,$c.form,$d.lemma,$c.lemma,$c.seg,$c.id
```

Post-processing: The queries provide the sentence id numbers, so that the output sentences can be analysed in their full-text context when post-processing them. Since the annotation style of LLCT does not allow distinguishing between a genuine genetive construction (e.g. *manus presbiteri* 'the priest's hand') and an attribute chain (e.g. *Teutperti presbiteri* 'of the priest Teutpertus', where both the genitives modify a common head, such as *signum* 'cross'), the phrases of this type have to be disambiguated manually. A large group to be disambiguated are the NPs with *ipse*, *ille*, and *hic*, where the pronoun can be either pronominal or adjectival: *ecclesie ipsius* 'of his church' is a genuine possessive construction while *ecclesie ipsius* 'of that church' is not.

5. Accusative and infinitive vs. complement clauses

5.1 Accusative and infinitive with non-coordinated subject

```
# ACI with non-coordinated subject
aldt-sentence
[ descendant aldt-word $a :=
    [ status = "normal", mood = "infinitive", pos = "verb",
        aldt-word $b :=
       [ relation = "SBJ" ] ] ];
 >> give $a.form,$b.form,$b.seg
```

5.2 Accusative and infinitive with coordinated subject

```
# ACI with coordinated subjects
aldt-sentence
[ descendant aldt-word $a :=
    [ status = "normal", mood = "infinitive", pos = "verb",
        aldt-word
       [ (relation = "COORD" or relation = "APOS"),
            aldt-word $b :=
           [ (relation = "SBJ_CO" or relation = "SBJ_AP") ] ] ] ];
 >> give $a.form,$b.form,$b.seg
```

5.3 Phrasal complementation

```
# clausal object complements with conjunction (instead of ACI)
aldt-sentence
[ descendant aldt-word $a :=
    [ pos = "verb", !mood = "infinitive",
        aldt-word $b :=
       [ relation = "AuxC", lemma != "si1",
            aldt-word $c :=
           [ pos = "verb", relation = "OBJ", !mood = "infinitive" ] ] ] ];
 >> give $a.lemma,$a.form,$b.form,$c.form,$b.seg
```

Post-processing: The complementiser clauses were disambiguated in their full-text context to isolate subordinate interrogative clauses and comparative clauses, which cannot be excluded on the basis of the treebank annotation. Six cases were discarded (facias comodo placueret (CDL 51), cogitare qualiter redimere (CDL 171), iscimus qualiter occhurra (CDL 230), iscio qualiter hoccurra (CDT 12), considerastis qualiter perveniret (MED 523), dicimus quod fiat (CDL 68)).

6. Absolute constructions

6.1 Absolute constructions with non-coordinated subject

```
# absolute constructions with non-coordinated subject
aldt-sentence
[ descendant aldt-word $a :=
    [ pos = "participle", !mood = "gerundive", (relation = "ADV" or relation = "ADV_CO" or relation = "ADV_AP"),
       0x  aldt-word
       [ relation = "AuxV" ],
        aldt-word $b :=
       [ relation = "SBJ" ] ] ];
```

```
 >> give $a.form,$b.form,$a.seg
```

6.2 Absolute constructions with coordinated subjects

```
# absolute constructions with coordinated subjects
aldt-sentence
[ descendant aldt-word $a :=
    [ pos = "participle", !mood = "gerundive", (relation = "ADV" or relation = "ADV_CO" or relation = "ADV_AP"),
        0x  aldt-word
        [ relation = "AuxV" ],
         aldt-word
        [ (relation = "COORD" or relation = "APOS"),
             aldt-word $b :=
            [ (relation = "SBJ_CO" or relation = "SBJ_AP") ] ] ] ];
 >> give $a.form,$b.form,$a.seg
```

7. Second-person singular form

7.1 Second-person singular form in -s

```
# active second person singular -s, non-perfect
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ status = "normal", person = "second_person", tense != "perfect", number = "singular", voice = "active", mood !=
"imperative", form ~ "s$" ] ];
 >> give $a.document_id,$a.subdoc,$a.date,$b.form,$b.lemma,$b.seg
```

7.1 Second-person singular form not ending in -s

```
# active second person singular non-s, non-perfect
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ status = "normal", person = "second_person", tense != "perfect", number = "singular", voice = "active", mood !=
"imperative", form !~ "s$" ] ];
 >> give $a.document_id,$a.subdoc,$a.date,$b.form,$b.lemma,$b.seg
```

8. Dative singular form

```
# dative singular forms
aldt-sentence
[ descendant aldt-word $a :=
    [ status = "normal", case = "dative", number = "singular", !declension = "0" ] ];
 >> give$a.form,$a.seg
```

9. Subject case encoding

```
# non-coordinated non-pronominal imparisyllabic 3rd-declension singular subjects
aldt-sentence $a :=
[ descendant aldt-word
    [ mood !~ "infinitive", !(pos = "participle" and tense = "present"), !(pos = "participle" and case = "ablative"), (pos =
"verb" or pos = "participle"),
        aldt-word $b :=
```

```
        [ status = "normal", declension = "3", number = "singular", relation ~ "^SBJ", gender != "neuter", pos !=
"pronoun", animacy in {"-", "0"} ] ] ];
  >> give $a.document_id,$a.subdoc,$a.date,$b.form,$b.lemma,$b.seg,$b.case

# coordinated non-pronominal imparisyllabic 3rd-declension singular subjects
aldt-sentence $a :=
[ descendant aldt-word
    [ mood !~ "infinitive", !(pos = "participle" and tense = "present"), !(pos = "participle" and case = "ablative"), (pos =
"verb" or pos = "participle"),
         aldt-word
        [ (relation = "COORD" or relation = "APOS"),
             aldt-word $b :=
            [ status = "normal", declension = "3", number = "singular", relation ~ "^SBJ", gender != "neuter", pos !=
"pronoun", animacy in {"-", "0"} ] ] ] ];
  >> give $a.document_id,$a.subdoc,$a.date,$b.form,$b.lemma,$b.seg,$b.case
```

10. Verb/object order

10.1 Verb/object order, formulaic parts

```
# VO/OV order with finite V, formulaic parts, only non-coordinated arguments
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ mood !~ "infinitive", !(pos = "participle" and tense = "present"), !(pos = "participle" and case = "ablative"), (pos =
"verb" or pos = "participle"),
         aldt-word $d :=
        [ relation = "OBJ", iobj != "1", seg = "formulaic", (pos = "noun" or pos = "adjective") ] ] ];
  >> give distinct $a.id,$b.form,$d.form,$b.id,$d.id
```

10.2 Verb/object order, free parts

```
# VO/OV order with finite V, free parts, only non-coordinated arguments
aldt-sentence $a :=
[ descendant aldt-word $b :=
    [ mood !~ "infinitive", !(pos = "participle" and tense = "present"), !(pos = "participle" and case = "ablative"), (pos =
"verb" or pos = "participle"),
         aldt-word $d :=
        [ relation = "OBJ", iobj != "1", seg = "free", (pos = "noun" or pos = "adjective") ] ] ];
  >> give distinct $a.id,$b.form,$d.form,$b.id,$d.id
```

Post-processing: The query outputs the sentence identifier numbers of V and O which indicate the linear position in the sentence. The relative order of the arguments must be defined on the basis of these numbers.