Time Series of Linguistic Networks in the Patrologia Latina*

Alexander Mehler, Rüdiger Gleim, Ulli Waltinger & Nils Diewald
Bielefeld University

Abstract: We analyze a corpus of historical texts in terms of complex network theory. This is done by means of the *Patrologia Latina*, a collection of Latin documents that were written over a period of more than 1,000 years. We perform a lemmatization of this corpus and map its documents onto the end of the productive period of the corresponding authors. By means of this temporal ordering, we perform a transitivity analysis on the level of lexemes and sentences as a function of time. This analysis shows a remarkable law-like behavior of transitivity on the lexical and sentential level.

1 Introduction

A long-term historical text corpus comprises natural language texts that were produced over a long period of time in the past. An example is the *Patrologia Latina* (PL) [Mig55], which is a collection of writings of the Church Fathers and other ecclesiastical authors that were written over a period of more than 1,000 years. With the growing availability of tools for automatic text analysis, the processing of historical corpora comes more and more into the focus of text-technology [LPF05]. There are four areas of advancements in this area: *corpus building*, *resource formation*, *corpus management*, and *corpus mining*. With a focus on Latin, these approaches can be reviewed as follows:

- *Corpus building and resource formation:* a multitude of projects aim at building text collections and related lexical and syntactic resources. This includes, for example, corpora of Latin texts of the early modern time as provided by the *Camena* project [Sch01]) as well as full-text databases of Latin literature as included in the *Bibliotheca Teubneriana Latina* (BTL-1) [Tom99] and its companion editions. As one of the most sophisticated projects in this area, the *Perseus Digital Library* [SRCC00] provides not only historical corpora, but also syntactic annotations [BPBC08]. See [Kos05] and [PD10] for related endeavors. Note that all these projects integrate full-form lexica of Latin.
- *Corpus management and mining:* the Perseus project has been built as a digital library that operates on historical corpora of several languages. Recently, two additional projects have started that also integrate facilities for corpus management in order to support text mining on Latin texts. Firstly, this relates to the *eAQUA* project

*Financial support of the German Federal Ministry of Education and Research (BMBF) through the project *Linguistic Networks* (see www.linguistic-networks.net) is gratefully acknowledged.

[BHG08], which aims at extracting linguistic knowledge from ancient resources by means of explorative data analysis. Secondly, the *Linguistic Networks Project* and its eHumanities Desktop [GM10] provides, amongst others, access to the PL by means of related methods. Both approaches leave the narrow focus on present-day corpora in order to explore historical corpora and, thus, gain access to language change.

In this paper, we analyze the PL as a long-term historical corpus of Latin texts. This is done with the help of mapping all PL documents onto the dates of the end of the productive period of their authors. By means of this temporal ordering we derive a subcorpus of 1,000 texts in order to perform a cluster analysis on the level of lexemes and sentences as a function of time (Section 3): starting from a graph model of layered linguistic networks (Section 2), we utilize complex network theory to get insights into the temporal dynamics of the PL as a long-term historical corpus (Section 4). We show a remarkable law-like behavior of clustering on both the lexical and the sentential level. In order to provide comparative values, we analyze a corpus of present-day texts in the same framework.

2 A Graph Model of 2-layer Linguistic Networks

A natural language is a multiresolutional system that is structured on various, interwoven linguistic levels (e.g., the morphological, lexical or syntactic level). From a formal point of view, such a system can be modeled as an instance of the class of *multilayer graphs*. Generally speaking, a directed *k-layer graph*

$$D \quad = \quad (V, A, B) \tag{1}$$

is a digraph whose vertex set is partitioned into non-empty, pairwise disjoint subsets $V_1, \ldots, V_k$ such that the following conditions hold:

- *Layer-external arcs:* $\forall a \in A \exists! V_i, V_j \in \{V_1, \ldots, V_k\} : \mathrm{in}(a) \in V_i \neq V_j \ni \mathrm{out}(a)$.
- *Layer-internal arcs:* $\forall a \in B \exists! V_i \in \{V_1, \ldots, V_k\} : \mathrm{in}(a) \in V_i \ni \mathrm{out}(a)$.

