

Visualizing linguistic variation in a network of Latin documents and scribes

Timo Korkiakangas, Matti Lassila

▶ **To cite this version:**

Timo Korkiakangas, Matti Lassila. Visualizing linguistic variation in a network of Latin documents and scribes. Journal of Data Mining and Digital Humanities, 2018, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, 10.46298/jdmdh.4472 . hal-01645124v2

HAL Id: hal-01645124
https://inria.hal.science/hal-01645124v2

Submitted on 24 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Journal of Data Mining and Digital Humanities

Visualizing linguistic variation in a network of Latin documents and scribes

Timo Korkiakangas[1*], Matti Lassila[2]

1 University of Oslo, Norway

2 Open Science Centre, University of Jyväskylä, Finland

*Corresponding author: Timo Korkiakangas timo.t.korkiakangas[ät]gmail.com

Abstract
This article explores whether and how network visualization can benefit philological and historical-linguistic study. This is illustrated with a corpus-based investigation of scribes' language use in a lemmatized and morphologically annotated corpus of documentary Latin (Late Latin Charter Treebank, LLCT2). We extract four continuous linguistic variables from LLCT2 and utilize a gradient colour palette in Gephi to visualize the variable values as node attributes in a trimodal network which consists of the documents, writers, and writing locations underlying the same corpus. We call this network the "LLCT2 network". The geographical coordinates of the location nodes form an approximate map, which allows for drawing geographical conclusions. The linguistic variables are examined both separately and as a sum variable, and the visualizations presented as static images and as interactive Sigma.js visualizations. The variables represent different domains of language competence of scribes who learnt written Latin practically as a second language. The results show that the network visualization of linguistic features helps in observing patterns which support linguistic-philological argumentation and which risk passing unnoticed with traditional methods. However, the approach is subject to the same limitations as all visualization techniques: the human eye can only perceive a certain, relatively small amount of information at a time.

keywords
network visualization; Latin linguistics; Early Middle Ages; philology

INTRODUCTION[1]

The objective of this article is to investigate whether and how network visualization can benefit philology and historical linguistics. This will be implemented by examining early medieval Latin scribes' language competences in terms of how the scribes mastered four linguistic features which reflect language change that took place in Late Latin. The four linguistic features, i.e. spelling correctness, classical prepositions, genitive plural form, and <*ae*> diphthong, are extracted corpus-linguistically from the Late Latin Charter Treebank (version 2, LLCT2), which also constitutes the network. Our primary interest is to find out whether the network visualization approach has demonstrable advantages compared to ordinary cross-tabulations as far as support to philological and historical-linguistic argumentation is concerned. As a necessary part of this enterprise, we seek to clarify the scientific premises of network visualization: to be utilized for research purposes, network visualization must be objective and replicable.

I BACKGROUND AND MOTIVATION OF THE OBJECTIVES

[1] Writing this article has been teamwork. Korkiakangas answers for the linguistic-philological substance and for the research setting in general. Lassila is responsible for the questions related to network visualization and its technical realization.

Networks are nowadays present everywhere where complex data sets are illustrated. In linguistics, network analysis has also found certain fields of application, albeit not so much in visualization itself. In spite of the popularity of the network approaches, network presentations often perplex audiences with their seemingly haphazard way of visualizing different aspects of a single network. Apparently, the nodes are often coloured and sized only in order to reach a maximal visual effect. This kind of utilitarian practice leads to a situation where the various visualizations of the same data are scarcely comparable. This may not be problematic in rough exploratory studies, but rigorous scientific research can only rely on network visualization to the degree it is grounded in objective, scrupulously defined principles which, at least in theory, also make it replicable. The present article seeks to clarify the premises of combining visual effectiveness and scientific rigour. The central practical method to be discussed is the use of gradient colour to represent node attribute values.

Networks are of various kinds. [Araújo and Banisch, 2016] emphasize that the linguistic networks proper, i.e. those induced directly from textual data sets themselves, are usually far more abstract than the so-called social networks analyzed in sociolinguistics ([Bergs, 2005]). The nodes of social networks represent, for example, people and the edges between the nodes, for example, their family ties or acquaintanceships. However, networks of diverse systems like social networks, neural networks, or linguistic networks expose similar attributions showing that they have similar organizing principles inside ([Barzel and Barabási, 2013], [Jackson, 2008]). Yet, linguistic networks in the sense used by Araújo and Banisch are only one type of networks related to language and linguistics. These linguistic text networks proper are constituted of nodes which are grammatical items, such as word forms, lemmas, or sentences, linked to each other by their co-occurrence or adjacency in a text. Figure 1 shows a simple linguistic network of the lemma *peccatum* 'sin' (24 occurrences) in the first book of St. Augustine's *De civitate Dei*, The City of God. The network was created by using the Linguistic Networks online tool at [http1]. The nodes orbiting *peccatum* are collocate lemmas which co-occur with *peccatum*. Perhaps not surprisingly, these include concepts such as *mundus* 'world', *iniquitas* 'iniquity', and *voluptas* 'pleasure'. Entire large text corpora form much larger and more complex networks (e.g. [Passarotti, 2014]).

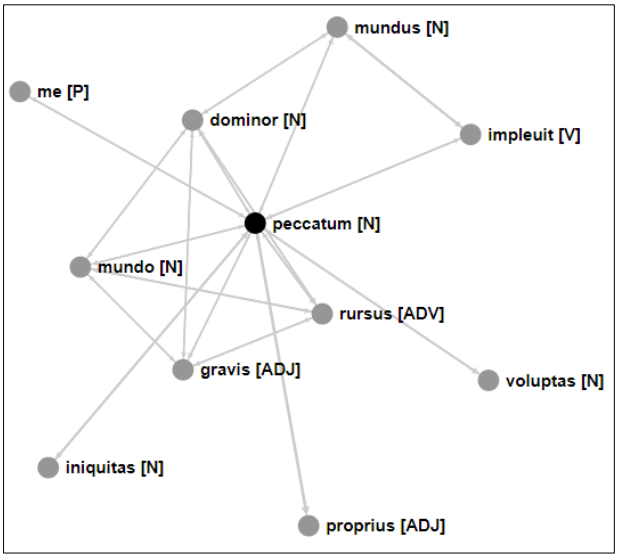


Figure 1. Network of lemmas co-occurring with the lemma *peccatum* 'sin' in the first book of St. Augustine's *De civitate Dei*. The pos tagging is partly defective: *dominor* 'to dominate' should be a verb [V] and an ablative form *mundo* has failed to map with the lemma *mundus* 'world'.

In this article, we will reserve the term "linguistic network" to this kind of text networks which are directly induced or synthesized from a textual data set and represent abstract relations between linguistic units. Our objective will be, instead, the visualization of external linguistic variables as the attributes of the nodes in a social network of documents, scribes, and locations. On linguistic networks, such as the lemma network of Figure 1, various network metrics, including betweenness centrality, closeness, diameter, and clustering coefficient, can be calculated. These are basic concepts of network theory, a subset of graph theory ([Ferrer i Cancho, 2008]). For example, [Passarotti, 2014] discusses the role of the most frequent lemma (*sum* 'to be') in a network built on a large Latin treebank. He focuses on betweenness centrality, closeness centrality, average edge/node degree, clustering coefficient, and average path length. However, [Araújo and Banisch, 2016] conclude that, at least in its current state, the application of available statistical network analytics to linguistic networks can only contribute to the development of linguistic theory in a limited manner. This is because the higher is the abstraction level of network induction, the lower is the intelligibility of the statistical indicators.

This article does not seek to calculate network metrics on a linguistic network induced from documentary Latin. Instead, we will visualize distributions of linguistic variables that do not arise from the LLCT2 network itself, but are derived externally from the text corpus that underlies the network. The linguistic features are extracted by corpus-linguistic methods which exploit the lemmatic and morphological annotation layers of LLCT2. The features represent the fields of orthography (spelling correctness and <*ae*> diphthong), lexicon (classical prepositions), and morphology (genitive plural form) and are meant to measure together the scribes' general language competence. The syntactic layer of the here examined corpus is still partly under construction. Hopefully, in a near future, even syntactic variables can be introduced in the analysis, which will expand the possibilities and the generalizability of the results based on the visualizations. The visualizations will be implemented in Gephi, a widely used open-source software package for network analysis and visualization [http2].

The linguistic variables which will be visualized are four. Section II will present the data, section III the nodes, and section IV the linguistic features. Section V will discuss the theoretical and statistical prerequisites for the visualization. In section VI, each linguistic variable will be first visualized separately and then merged into a sum variable. Section VII will present two case studies to deal more extensively with the philological and historical-linguistic import of the visualizations. Conclusion will close the article by discussing the relative merits of the experimented methods.

II DATA

The data utilized in this article is the lemmatized and morphologically tagged Late Latin Charter Treebank (v. 2, LLCT2), which consists of 1,040 early medieval Latin documentary texts (c. 480,000 words) written between AD 714 and 897. The documents were written in historical Tuscia, which mostly corresponds to modern Tuscany in Italy. The documents are sale or purchase contracts or donations, accompanied by a few judgements as well as some humble lists and memoranda. LLCT2 is still under construction, and only the first half of it is already provided with the syntactically annotated layer, thus making it a treebank (i.e. LLCT, version 1, see [Korkiakangas and Passarotti, 2011] and [Korkiakangas and Lassila, 2013]). The lemmatization as well as the morphological and syntactic annotation styles are based on the style of the Ancient Greek and Latin Dependency Treebank (AGLDT), which can be deduced from the *Guidelines for the Syntactic Annotation of Latin Treebanks* [Bamman et al.,

2007]. [Korkiakangas and Passarotti, 2011] define a number of additions and modifications to these general guidelines which are designed for Classical Latin. The LLCT2 documentary texts come from relatively old copyright-free editions: [Schiaparelli, 1929–1933] (CDL), [Brunetti, 1833] (CDT), [Barsocchini, 1837] (MED), [Bertini, 1836] (MED2), [Barsocchini, 1841] (MED3). Therefore, we decided to check the texts against the modern *Chartae Latinae Antiquiores* editions [ChLA]. For a more detailed description of LLCT2 and the underlying text editions, see [Korkiakangas, 2017].

The LLCT2 documents are part of a unique patrimony of texts written in a variety of Latin with considerable orthographical and grammatical oscillation affected by the spoken Romance-type vernacular of the time. The spoken language had evolved far from Classical Latin, which was still considered the grammatical ideal, judging, for example, from the most classically written LLCT2 documents. Many documents abound, however, with features that derive from the spoken idiom, such as Romance-type prepositions or spellings motivated by phonological change. The here utilized four linguistic variables are meant to measure the linguistic competence understood as an adherence to the Classical Latin grammar and spelling, as will be explained in section IV. The linguistic variation together with the fact that the scribes' names, writing locations, and writing years are almost always known makes the corpus particularly suited both for linguistic study and network approach.

III NODES

The LLCT2 network is relatively small. Its 1,326 nodes represent documents, scribes, and locations. The nodes are connected to each other by unweighted edges. The definition of the document nodes is straightforward since each of the 1,040 LLCT2 documents forms a node. The document nodes are labelled with a unique identifier that consists of the acronym of the original edition (CDL, CDT, MED, ChLA; see above and References) plus the document's number within that edition. In most cases, the definition of the 220 scribe nodes is equally straightforward, given that the scribes scrupulously signed what they wrote, with the exception of eight documents. The writers of these eight documents are, nevertheless, included in the network as Anonymous 1, Anonymous 2, etc., so that each document node is connected to one scribe node. The attributes of the scribe nodes also indicate scribal status, i.e. whether the scribe was lay or ecclesiastical.

A bigger challenge for the definition of the scribe nodes is, however, the fact that there are several scribes with the same name and that some scribes wrote their own names in several ways. The ChLA editors have disambiguated the scribes by way of paleographical analysis of their handwriting and by diachronic inferences about the scribes' periods of activity: two documents signed by a scribe with the same name are not likely to have been written by the same person if they are separated by one hundred years. For example, there seem to be eight scribes called Petrus in the LLCT2 corpus. These are encoded as "Petrus I", "Petrus II", etc., following the ChLA convention. Likewise, there are documents written clearly by one and the same hand, but with three (slightly) differing signatures. Consequently, these documents are grouped under a joint scribe label, such as "Ratfusu, Ratfonso, Ratfonsus I", again following the ChLA convention.

The 66 location nodes constitute the most difficult task. 79% of the documents have been written in the city of Lucca. However, the definition of several location nodes required a time-consuming disambiguation and occasional merging of small localities of which little is known with better-known, usually larger, localities. The Repetti online database of Tuscan toponyms

[http3] was of great help in disambiguating the locations. Also, the geographical coordinates of the locations are derived from the Repetti database whenever available. However, the Repetti online does not provide some small localities with coordinate data. The geographical coordinates are utilized to settle the location nodes on the graph background in the way that they create an approximate map of Tuscia (see the figures in sections VI and VII). The node labels will not be displayed in the graphs of this article but can be consulted in the interactive Sigma.js visualizations [http4]. The map of Figure 2 presents some important centres of historical Tuscia that will appear in the discussion of sections VI and VII.



Figure 2. Some writing places of LLCT2 documents.

The writing location is not mentioned at all in six documents and the location mentioned in some others is ambiguous because it is too superficial (e.g. *ad ecclesia sancti Iuliani* "in the church of St. Julian") or its exact location is unknown (e.g. Laucinianu). Yet, it is often possible to deduce the location quite reliably on the basis of various clues and other place names occurring in the document. A scrutiny of the contents and the scribe revealed that the above-mentioned church must have been near Monte Pisano and that Laucinianu is close to Monte Amiata. We mark some particularly uncertain localizations with question mark after the location name.

As said, we sometimes merge infrequent minor localities with adjacent major localities in order to multiply location nodes unnecessarily. This merging of location names that, in practice, denote the same location or a location which is close to or part of a certain other location has been the most demanding challenge. For example, the suburban villages of the city of Lucca, such as Capannori and Lunata, are not far away from Lucca, but given that they occur as writing locations in several documents, we decided to create the respective location nodes, but with the labels showing the close-by city name in brackets: "Capannori (Lucca)" and "Lunata (Lucca)". Instead, Laucinianu, where only one document has been written, was subsumed under the Monte Amiata node. Thus, we define the location nodes by following a system based on how often the locations occur within the corpus. Small locations that only one or two documents come from are merged into the adjacent bigger location. It is sometimes unclear on which close-by centre a certain small location is dependent. In these cases, the Repetti database often provides information on which diocese the location belonged

to. For example, Vaccoli belongs to the sphere of influence of Lucca instead of Pisa and receives, thus, the label "Vaccoli (Lucca)". All this said, it is obvious that as far as the location nodes are concerned, the network is far from accurate. On the other hand, it must be remembered that Tuscia is relatively compact and scribes sometimes seem to have written documents in several locations, so the attribution of a document to a certain location does not necessarily reflect the documentary tradition of that location.

For the geographical positioning of the nodes, we utilize the Geo Layout plugin in Gephi. Those locations that are given geographical coordinates in the Repetti online occupy the position defined by those coordinates on the graph surface. Those generally more peripheral location nodes whose coordinates are not available in the Repetti online are provided with the coordinates of the nearest adjacent location with known coordinates. We find this procedure tolerable since the main idea of the map layout is to illustrate the adjacency of places, not so much their absolute positions. In order to make the graph compact, we dragged the twenty southernmost location nodes with a latitude coordinate below Capannoli (WGS 1984 43.58431) closer to Lucca by reducing their location coordinates by 40% while the rest of the map remains as it is. This procedure does not distort the graph projection but makes it more easily readable, given that the map scale keeps modest while the plethora of nodes around Lucca remain perceptible. The graph is created by applying first the Geo Layout algorithm to establish the location nodes. Subsequently, the ForceAtlas 2 layout algorithm is run with the Dissuade Hubs box checked and with Scaling 7.0.

IV LINGUISTIC FEATURES

The four linguistic features to be visualized are spelling correctness, classical prepositions, genitive plural form, and <*ae*> diphthong. The choice of the features is motivated by the language change that took place in late Latin and early medieval Latin. The Latin language was evolving all the time, as every language, while the most striking developments that finally led to the emergence of Romance are assumed to have accelerated roughly from the 4th century onwards. The language change led to a situation where the early medieval writers of Latin had to learn the written standard practically as a second language. Indeed, Latin continued to maintain a huge prestige as the language of law and juridical practice until the full Romance-writing Middle Ages. The LLCT2 scribes certainly thought they were writing the same language they spoke, i.e. Latin, although the distance between the spoken form and the written standard had grown considerable. The metalinguistic shift and the subsequent emergence of the Romance writing systems took place only later ([Wright, 1991]).

The mentioned four features were selected because they represent different aspects of language use: spelling, lexicon, and morphology. Genitive and preposition use are, of course, related to syntax as well, insofar as the genitive case was being gradually replaced by prepositional constructions. Since prepositional constructions gained ground at the expense of several inflexions, even the new, Romance-type prepositions, which originated from the spoken language, became more frequent ([Adams, 2013: 257 ff.]). As regards spelling, it is also influenced by various domains of grammar: phonology, morphology, and even syntax. With the exception of pure typos and incidental distractions of the scribe, misspellings can be motivated by increasing phonological distance between oral use and written standard ([Adams, 2013: 32–36], [Rutkowska and Rössler, 2012]). Likewise, some spellings may be due to specific local traditions of which little is known. The matter is further complicated by the fact that the inflexional system of late Latin syntax was undergoing a profound reorganization. In cases where the concomitant phonological change affected inflexional

morphemes, which carried syntactic information, it is often impossible to decide whether the misspellings are motivated by phonology, morphology, or syntax. In any case, the four linguistic features that will be visualized in this article can be claimed to cover various aspects of the Latin language.

The linguistic features are operationalized as variables which quantify the variation of those features in LLCT2. This quantification is based on the numerical output of corpus queries which extract from LLCT2 all constructions or forms that meet certain criteria. The variables indicate the relative frequency of the examined feature in each document, scribe, and writing location. The percentages of scribes and writing locations are calculated by counting the occurrences within all the documents written by a certain scribe or in a certain location, respectively. The same way of operationalization is followed in [Korkiakangas, 2017]. The thus acquired linguistic variables are subsequently encoded in the network data as attributes of the nodes.