The ith layer of a k-level graph, $i \in \{1, \ldots, k\}$, is defined by the subgraph $D(i) = D_i = (V_i, B_i)$ of (V, B) such that $a \in B_i \leftrightarrow \mathrm{in}(a) \in V_i \ni \mathrm{out}(a)$. Likewise, the bipartite graph that is induced by the layers i, j is defined by the subgraph $D(i, j) = D_{ij} = (V_i \cup V_j, A_{ij})$ of (V, A) such that $a \in A_{ij} \leftrightarrow (\mathrm{in}(a) \in V_i \wedge \mathrm{out}(a) \in V_j) \vee (\mathrm{in}(a) \in V_j \wedge \mathrm{out}(a) \in V_i)$ (note that $D_{ij} = D_{ji}$). Subsequently, we experiment with two linguistic layers, the lexical and the syntactic layer. That is, we deal with 2-layer graphs $D = (V, A, B)$ where A is the set of arcs that either connect words with sentences or vice versa, while B is the union of lexical and sentential arcs. Thus, we have to induce three graphs D_1, D_2 and D_{12} to build a 2-layer graph where D_1 is a co-occurrence graph, D_2 is a graph of sentential relations and D_{12} links sentences with their lexical constituents:

- $D_1 = (V_1, B_1)$ is induced as a lexical co-occurrence network by means of the co-occurrence measure $\sigma : V_1 \times V_1 \to \mathbb{R}_0^+$ of [HQW06]. That is, vertices $v \in V_1$ denote

lexical units that are output by lemmatizing the PL, while arcs denote significant lexical co-occurrences in the sense of σ. This procedure of network induction is described in detail in [MDW$^+$10]. Note that in contrast to [HQW06], we do not operate on word forms, but on lexemes. Note further, that for every arc $a \in B_1$ there is an arc $b \in B_1$ such that $\mathrm{in}(a) = \mathrm{out}(b)$ and $\mathrm{in}(b) = \mathrm{out}(a)$ since lexical associations are, as mapped here, symmetric.

- $D_{12} = (V_{12}, A_{12}) = (V_{12}, A)$ is induced as follows: the vertex set of D_{12} is the union of the set of lexemes V_1 and the set of sentences V_2. Further, for any lexeme $v \in V_1$ for which there exists a word form as a lexical constituent of the sentence $w \in V_2$, we generate two non-parallel but multiple arcs as elements of A_{12} that end at v and w, respectively.

The third step is to induce the sentence network $D_2 = (V_2, B_2)$ for which $B_2 = B \setminus B_1$. Our starting point is to connect sentences that express similar concepts, that is, sentences with similar conceptual meanings. Evidently, the automatic detection of these meanings is out of reach so that we need a heuristic. We do this by means of the following hypothesis:

> **H1:** *The more lexical units two sentences have in common and the higher the degree of semantic specificity of these units, the higher the contribution of these common lexical constituents to the conceptual similarity of both sentences.*

Obviously, sentences that have no lexical constituents in common can still be semantically related, for example, by a relation of entailment. Such relations are currently out of reach of being automatically detected in huge corpora such as the PL. Therefore, we start from H1 by saying that the probability of a conceptual similarity of two sentences is a function of the semantic specificity and the number of the lexical constituents they share. This approach can be implemented straightforwardly. In set-theoretical terms, sentences can be represented as collections of lexical units so that multisets can be used to model the first parameter, that is *number*. Further, the notion of semantic specificity can be modeled by means of *idf*-scores [SB88]. Now, let $v_i, v_j \in V_2$ be two sentences that are represented as multisets S_i and S_j. Then, the *lexical overlap* of v_i and v_j is computed as

$$\omega(S_i, S_j) = \frac{\sum_{x \in S_i \cap S_j} i(x)}{\sum_{x \in S_1} i(x) + \sum_{x \in S_2} i(x) - \sum_{x \in S_1 \cap S_2} i(x)} \tag{2}$$

where $S_i \cap S_i$ is the multiset intersection of the set representations of v_i and v_j. $i(x)$ is the *inverse document frequency* (idf) of x in the underlying reference corpus, that is, the PL. σ has the property that if S_1 and S_2 are identical, then $\sigma(S_1, S_2) = 1$. Otherwise, if their intersection is empty, then $\sigma(S_1, S_2) = 0$. Note that we do not say that the conceptual similarity of sentences equals their lexical overlap. Rather, we use $\omega(S_i, S_j)$ to approximate the impact of shared lexical constituents to conceptual similarity in the sense of H1.

Variable	Value	Variable	Value
number of authors	1,320	number of sentences	7,727,864
number of texts	4,555	number of tokens	121,722,687
number of paragraphs	674,718	number of word forms	1,094,850

Tabelle 1: Quantitative characteristics of C1, the subset of the PL without commentaries.