The **spelling correctness** variable indicates the percentage of characters which are spelled according to the Classical Latin spelling in relation to all the characters of a document. For example, the word form *atmodo* differs from the classical standard form *admodum* "greatly" by three characters. The correct characters are four and, thus, the spelling correctness percentage of the form *atmodo* is 57 (i.e. 4 in 7). This percentage is then defined for documents, scribes, and writing locations by calculating the proportion of the non-classically and classically spelled characters that are present in each document, scribe, and location.

Technically, the number of misspelled characters is obtained by calculating the Levenshtein edit distance, i.e. the number of single-character changes, between each word attested in LLCT2 and the normalized, standard version of that word. The normalized forms were produced by matching each lemma/morphological tag pair, extracted from the LLCT2 annotation, with the entries of a two-million-line Classical Latin word-form library. This word-form library was created by feeding the Open Office Latin spelling lexicon [http5] into the Whitaker's Words (Words 1.97FC) tagger [http6]. For a detailed description of the procedure, see [Korkiakangas, 2017]. Any calculation based on simple edit distances is necessarily a coarse estimate of the linguistic reality underlying the attested spellings. A more sophisticated way would be to distinguish different types of spelling variants by annotating them as phonetic, morphological, syntactic, and other misspellings. However, even this would be a rough attempt because the motivations of spelling variation are often impossible to isolate, as stated above. For a broader discussion on this issue, see [Korkiakangas, 2017].

The often-quoted word "standard" requires a clarification. Latin was never bound by a standard comparable to the standards and norms that are imposed and preserved by the authorities in most modern societies. Yet, the best-written documentary texts of LLCT2 still essentially followed the orthography and morphology of the Classical Latin, as cultivated by ecclesiastical authors of the Late Antiquity. Given this kind of consensus about correct spelling and grammar obtained, which was also codified by grammarians to some extent, we deem it justified to speak here of standard spelling and standard grammar ([Korkiakangas, 2016: 36], [Lazard, 1993: 391]).

In late Latin, some traditionally used **classical prepositions** were replaced by new innovative, Romance-type formations which were often originally adverbs. At the same time, both old and new prepositions also increasingly substituted simple case inflexions. We count the number of twenty innovative prepositions in each document and compare it to the number of

classical prepositions in the same document. The result is a variable which gives the percentage of the classical prepositions in all the prepositions of a document. The twenty innovative prepositions are *anteposito*, *da*, *desub*, *desuper*, *excepto*, *fine*/*fini*, *foras*, *foris*, *hactenus*, *inante*, *insimul*, *insuper*, *intro*, *longinquo*, *recta*, *rectum*, *retro*, *sequenter*, *subtus*, and *usque*. They form 2.3% of the 43,406 prepositions in LLCT2. Some of these prepositions were likely to be only temporary formations. No overview study on innovative late Latin prepositions has been yet written. See, however, [Adams, 2013: 582–593], [Rovai, 2013], and [Korkiakangas, 2017].

The Latin genitive case was gradually replaced in late Latin by the prepositional construction with *de*, a development which finally led into complete disappearance of the genitive case in Romance ([Valentini, 2017]). The decline of the **genitive plural** form (*-orum*/*-arum*/*-um*) is likely to have begun earlier than that of the genitive singular forms, which still appear to be syntactically functional even in early medieval Latin, perhaps because they were supported by the resembling dative singular morphology in certain declensions. The decline of the genitive plural form is witnessed by its hypercorrect use in some Late Latin texts as well as by its restriction to certain fixed expressions ([Sornicola, 2012: 59]).

Documentary texts are a good habitat for possessive constructions, natural loci of the genitive case, because they deal with possession and the transfer thereof. The great majority of the LLCT2 documents contains phrases, such as *filiis filiorum meorum* "to the sons of my sons" (genitive) and *terra de filiis Guntiperti* "the plot of the sons of Guntipertus" (prepositional construction). Thus, the frequency of genitive plural forms in a document can be taken to be a rough inverse measure of the frequency of the prepositional construction and, consequently, of the faithfulness of that document to the classical grammatical standard. The genitive plural is attested 1,821 times in LLCT2 and seems to have been felt a prestige form. Our genitive plural variable indicates the percentage of the genitive plural form in all the words of a document. Once the syntactic annotation layer of LLCT2 is completed, it will be possible to calculate a corresponding percentage of the genitive plural in all the plural possessive constructions, which is, obviously, a more precise way. Nevertheless, the share of the total number of words is also likely to be descriptive enough, given that the documents are generally rich in possessive constructions.

The original Latin /*ae*/ diphthong had become monophthongized to open /ɛ/ in ordinary speech by the early 1st century AD, the development being originally no doubt subject to variation by location, register, and social stratification. In writing, the **<*ae*> diphthong** continued to be the standard until Late Antiquity and the monophthongized spelling <*e*> was sanctioned by grammarians ([Adams, 2013: 71–80]). However, in LLCT2, the attempts to preserve <*ae*> seem to have been largely abandoned and <*e*> prevails: only 6.5% of the <*ae*> diphthongs are written correctly, the best decades being the 760s and the 770s with 24% and 33% of correctness, respectively. Additionally, 0.3% of all the <*ae*> in LLCT2 are hypercorrect. This means that the diphthong <*ae*> has been used in environments where Classical Latin had no diphthong but a simple <*e*>. Even these date mostly back to the 760s/770s.

The <*ae*> diphthong is, of course, part of spelling correctness and included in the above general spelling correctness variable, but given its long history as a sign of learned language skills and the attention paid to it by all ancient grammatical treatises, we consider it useful to subject it to separate examination as well. The diphthong occurs both in word stems and prefixes and suffixes, the latter two being more easily memorizable. We will conduct no

statistical analysis on the here examined four variables, so two connected variables are not a problem. The <*ae*> diphthong variable indicates the percentage of the correctly written <*ae*> diphthongs in all the environments where that diphthong was required in Classical Latin. The <*ae*> diphthongs are counted using the same Levenshtein edit distance as with defining the spelling correctness as a whole.

V FROM THEORY TO PRACTICE

This section addresses the theoretical and practical issues that are met when visualizing linguistic variables in a network. We will focus on the use of colour in the representation of node attribute values. At its simplest, colour can be utilized to represent a discrete class membership: for example, in bibliometric networks, different colours may represent publications pertaining to different disciplines. The visualization of this kind of discrete variables is straightforward. The ambiguity begins, however, when the visualized attribute is continuous.

As was stated in section I, colours are often applied by gut feeling when visualizing the node attribute values. This is because the arising patterns are wished to support maximally the argument the author wants to promote at a given time. If no clear principle is applied to the colouring, one cannot know how the different colours are related to the distribution of the variable. In this and the following sections, we explore how continuous linguistic variables can be managed in networks in a visually effective and, at the same time, scientifically rigorous manner. As already mentioned, the edges between the 1,326 LLCT2 network nodes (documents, scribes, and locations) are unweighted. The edges link each document to the scribe wrote the document and to the location where the document was written. So, the scribes and the locations are not linked to each other except indirectly through the very document nodes.

The ways to convey information in a network are limited since the human eye is not capable of receiving much information at a time. On the other hand, it is not necessary to insist that the perception of all the aspects takes place simultaneously since a single network graph can be analyzed at leisure with the emphasis on various dimensions. In practice, the following means are available to convey information in a network graph: the text label, size, colour, shape, and position of the nodes. The edges also convey information on the relations between the nodes, but cannot be used to visualize the values of linguistic variables.

The node labels are a poor means because they require close reading. In the following graphs, we do not display the node labels, which can be, however, examined in the interactive Sigma.js visualizations (see below). In these interactive visualizations, the node label indicates the name of the scribe in the scribe nodes, the name of the location in the location nodes, and the abbreviation of the document in the document nodes. The node shape (e.g. circle, rectangle, triangle) would be a valid means of indicating a discrete attribute value, but only circle is utilized in this study which focuses on colour. The position of the node within the graph indicates the values of the geographical coordinate attributes in the LLCT2 network.

Since the linguistic variables visualized in section VI are continuous, such an indicator is wanted that reflects continuously the entire range of the variable. Only size and colour come into question because they can increase and decrease on a continuous scale. The human eye does not seem to be as good at discerning the relative sizes of a large set of objects at the same time as at perceiving colours (e.g. [Tarenskeen et al., 2015]). Thus, colour seems to be

the only really feasible means of visualization of linguistic variables. Consequently, we let the size to mark the node type, which only has three options: the document nodes are 7.0 pixels wide, the scribe nodes 10.5 pixels, and the location node 14.0 pixels. In this way, the most frequent nodes are the smallest and the most infrequent the biggest.

Gephi has several built-in gradient colour palettes. We will utilize the red-yellow-blue gradient where red stands for the low-value end of the range and changes gradually through yellow into blue, which stands for the high-value end of the range. This kind of three-colour hot-cold scale is better than a simple two-colour hue gradient scale, where a colour intensifies gradually from white to full saturation. The latter suits best the traditional black-and-white presentation in printed publications.

For scientific purposes, the use of gradient colours cannot be arbitrary, but must rely on some justifiable and theoretically valid principle. At the same time, this principle still has to lead into a visual outcome where the colours work in a maximally discerning way. This kind of theoretical grounding is the only way to make linguistic network visualization a scientifically reliable method that enables objectivity and replicability of the results. However, no network graph produced by way of a non-deterministic force-directed layout algorithm is strictly speaking visually replicable. This is because non-deterministic algorithms never produce two identical graphs although the underlying network data and the organizing principle is the same. This is also the case of the here applied ForceAtlas 2, which is utilized along with Geo Layout. Nonetheless, it seems to be clear that the gradient colouring has to be anchored to the statistical dispersion of the values of each visualized, continuous linguistic variable.

VI VISUALIZING THE LINGUISTIC FEATURES

In this section, we present the visualizations of the four linguistic features in the LLCT network and explain how we have realized the colouring. Each feature will be first visualized separately (Figures 4 to 7) and then subsumed under a sum variable (Figure 9). Each visualization will be interpreted in terms of its philological and/or linguistic contribution.

Table 1 presents the statistics which describe the four linguistic variables. The values of the variables are percentages, as explained in section IV. For convenience, Figure 3 shows the distributions of the variables as histograms. In Table 1, Minimum and Maximum stand for the lowest and highest attested value of each variable, respectively. Mean is the arithmetic mean of the values. Standard Deviation measures the dispersion of a set of data points from the mean. If the data points are close to the mean, the standard deviation is low. If they are spread across a wide range of values, the standard deviation is high.

	Spelling correctness	**Classical prepositions**	**Genitive plural**	**<*ae*> diphthong**
Minimum	85.68	61.88	0	0
Maximum	98.99	100	5.13	150.00
Mean	94.09	97.86	0.47	6.58
Standard Deviation	1.98	3.27	0.55	18.91

Table 1. Statistics describing the distribution of the variables.

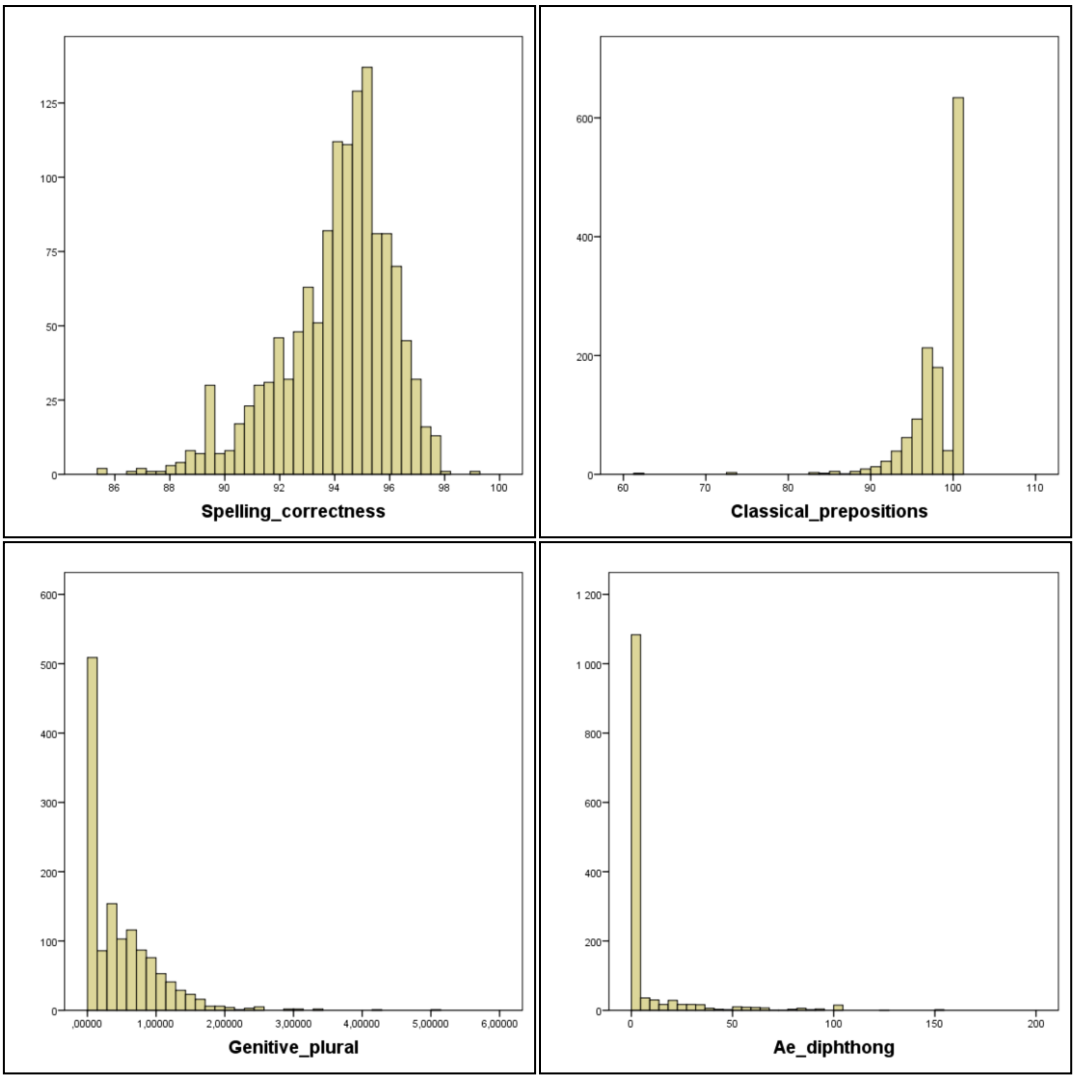


Figure 3. Histograms of the distributions of the four linguistic variables.

For statistical analysis, the sporadic radically low and high values of certain variables should be interpreted as outliers and left out of the examined range. For example, the normal procedure would be to discard at least the minimum value of the classical prepositions (61.88%) and the maximum value of the <*ae*> diphthong (150%). However, this cannot be done with network visualization where the idea is to make visible both the most typical and the most divergent traits of the data, so each node has to be present. Note that the 150% of the <*ae*> diphthong is due to the hypercorrect, i.e. excessively used, diphthongs (see section IV).

Spelling correctness is the only one of the four variables with the distribution roughly following the normal distribution. Being normally distributed is, however, not required by the present visualization task, which does not involve statistical analysis. The spelling correctness values of the document nodes vary between 85.68% and 98.99%, and the respective values of the scribe and location nodes settle themselves within this range because they are calculated by adding up the individual documents that these scribes wrote or that were written in that location. For the analyses of this article, the distributions of the linguistic variables were calculated across all the 1,326 nodes of the network and not separately for different types of nodes.

In order to ground the gradient colouring in the statistical dispersion of the variable values and to maintain maximal visual effect, we customized the Gephi colour palette so that the maximal yellow, which stands for the middle of the colour scale, marks the mean of the spelling variation distribution (94.09%). After that, we set the thresholds of the maximal red and maximal blue equally far from the mean. We chose the distance to be two standard deviations away from the mean. In so doing, only around 2.5% of the nodes with the lowest and highest values at both ends of the distribution are maximally saturated with red and blue while the rest, around 95%, of the nodes displays a gradient colour, including the maximal yellow in between. In this way, the maximal red and blue are not only reserved to the lowest value and the highest value of the variable. Figure 4 shows that, with two standard deviations, the extreme values come distinctly forward while the great majority of the values display a gradation which is easy to perceive. The network graphs are meant to be examined more closely by opening the interactive Sigma.js visualizations. The reader is advised to follow the links available in the caption of each graph figure.

When the layout algorithms are run, the documents written in each writing location are clustered close to the location node while being also connected to the scribe node, which places itself further apart. Thus, each document node is always connected with edges both to the scribe who wrote that document and to the place where that document was written while the scribe and location nodes are only connected to the document nodes (see section V). When observing the graphs, the reader should also bear in mind that the document nodes are represented by the smallest circles, the scribe nodes by the middle-sized circles, and the location nodes by the biggest circles.

Since 79% of the documents were written in Lucca, the location node of Lucca is surrounded with bunches of documents and scribes related to them. Some scribes have also written documents in other locations, mostly in small locations around Lucca, connecting thus the Lucca hub with most of Tuscia. On the one hand, the scribal mobility does not seem to have been considerable, given the many islets, such as Siena and Chiusi, which are fully detached from Lucca, the administrative and cultural centre of Tuscia. On the other hand, this picture is based only on the data that survives, most of which comes from the Luccan episcopal archives. It is noteworthy that, in the South, there seems to be another network of documents and scribes related to the abbatial archives of the monastery of Monte Amiata. For the place labels, see the interactive visualization.



Figure 4. Spelling correctness: red – low percentage, blue – high percentage. Open interactive version.