3 The Corpora

In order to instantiate our model of 2-layer graphs, we start from a subset of 4,555 texts of the PL by excluding all commentaries (see Table 1). Henceforth, this subcorpus is denoted by *C1*. It is input to a preprocessing module that includes a sentence boundary detection, named entity recognition, lemmatization and an annotation of the logical document structure of all texts in *C1*. The details of this preprocessing and of the preceding formation of a full-form lexicon are described in [MDW$^+$10]. In this section, we concentrate on describing the process of network induction based on the PL. In order to arrive at a manageable time series analysis, we perform two steps of corpus selection: firstly, we rank all texts in *C1* by their number of tokens and select the 1,000 highest ranked documents that follow the first 14 ranks. This subcorpus is denoted by *C2*. As we induce for each of the documents in *C2* a 2-layer graph, we have to analyze 2,000 networks in terms of their topological characteristics. Obviously, this is a huge comparative network analysis. Thus, the reason to build *C2* is to reduce computational effort and to keep the corpus more balanced. Secondly, we perform a mapping of the documents of the PL onto the date of the end of the productive period of the corresponding author. This is done to get a temporal ordering of *C2*. As a result of additionally applying a random ordering of documents with the same time values, we get a linear ordering $C = (x_1, \ldots, x_{1000})$ of *C2*. The time values are annotated manually where we use the *Documenta Catholica Omnia* to get the author information. As a result of this selection, we get the subcorpus *C3* of 1,000 documents that are ordered by the end of the productive period of their author. *C3* is input to induce a separate 2-layer graph for each of the 1,000 documents according to Section 2. That is, we map *C3* onto a corresponding time series of 2-layer graphs

$$(D^{[1]}, \ldots, D^{[n]}) \tag{3}$$

where for each $t \in \{1, \ldots, n\}$: $D^{[t]} = (V^{[t]}, A^{[t]}, B^{[t]})$, $n = 1000$, is the two-level graph of lexical and sentential networking in document x_t at time t such that $D_1^{[t]} = D^{[t]}(1)$ is the lexical and $D_2^{[t]} = D^{[t]}(2)$ is the sentential layer of $D^{[t]}$.

As a basis of comparison we analyze a corpus of newspaper articles of the *Süddeutsche Zeitung* (SZ) from 1994 by the same procedure. The difference is that 2-layer networks are not induced for each article as this would get tiny graphs, but not complex networks. Instead, we deal with each daily issue as a single document for which we induce a lexical and sentential network as described above. This gives a time series of 301 2-layer networks.

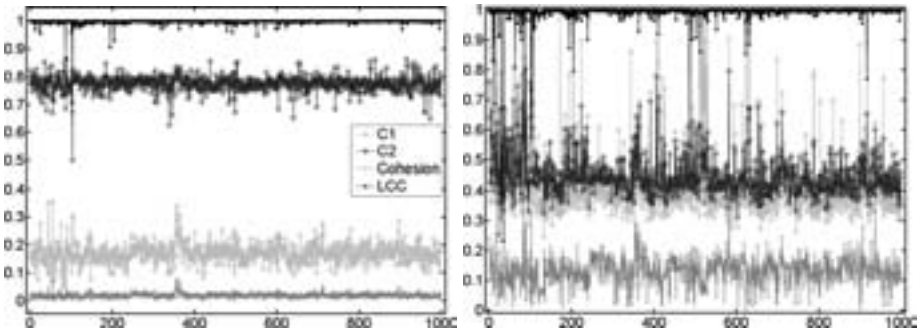

Abbildung 1: Time series of indices of lexical (left) and sentential networks (right) in the PL.

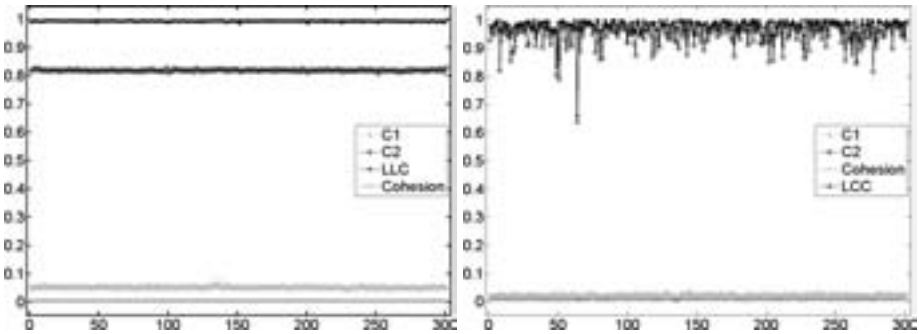

Abbildung 2: Time series of indices of lexical (left) and sentential networks (right) in the SZ.