Since the distribution of the spelling correctness variable is approximately normal, the colour distribution is also relatively even when compared to that of the following graphs: here, all the colour grades are present and no one dominates. The graph shows that high and above-mean spelling correctness values are mostly concentrated in Lucca. Some scribes who are active both in Lucca and in an adjacent location also write blue or bluish documents. On the other hand, some minor locations in the immediate vicinity of Lucca perform rather poorly, i.e. the nodes are red or reddish. This results mainly from the fact that the few scribes active in those locations were non-classical spellers. It might have been best to merge even these locations with Lucca (see section III).

The spelling correctness percentage of Lucca is, indeed, 94.71, i.e. slightly over the mean of all the nodes (94.09%). What is perhaps more important is that all the substantial blue clusters of high-value documents are written in Lucca, whereas the blue location nodes outside Lucca are due to sporadic high-value documents (and scribes). Conversely, most of the red and reddish low-value nodes are situated outside Lucca, in Pisa on the coast, Garfagnana in the North, and in southern and south-western localities, such as Siena, Chiusi, and Monte Amiata. On the other hand, red is also found in Lucca and its suburbs, such as Vaccoli and Salisciamo, but to a lesser degree. All this elicits the conclusion that classical spelling was cherished in

Lucca more than in most other locations. In sum, the applied distributionally-based principle of gradient colouring seems to be suitable at least for variables which are not too badly divergent from normal distribution. The result is a graph with easily observable colour patterns that are, at the same time, grounded in statistical reality.

However, several corpus-linguistically noteworthy phenomena within LLCT2 display distributions which are far from normally distributed. The classical preposition, genitive plural, and <*ae*> diphthong variables have highly skewed distributions, as can be seen in Table 1 and in the histograms of Figure 3. Anyway, we continue to utilize the same red-yellow-blue palette as above in order to emphasize that the colours have the same meaning in each graph: red stands for low values and blue for high values, i.e. red means a weak faithfulness to the classical Latin grammar and blue a high faithfulness. On the other hand, it has to be reminded that, once the distributions are skewed and the standard deviations highly unequal, a certain colour gradation, as such, does not represent the same change in percentage points on the percentage scales.

The distribution of the **classical prepositions** ranges between 61.88% and 100%, with 644 (47.9%) of the nodes containing only classical prepositions. The variable is visualized in Figure 5. The most intense red begins again two standard deviations away from the mean towards the left tail (i.e. at value 91.31%) but, this time, the most intense blue cannot be two standard deviations away from the mean towards the right tail since the distribution is so skewed that the mean is closer to the highest value (100%) than that (see Figure 3). Thus the intense blue represents exclusively the value 100%.

In Figure 5, the blue colour, which represents the overrepresented value of the fat tail, predominates the graph surface while red, which represents the infrequent values of the long tail, has only low presence. On the other hand, the sporadic red nodes seem striking, given that they are location nodes which are the biggest-sized node category (see above). This is because, in the case of locations that only one document survives from, the colour of the document node equals the colour of the location node. Indeed, there are also fully red document nodes written in Lucca, but they do not dominate the picture since their effect is mitigated by the other Luccan documents with more classical prepositions, i.e. with a bluer hue. A corresponding phenomenon is observed with the blue location nodes.



Figure 5. Classical prepositions: red – low percentage, blue – high percentage. Open interactive version.

If compared to the spelling correctness graph in Figure 4, it is noticed that the blue colour, i.e. the high conservation of classical prepositions, is more evenly spread geographically. It is not particularly restricted to Lucca, but is also found in those locations where spelling correctness was relatively low. This may be at least partly explained by the fact that words are particularly salient items of language. Consequently, language users are typically more aware of the vocabulary than of the morphology and syntax they use (e.g. [Labov and Harris, 1986: 21]). It has been relatively easy to memorize the closed inventory of classical prepositions and, thus, to avoid prepositions alien to that. This would explain the high proportion of the value 100%.

On the other hand, the colour variation of the substantial document clusters shows that although there are a few productive scribes who never utilized innovative prepositions in the documents that survive to us, most of these scribes did utilize both classical and innovative prepositions. It can be hypothesized that innovative prepositions were chosen especially in cases where classical prepositions were felt to be ambiguous. The meaning of the innovative prepositions was usually more precise than that of classical prepositions, which often had a broad semantic scope ([Herman, 2000: 96]). Therefore, the use of innovative prepositions is likely to be in part related to the different semantic contents of the documents: some propositional contents are better expressed by using innovative prepositions, some of which

described, for example, precise relations between locations (e.g. *desub* "from under", *recta* "along "), an important issue with estate sale contracts.

Thus, it can be argued that the classical preposition variable does reflect the scribes' learned language competence at least with those scribes who were able to produce documents with 100% of classical prepositions. At the same time, it suggests that at least some scribes considered it legitimate to choose, if necessary, an innovative preposition instead of a classical, perhaps less specific, preposition. Yet others were clearly not aware of the relative prestige statuses of prepositions or did not care. Consequently, it can be concluded that the classical prepositions and their functionalities were no longer fully understood while the local language attitudes did not sanction resorting to innovative prepositions. The motivations underlying the preposition use are, thus, various and would deserve a full comparative study of their own.

The distribution of the **genitive plural** form ranges between 0% and 5.13%, with 691 (51.4%) of the nodes showing no occurrence of genitive plural. The variable is visualized in Figure 6. The most intense blue begins two standard deviations after the mean (i.e. at value 1.57%), whereas the most intense red is reserved for the value 0% (i.e. no genitive plurals).



Figure 6. Genitive plurals: red – low percentage, blue – high percentage. Open interactive version.

The graph is rather reddish, given that so many documents contain no genitive plural form. It is to be noticed that the genitive plural variable is different from the other three in that here the percentage expresses the share of the genitive forms in the total number of words in each document, scribe, and location, as was explained in section IV. The highest percentage (5.13%) is, thus, by no means the highest possible genitive plural share of a text, but just the maximum of the LLCT2 data. A share of 5.13% is quite a lot. This is explained by the fact that the document in question (CDL-208) is the shortest document of LLCT2, a memorandum of only 39 words, which contains two genitive plurals. The genitive plural percentage of an average LLCT2 document is 0.47%. It has to be remembered that the number of the genitive plural forms is also likely to depend, to a certain extent, on the propositional contents of each document as well as on the documentary formulae utilized in each document.

As far as the geographical distribution is concerned, it seems that, in Lucca, it was felt to be wholly acceptable to replace the genitive plural with the prepositional phrase with *de*, witness the large red clusters of documents written by productive Luccan scribes. On the other hand, it is true that there are some productive Luccan scribes, such as Rachiprandus I, Rachiprandus III, and Gumpertus I, who made frequent use of the genitive plural in all their documents. This proves that the use of the genitive plural form is an idiosyncratic phenomenon which is likely to tell about the scribe's learned language skills. Instead, the genitive plural seems to be much more frequent in certain non-Luccan locations, such as Siena, Monte Amiata, Pistoia, insofar as it is allowed to generalize on the basis of relatively few documents.

The skewest distribution of the four linguistic variables is that of the **<*ae*> diphthong**, which ranges between 0% and 150%, with 1,133 (84.3%) of the nodes containing no occurrence of the <*ae*> diphthong. The variation is, indeed, enormous, given that, as was mentioned in section IV, two documents use <*ae*> in excess and reach, thus, 125% and 150% of <*ae*>. The variable is visualized in Figure 7. The most intense blue again begins two standard deviations after the mean (i.e. at value 44.40%), whereas the most intense red is reserved for the value 0% (i.e. no <*ae*> diphthong).



Figure 7. <*ae*> diphthong: red – low percentage, blue – high percentage. Open interactive version.

The skewness of the distribution makes the graph surface mostly red. Only a few concentrations of blue and other hues are attested. Additionally, there are again a few salient blue and bluish location nodes, which receive their colour from a single blue document node. Based on the almost monochrome overall impression, it can be questioned whether the distribution of the <*ae*> diphthong should be considered anything but a dichotomous one: either a document makes use of the diphthong or it does not. The same question can also be raised regarding the classical prepositions and the genitive plural, although their distributions are not so extremely skewed.

Treating skewed variables as dichotomous could be justified by the argument that, if a scribe utilized once a certain classical feature which had to be learnt by training, he knew that feature. This is particularly true with single forms and conventions (e.g. genitive plural, <*ae*> diphthong), less so with categories which involve several items (e.g. spelling correctness as a whole and the group of classical prepositions). With the genitive plural, the dichotomization makes sense because, if a scribe did not know the obsolescent genitive plural ending, he could not use it (0%). Instead, if he used it once or more often (> 0%), he inevitably was familiar with the ending, however superficial that familiarity was. Nonetheless, the scribe's actual

productivity of the genitive plural form can only be approximated by observing the degree of the correct use of the feature, i.e. by examining the continuous variable.

Likewise, the dichotomization of the <*ae*> diphthong variable distinguishes those scribes who did not know the diphthong from those who knew it, regardless if they were able to utilize it only once or in all the words where it was required. This dichotomous distinction does not tell, however, whether a scribe knew only a few sporadic words where <*ae*> was required or whether he had memorized all the contexts. This is related to the fact that the occurrences of the diphthong had to be memorized case by case by heart, excluding those inflexional endings where it belonged. On that account, the <*ae*> diphthong is an important indicator of learned second-language competence in late Latin, which calls us to pay attention to the entire range of the variable. We will, therefore, not present dichotomized red-blue visualizations of the <*ae*> diphthong or the classical preposition and genitive plural variables, but go on treating them as continuous variables that can be visualized with gradient colour. The visualization of the <*ae*> diphthong will be discussed in more detail in the first case study of section VI.

To close this section, we seek to visualize the aggregate effect of the four linguistic features. We implement this by way of a **sum variable**, which is meant to measure the general language competence level of the documents, scribes, and locations. It is obvious that four features cannot describe thoroughly the entire field of the scribes' learned language skills, but they can still be claimed to form a rough estimate of their Latin second-language competences. It is equally obvious that a further study is needed to complement the feature repertoire with syntactic variables, but also with a larger set of relevant morphological and other variables.

We recoded the above four linguistic variables by standardizing them and, then, aggregated them into a sum variable. Standardization of variables (also known as z-score or a standard score) means rescaling their scale so that they have a mean of zero and a standard deviation of one. Once standardization has made the variables commensurate, they can be easily added up into a sum variable which is more nuanced than any sum variable built out of non-standardized variables would be. The distribution of the standardized sum variable is illustrated in Figure 8. The mean is, of course, still 0 while standard deviation is 0.588. The values around the mean are pronouncedly frequent.

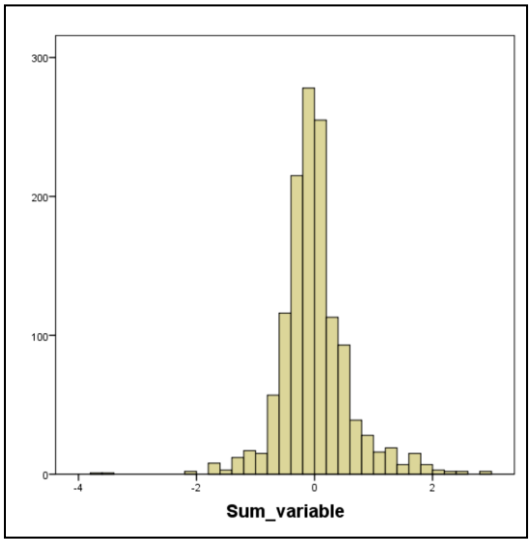


Figure 8. Histogram of the distribution of the standardized sum variable.



Figure 9. Sum variable of the four standardized linguistic features: red – low percentage, blue – high percentage. Open interactive version.

With the sum variable of Figure 9, the yellow values close to the mean are notably frequent, leading to a general yellowish tone of the graph surface. The most intense red and blue colours again begin two standard deviations away from the mean. As far as the sum variable can be considered representative of the scribes' general language competence, this competence shows no particular geographical variation, in spite of the fact that there is notable geographical variation in the spelling-related variables, i.e. spelling correctness and <*ae*> diphthong, as was seen in Figures 4 and 7. The main geographical observation based on Figure 9 is that the most competent prolific scribes seem to have been active in Lucca. It might also be justified to state, with caution, that the eastern and south-eastern parts of Tuscia show lower language competence than Lucca.

It seems to be useful to pay special attention to the clusters of documents written by the most productive scribes. These are only found in Lucca and neighbourhood. The output of each single scribe usually shows relatively moderate colour variation, which is, of course, to be

expected. It is noteworthy that no large cluster is fully red, but even the reddest clusters also contain yellow nodes. Instead, a few clusters are almost fully blue, with Austripertus I performing best. This means that the most competent productive scribes managed to keep within the highest 2.5% (threshold of the intense blue) of the nodes with most of their documents, whereas even the least competent scribes escape scoring within the lowest 2.5% with all their documents. Only the five documents written by Teutpert/Teutpertu II are either reddish or fully red.

VII TWO LINGUISTIC-PHILOLOGICAL CASE STUDIES

The goal of this section is to show which kind of philological and linguistic conclusions can be drawn by filtering the network visualizations presented in the previous section. In the first case study, we examine the <*ae*> diphthong in relation to the overall spelling correctness. As was stated above, the <*ae*> diphthong is an important indicator of learned second-language skills, given that it had not been at all supported by pronunciation for centuries. Figure 7 showed that the <*ae*> diphthong distinguishes nodes radically in the LLCT2 network, with the great majority of the scribes making no use of it. Figures 10 and 11 present an <*ae*> diphthong visualization where the nodes are filtered by spelling correctness. Figure 10 presents only the nodes with the general spelling correctness value below the mean (94.09%) and Figure 11 the nodes with the spelling correctness value over the mean.

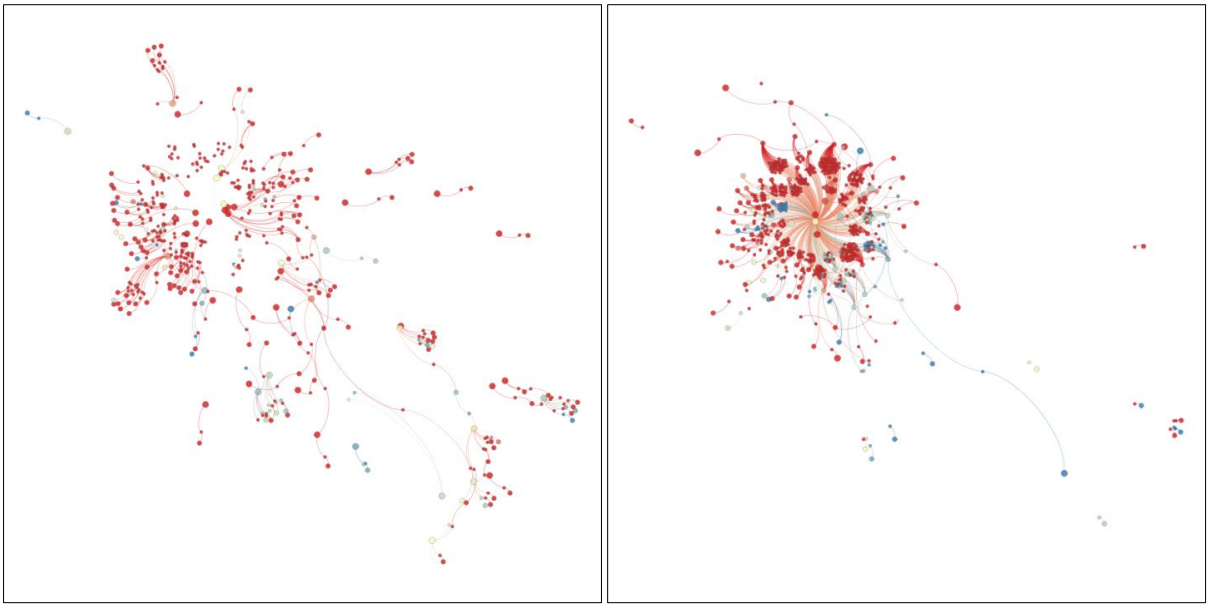


Figure 10. <*ae*> diphthong percentage in nodes with spelling correctness below the mean. Open interactive version.
Figure 11. <*ae*> diphthong percentage in nodes with spelling correctness above the mean. Open interactive version.

The gradient colour is applied to the distribution of the <*ae*> diphthong in the same way as in Figure 7. The general colour of the Figures 10 and 11 is still firmly red but, additionally, we notice a surprising pattern: Those nodes that show a high <*ae*> diphthong percentage but a below-mean general spelling correctness value (Figure 10) are mostly located outside Lucca, whereas the nodes with a high <*ae*> diphthong percentage and an above-mean spelling correctness percentage (Figure 11) are found both in and outside Lucca. In other words, there were scribes who knew to produce <*ae*> diphthongs both in Lucca and in other, smaller cities

and localities, such as Volterra, Chiusi, Grosseto, Populonia, and Fucecchio, including some locations dependent on Lucca, like Griciano. The difference is that in Lucca the diphthong is connected to a high general level of spelling while outside Lucca there is no such connection. Some sporadic scribes just seem to have known the diphthong, although they have not been particularly classical spellers in other respects.

This is likely to reflect local traditions in spelling and grammar as well as the relative standardization of the Luccan scribal training (see above Figure 4 and our conclusion on the general high spelling correctness level in Lucca). To our view, this suggests that, in terms of the local norm of documentary Latin in Lucca, it was, indeed, acceptable to ignore the <*ae*> diphthong, witness the large clusters of completely red documents written by established scribes who did relatively well with other linguistic features (note the spelling correctness percentage over the mean). However, some seemingly competent Luccan scribes, such as Osprand/Osprandus I, Filippus/Filippo I, Rachiprandus I, Gheipertus I, Richiprandus II, and, most strikingly, Austripertus I, utilized the diphthong. Thus, it was likely to have become an idiosyncratic feature, perhaps something that these scribes were proud of and considered a status symbol. The very same scribes show high percentages with all or most of the examined features, as proven by the sum variable of Figure 9. (See the scribe names by zooming in on the interactive visualizations.)