4 Variation of Clustering of Textual Units in Time

In order to get insights into the temporal dynamics of the PL by example of the time series of 2-layer graphs (see Equation 3) we compute four topological indices of complex networks. The density or cohesion of a graph $G = (V, E)$ is the number of its edges in relation to the number of all possible edges in G: $coh(G) = \frac{\sum_{v \in V} d_G(v)}{|V|^2 - |V|} \in [0, 1]$ where $d_G(v)$ is the degree of $v \in V$. Note that $coh(G)$ is computed for the undirected variant of $D_1^{[t]}$ and $D_2^{[t]}$, respectively, $t \in \{1, \ldots, n\}$. Our expectation is that linguistic networks of the sort considered here are sparse networks by analogy to small-world graphs [New03]. This expectation is confirmed on the lexical and sentential level by example of the newspaper corpus (see Figure 2). In both cases we observe a cohesion near to zero. Interestingly, this observation correlates with the fact that the fraction lcc of vertices that belong to the *Largest Connected Component* (LCC) is near to 1 – although in the sentential network it is smaller than in the lexical network. Thus, lexical and sentential networking results in highly sparse though connected graphs in the case of present-day language.

In the case of the PL, we observe a remarkably high value of lcc and a near to zero cohesion on the lexical level. However, if we look on the sentential level, we observe a difference

to the newspaper corpus: although a fraction of nearly 100% of vertices that belong to the LCC is retained in almost all cases, the cohesion is much larger. One reason for this difference is that while in the case of the Pl, sentential networks are computed by including all links of at least average specificity (in the sense of Equation 2), in the case of the SZ we retain a connectedness of nearly 100% if we include only those sentence links whose weight is $\geq \mu + 5\sigma$. If we apply the same criterion to the sentence networks of the PL, we get disconnected networks where $lcc \leq 70\%$. Thus, we conclude that networking is much more stable in the present-day corpus than in the historical corpus. However, we also observe a remarkably stable network pattern in the case of historical lexical networks. To a minor degree this also holds for the historical sentential networks.

The second pair of indices that we consider relates to what is called transitivity in quantitative sociology, that is, the probability by which neighbors of the same vertex (e.g., a person) are themselves neighbors (e.g., friends). More formally, the transitivity or cluster value of an undirected graph $G = (V, E)$ is as follows [WS98]:

$$C_2(G) = C_{ws}(G) = \frac{1}{n}\sum_{i=1}^{n} c_{ws}(v_i) = \frac{1}{n}\sum_{i=1}^{n} adj(v_i) / \binom{d_G(v_i)}{2} \in [0,1] \tag{4}$$

where $adj(v_i)$ is the number of edges ending only at neighbors of v_i. [BR03] discuss an interesting alternative to C_2 that is computed as follows:

$$C_1(G) = C_{br}(G) = \left(\sum_{v \in V} \binom{d_G(v)}{2} c_{ws}(v)\right) / \sum_{v \in V} \binom{d_G(v)}{2} \tag{5}$$

The difference of C_2 and C_1 lies in the fact that the latter is a weighted mean in contrast to C_2, which is an arithmetic mean. C_1 weights the impact of the transitivity of each vertex v by its degree: the higher $d_G(v)$, the higher the impact of $c_{ws}(v)$ to C_1.

If we look on Figure 1 and 2 we observe a remarkably stable distribution of C_2 and C_1 on the level of lexical networks (though to a minor degree in the case of lexical networking in the PL). This patterned distribution is also observable in the case of sentential networks derived from the present-day corpus and to a much smaller degree in the case of the historical corpus. Interestingly, this exception correlates with the fact that in the case of sentential networking in the PL, C_1 and C_2 have a similar spectrum of values. In all other cases we observe a tremendous divergence of both cluster values. This is a hint that in these networks, high-degree vertices have small cluster values, while in the sentence networks of the PL, high-degree and low-degree vertices tend to have the same impact. These findings are in support of [BR03] who stress the need to use weighted cluster coefficients, which are more expressive in terms of network topology.