Of course, we cannot exclude the same idiosyncracy explanation when it comes to the non-Luccan scribes who utilized <*ae*>, but given that the spelling correctness outside Lucca is generally lower, it seems more likely that the sporadic high values of the diphthong are residues of differing scribal traditions, with different spelling preferences and different considerations about prestige. We do not claim that non-Luccan locations were in cultural isolation, but the local scribes may just have wanted to cherish their centuries-old documentary traditions by imitating their own ancient and venerable documentary funds. This kind of sticking to local traditions is attested, for example, with the continuous use of old Roman documentary formulary in Piacenza of the 8th century ([Schiaparelli, 1933: 10 ff.]). However, the interpretation is again complicated by the fact that much fewer documents survive from writing centres outside Lucca.

Although we suggested that using the monophthong <*e*> instead of the diphthong <*ae*> had *de facto* become the local Luccan norm, we do not claim that this norm was binding or imposed by some authority. On the contrary, the fact that certain scribes took pains with memorizing the word stems and morphological contexts where <*ae*> was classically utilized proves that at least some of the scribes had not given up pursuing classical grammatical models even when it concerned completely archaic and obsolete conventions, such as <*ae*>. These scribes must have been exposed to other texts than documents to be able to adopt the diphthong correctly. In practice, this means old Latin literature from Late Antiquity, which was being copied in the *scriptorium* of Lucca, and perhaps even grammatical treatises on spelling. [Schiaparelli, 1924: 56 ff.] has proposed, with good reason, that some Luccan documentary scribes also worked in the Luccan *scriptorium*, where they copied ecclesiastical and lay texts, both of which survive from early medieval Lucca.

Let us now stop to analyze the output of the above-mentioned scribe Austripertus I for a while. Austripertus might well have been one of these book copyists because he performs best in the entire LLCT2 network if measured by the spelling correctness variable of Figure 4, and is the fourth in the overall language competence reflected by the sum variable of Figure 9. Actually, he even has the highest score in the overall language competence if only prolific

scribes with more than two surviving documents are counted. Austripertus also utilizes the <*ae*> diphthong fully correctly in several documents, but his <*ae*> percentage still remains 89.4%. This is mostly because of one document with the <*ae*> diphthong percentage of 0, i.e. the one red node seen within Austripertus' document cluster in Figure 11. This kind of departure from the general trend requires an explanation and, indeed, there is one: the red document is one of the few copies included in LLCT2. The document is written originally by Austripertus I but copied later by Rachiprandus I who has replaced each <*ae*> with <*e*>, but retained everything else as it was. This interesting detail would hardly have been noticed without the network visualization.[2]

We now proceed to the second case study, where we utilize non-linguistic metadata attributes to filter the nodes whose colour represents the spelling correctness variable visualized in Figure 4. The metadata attributes will be time and scribe title, which introduce the chronological and slight sociolinguistic aspect into the analysis. Figures 12 to 15 present a visual "cross-tabulation" of the spelling correctness variable. The first two graphs (Figures 12 and 13) visualize the values of the spelling correctness variable for lay and ecclesiastical scribes and for the documents they produced until the year 812. The graphs of Figure 14 and Figure 15 visualize the respective values after 812. The few scribes with no title are combined with the lay scribes. Note that the location nodes have been removed from Figures 12 to 15. This is because the location nodes have a time interval that ranges across the entire time span of LLCT2. Consequently, the location nodes would disturb the interpretation of the chronologically filtered visualizations.

There were two types of scribes in early medieval Tuscia: ecclesiastical scribes, who were also clerics and worked by cathedrals or other churches, and lay scribes, who were employed by lay rulers, such as dukes, counts, and minor governors. The year 812 is a historical watershed since it is known that around 812 and 813 Tuscia got a new Frankish count Bonifatius I who began to contend for authority with the Luccan archbishop. Bonifatius seems to have excluded the clergy from documentary production, which had earlier been predominantly the playground of the church at least in Lucca [Keller, 1973: 119–124].

[2] In general, scribes copied documents very scrupulously [Sanga and Baggio, 1995: 250]. Only 4% of the LLCT2 documents are copies, and the great majority of them made by the same scribe who had written the original.

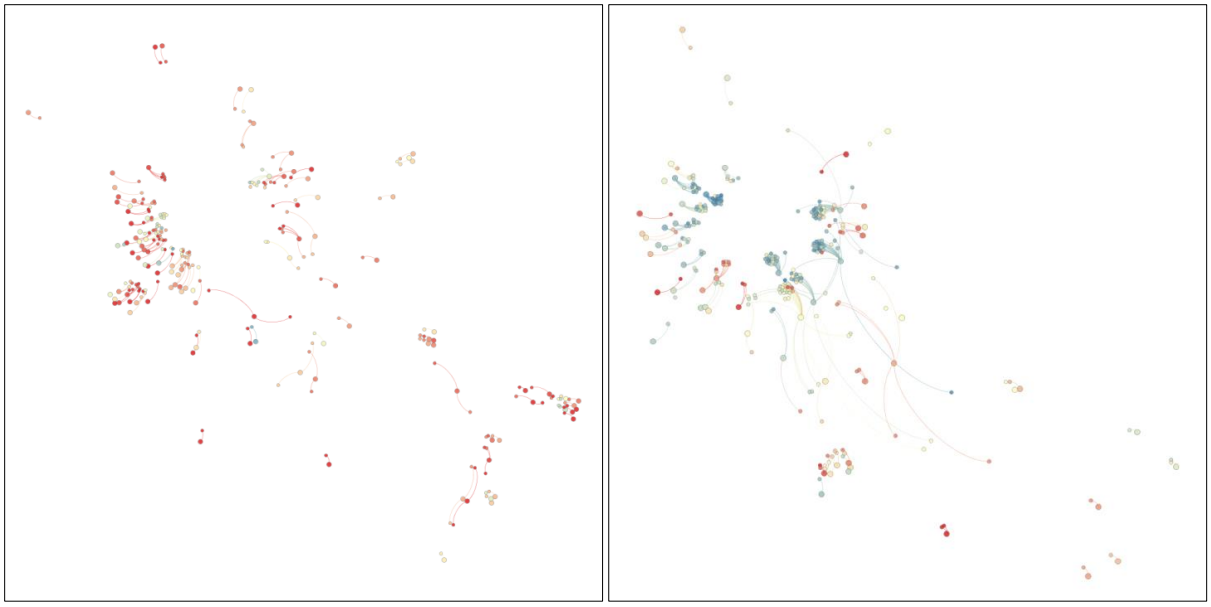


Figure 12. Spelling correctness of lay scribes and their documents until AD 812. Open interactive version.
Figure 13. Spelling correctness of ecclesiastical scribes and their documents until AD 812. Open interactive version.

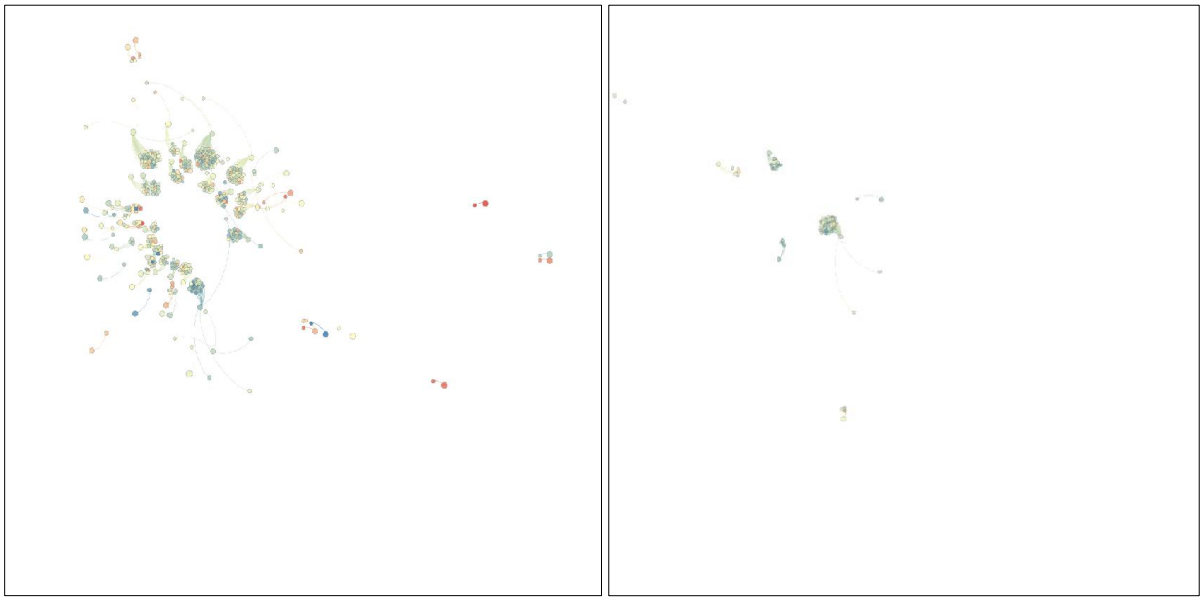


Figure 14. Spelling correctness of lay scribes and their documents after AD 812. Open interactive version.
Figure 15. Spelling correctness of ecclesiastical scribes and their documents after AD 812. Open interactive version.