In summary: with the exception of historical sentence networks, we observe an almost constant transitivity of linguistic networks *irrespective of time*. Thus, clustering, cohesion and connectedness evolve as stable topological indicators that hint at a law-like linguistic networking. However, we also observe remarkable differences between historical and present-day language. There may be at least four reasons for these differences: in the case of the newspaper corpus we can expect much more stable patterns of text production.

There is certainly a tendency to highly stabilized text types due to a standardized process of text production. This level of standardization may result in the almost constant cluster coefficients of lexical and sentential networks. Secondly, we have to expect less heterogeneous genres in the case of newspaper articles in contrast to the PL with its indices, sermons, handbooks etc. Thus, we need a more precise *text-typological* analysis of the PL and a sophisticated look at *clustering in different genres* in order to underpin our findings. Thirdly, the SZ is a short-term corpus, while the PL is a long-term corpus. This difference in temporal scale may be an additional source of differences in the sense that the PL manifests language change that is not present in the SZ. Finally, the PL-based networks are computed per document, while the SZ-based networks are computed per daily issue. The corresponding differences in text coherence may also be a source of the deviations.

5 Conclusion

We analyzed a subset of the Patrologia Latina in terms of its temporal dynamics. We found a pattern of clustering on the lexical level and on the sentential level, though to a minor degree. In the case of lexical networking, this pattern is stable irrespective of the period of time. Future work aims at investigating this pattern using larger sets of reference corpora.

Literatur

[BHG08] Marco Büchler, Gerhard Heyer und Sabine Gründer. eAQUA: Bringing modern Text Mining approaches to two thousand years old ancient texts. In *Proc. of the e-Humanities workshop at the 4th IEEE Int. Conf. on e-Science*, 2008.

[BPBC08] David Bamman, Marco Passarotti, Roberto Busa und Gregory Crane. The Annotation Guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. In *Proceedings of LREC 2008*, Marrakech, Morocco, 2008. ELRA.

[BR03] Béla Bollobás und Oliver M. Riordan. Mathematical Results on Scale-Free Random Graphs. In Stefan Bornholdt und Heinz Georg Schuster, Hrsg., *Handbook of Graphs and Networks*. Wiley-VCH, Weinheim, 2003.

[GM10] Rüdiger Gleim und Alexander Mehler. Computational Linguistics for Mere Mortals — Powerful but Easy-to-use Linguistic Processing for Scientists in the Humanities. In *Proceedings of LREC 2010*, Malta, 2010. ELDA.

[HQW06] Gerhard Heyer, Uwe Quasthoff und Thomas Wittig. *Text Mining: Wissensrohstoff Text*. W3L, Herdecke, 2006.

[Kos05] Cornelis H. A. Koster. Constructing a Parser for Latin. In Alexander F. Gelbukh, Hrsg., *Proceedings of CICLing 2005*, Seiten 48–59, 2005.

[LPF05] Anke Lüdeling, Thorwald Poschenrieder und Lukas C. Faulstich. DeutschDiachron-Digital. *Jahrbuch für Computerphilologie*, Seiten 119–136, 2005.

[MDW+10] Alexander Mehler, Nils Diewald, Ulli Waltinger, Rüdiger Gleim, Dietmar Esch, Barbara Job, Thomas Küchelmann, Olga Pustylnikov und Philippe Blanchard. Evolution of Romance Language in Written Communication. In *Arts, Humanities, Complex Networks: a Leonardo satellite symposium at NetSci 2010*, 2010.

[Mig55] Jacques-Paul Migne, Hrsg. *Patrologiae cursus completus: Series latina*, Jgg. 1–221. Chadwyck-Healey, Cambridge, 1844–1855.

[New03] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[PD10] Marco Passarotti und F. Dell'Orletta. Improvements in Parsing the Index Thomisticus Treebank. In *Proceedings of LREC 2010*, 2010.

[SB88] Gerard Salton und Chris Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing Management*, 24(5):513–523, 1988.

[Sch01] Wolfgang Schibel. CAMENA-Neulateinische Dichtung im WWW. *Neulateinisches Jahrbuch*, 3:211–219, 2001.

[SRCC00] David A. Smith, Jeffrey A. Rydberg-Co und Gregory R. Crane. The Perseus Project. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[Tom99] Paul Tombeur, Hrsg. *Bibliotheca Teubneriana Latina (BTL 1)*. Teubner/Brepols, Stuttgart, Leipzig/Turnhout, 1999.

[WS98] Duncan J. Watts und Steven H. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, 1998.