First of all, the graphs of Figures 12 to 15 show that the volume of lay document production increases to some degree after 812 while the volume of ecclesiastical document production decreases radically, with only a few ecclesiastical scribes after 812. The graphs also testify to a geographical centralization: documents originating from outside Lucca are fewer after 812.

When we analyze the values of the spelling correctness variable, it turns out immediately that lay scribes have a much lower spelling correctness level before 812 than ecclesiastical scribes. This observation refers mainly to Lucca, though, given that the ecclesiastical scribes in the southern and south-western parts of Tuscia do not perform particularly well before 812. What

is noteworthy is that, after 812, the spelling correctness level of both the lay and the few remaining ecclesiastical scribes is close to or above the mean. All this confirms the conclusions drawn in [Keller, 1973] about the chronological discontinuity of document production and spelling in Tuscia. The pattern that arises from the above graphs also corroborates the hypothesis of [Korkiakangas, 2017] on the centralization and consolidation of document writing in Tuscia following the ouster of ecclesiastical scribes as main protagonists of documentary production. Conclusions of this kind show that the detailed visualization with carefully chosen filters allows for relevant philological and historical-linguistic inferences.

Conclusion

This article explored whether and how network visualization could benefit philological and historical-linguistic study. This was realized by applying gradient colour to visualize values of continuous linguistic variables in the LLCT2 network, a network of early medieval Latin documents, scribes, and locations. The gradient colour was anchored to the statistical dispersion of the variable values. Filters were utilized to filter nodes which met certain criteria. Fixing the location nodes on their geographical coordinate positions made geographical conclusions possible.

The network visualization turned out to be helpful in detecting patterns which are essential for philological and linguistic analysis and which easily remain unnoticed with traditional methods, such as cross-tabulation. Especially, it is practically impossible to draw geographical generalizations from tabular data. Network visualization can be, thus, an exploratory tool even in data sets which are relatively well known to the researcher. The interactive Sigma.js visualizations proved to be particularly effective in examining a graph dynamically on various levels of detail thanks to the zooming possibility.

The network visualization approach is, however, subject to the same limitations as all visualization techniques: the human eye can only catch a certain, relatively small amount of information at a time. At their best, gradient colour visualizations elicit important conclusions on surprising patterns, but they cannot replace a sound statistical analysis. While being illustrative, the visualizations may also easily mislead an incautious observer, especially if no clear pattern arises.

The philological and linguistic results of the article showed geographical differences between Lucca and other locations, especially in terms of general spelling correctness. Geographical variation was also attested with genitive plural, <*ae*> diphthong, and the sum variable. The visualizations made it easy to examine clusters of documents written by productive scribes. The application of time and scribe title filters to the spelling correctness variable confirmed results of existing historical-philological studies about the evolution of documentary praxis in historical Tuscia.

References

Adams J.N. *Social variation and the Latin language*. Cambridge University Press (Cambridge), 2013.

Araújo T. and Banisch S. Multidimensional Analysis of Linguistic Networks. Mehler A., Lücking A., Banisch S., Blanchard P. and Job, B. (eds) *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Springer (Berlin, Heidelberg), 2016, 107-131.

Bamman D., Passarotti M., Crane G. and Raynaud S. *Guidelines for the Syntactic Annotation of Latin Treebanks* (v. 1.3), 2007 http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf.

Barzel B. and Barabási A.-L. Universality in network dynamics. *Nature Physics*. 2013;9:673-681.

Bergs A. *Social Networks and Historical Sociolinguistics: Studies in Morphosyntactic Variation in the Paston Letters*. Walter de Gruyter (Berlin), 2005.

ISSN 2416-5999, an open-access journal

CDL = Schiaparelli L. *Codice diplomatico longobardo* I-II. Fonti per la storia d'Italia, 62-63. Tipografia del Senato (Roma), 1929-1933.
CDT = Brunetti F. *Codice diplomatico toscano*. Parte II, tomo I. Leopoldo Allegrini e Giovanni Mazzoni (Firenze), 1833.
ChLA = Cavallo G. and Nicolaj G. (eds) *Chartae Latinae Antiquiores*. Facsimile-edition of the Latin Charters. 2nd Series: Ninth Century. Urs Graf Verlag (Dietikon, Zürich), 1997-.
Ferrer i Cancho R. Network theory. Hogan P.C. (ed.) *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press (Cambridge), 2010, 555–557.
Herman J. *Vulgar Latin*. Translated by Roger Wright. The Pennsylvania State University Press (University Park), 2000.
http1 https://hucompute.org/applications/linguistic-networks/
http2 https://gephi.org/
http3 http://stats-1.archeogr.unisi.it/repetti/index.php
http4 http://sigmajs.org/
http5 https://github.com/cisocrgroup/Resources/tree/ master/lexica
http6 http://mk270.github.io/whitakers-words/operational.html
Jackson M.O. *Social and economic networks*. Princeton University Press (Princeton), 2008.
Keller H. La marca di Tuscia fino all'anno Mille. *Atti del V Congresso internazionale di studi sull'alto medioevo*, Lucca, ottobre 3-7, 1971. CISAM (Spoleto), 1973, 117-136.
Korkiakangas T. *Subject Case in the Latin of Tuscan Charters of the 8th and 9th Centuries*. Societas Scientiarum Fennica (Espoo), 2016.
Korkiakangas T. Spelling Variation in Historical Text Corpora: The Case of Early Medieval Documentary Latin. *Digital Scholarship in the Humanities*. 2017. https://doi.org/10.1093/llc/fqx061
Korkiakangas T. and Lassila M. Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material. Mambrini F., Sporleder C. and Passarotti M. (eds) *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3), Sofia, December 13, 2013. Bulgarian Academy of Sciences (Sofia), 2013, 61-72.
Korkiakangas T. and Passarotti M. Challenges in Annotating Medieval Latin Charters. *Journal of Language Technology and Computational Linguistics*. 2011;26,2:103-114.
Labov W. and Harris W.A. *De facto* segregation of black and white vernaculars. Sankoff D. (ed) *Diversity and Diachrony*. John Benjamins (Philadelphia), 1986, 1-24.
Lazard S. Indices de la langue parlée à Ravenne au VIe siècle à travers le témoignage des chartes. *Actes du XXe Congrès international de linguistique et philologie romanes.* 1993;II,3:391-402.
LLCT2 = *Late Latin Charter Treebank*, version 2. See Annex 1.
MED = Barsocchini D. *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo V, parte II. Francesco Bertini (Lucca), 1837.
MED2 = Bertini D. *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo IV, parte II. Francesco Bertini (Lucca), 1836.
MED3 = Barsocchini D. *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomo V, parte III. Francesco Bertini (Lucca), 1841.
Passarotti M.C. The Importance of Being sum. Network Analysis of a Latin Dependency Treebank. Basili R., Lenci A. and Magnini B. (eds). *Proceedings of the First Italian Conference on Computational Linguistics*, CLiC-it 2014, Pisa, December 9-10, 2014. Pisa University Press (Pisa), 2014, 291-295.
Rovai F. The Development of Deverbal Prepositions in Latin: Morpho-Syntactic and Semantico-Pragmatic Factors. *Archivio Glottologico Italiano*. 2013;98,2:175-213.
Rutkowska H. and Rössler P. Orthographic Variables. Hernández-Campoy J.M. and Conde-Silvestre J.C. (eds) *The Handbook of Historical Sociolinguistics*. Blackwell (Chichester), 2012, 213-236.
Sanga G. and Baggio S. Sul volgare in età longobarda. Banfi E., Bonfadini G., Cordin P. and Iliescu M. (eds) *Italia settentrionale: crocevia di idiomi romanzi*. Niemeyer (Tübingen), 1995, 247-260.
Schiaparelli L. *Il codice 490 della Biblioteca capitolare di Lucca.* P. Sansavini (Roma), 1924.
Schiaparelli L. Note diplomatiche sulle carte longobarde, II: Tracce di antichi formulari nelle carte longobarde. *Archivio storico italiano.* 1933;19:3-34.
Sornicola R. Bilinguismo e diglossia dei territori bizantini e longobardi del Mezzogiorno: le testimonianze dei documenti del IX e X secolo. *Quaderni dell'Accademia Pontaniana*. 2012;59:1-102.
Tarenskeen S., Broersma M. and Geurts B. Overspecification of color, pattern, and size: salience, absoluteness, and consistency. *Frontiers in Psychology*. 2015;6.
Valentini C. *L'evoluzione della codifica del genitivo dal tipo sintetico al tipo analitico nelle carte del Codice diplomatico longobardo*. Ph.D. thesis, University of Florence, 2017.
Wright R. The conceptual distinction between Latin and Romance: invention or evolution. Wright R. (ed) *Latin and the Romance Languages in the Early Middle Ages* (2nd ed. 1996). The Pennsylvania State University Press (University Park), 1991, 103-113.

ANNEX 1

LLCT1 (version 1) is available at https://doi.org/10.5281/zenodo.1197357. LLCT2 (version 2) will be made available in Zenodo in 2018. Meanwhile, the LLCT2 data relevant for the present article, i.e. the Gephi files underlying the network graphs of Figures 4 to 15, can be downloaded from https://doi.org/10.5281/zenodo.1064693.