Evolution of Romance Language in Written Communication: Network Analysis of Late Latin and Early Romance Corpora

Alexander Mehler, Nils Diewald, Ulli Waltinger, Rüdiger Gleim, Dietmar Esch, Barbara Job, Thomas Küchelmann, Olga Pustylnikov, Philippe Blanchard; University of Bielefeld, Germany
E-mail: Alexander.Mehler@uni-bielefeld.de.

Submitted: <leave for Editor to date>

Abstract

In this paper, we induce linguistic networks as a prerequisite for detecting language change by means of the *Patrologia Latina*, a corpus of Latin texts from the 4th to the 13th century.

Corpus

The *Patrologia Latina* (PL) is the most important text corpus of Christian documents from the very beginning in Late Antiquity to the High Middle Ages, comprising 8,508 documents (including commentaries) dating from the 4th to the beginning 13th century, written by 2,004 authors. The collection has been compiled in the first decades of the 19th century by Jacques Paul Migne and has been printed in a first edition from 1844 to 1855 [1]. A digital version was published in 1993.

The texts represent the whole range of discourse traditions and communicative genres of the period: from private letters, biographies of church fathers, sermons, and ecclesiastical commentaries to juridical texts. The authors show a spectrum from famous church fathers and popes to nearly unknown or anonymous authors. In regard to diachronic, stylistic, and regional registers, the language of these texts varies considerably.

In order to induce linguistic networks from the PL, we used a subset of 4,555 documents by excluding commentaries (see Table 1). To get an idea of the PL, look at the German Wikipedia, which has more than 3 Mio authors who have produced more than 400 Mio tokens (till 2008). In contrast to this, the 2,000 authors of the PL have produced more than 100 Mio tokens – a quarter of the German Wikipedia, which shows the tremendous size of the PL as a historical corpus.

Table 1: Some characteristics of the PL.

Variable	Value
Author	1,320
Text	4,555
Paragraph	674,718
Sentence	7,727,864
Token	121,722,687
Word form	1,094,850

Language Change

The wide chronological and textual range of documents compiled in the *PL* allows for various analyses of language change. Although the *PL* represents only written texts, and language change is considered to take its origin primarily in spoken everyday language, several documents within this corpus are known to be closely related to spoken Latin, written sources of the so called 'Vulgar Latin'. An example of this variety is the *Historia Francorum* of Gregory of Tours. The evolution of classical Latin to Romance languages becomes transparent especially in these texts. For instance, the grammaticalization of *habere*: Starting from the full verbal meaning of possession, *habere* went through a process of auxiliation [6, 8]. This can be analysed by looking at combinations of the word with full verb forms.

On a lexical level, a change of word usage can be seen as a replacement of words in the same context. For example, the adjective *pulcher* (‘nice’ or ‘pretty’) was replaced by the word *bellus* (a diminutive form of bonus) with the same meaning used in similar contexts. This form remains in Romance language (e.g., *beau*) while *pulcher* vanished [9]. While the first change can be analysed with a syntagmatic view on the co-occurrence of the linguistic form, the second one needs a paradigmatic view. Both views give rise to network analysis. In this paper, we focus on the former by inducing networks of lexical units whose edges model relations of syntagmatic contiguity.

Table 2: Resources of lexicon formation.

Source	Lemmata	Word forms
AGFL	5,494	244,582
Perseus	45,194	225,360
LemLat	32,862	125,920
Names	8,731	40,630
Total [7]	70,846	442,372

Preprocessing

We put special emphasis on the preprocessing of the PL. The raw corpus data was reformatted by means of an XML Schema following the guidelines of the *Text Encoding Initiative* using TEI-P5 [5]. We automatically annotated the logical document structure, and linguistic categories by a Part-of-Speech tagger. For this purpose, a new flexible and extensible management system for linguistic data was established that is called *eLexicon*.

At present, the biggest part of the data stored in the *eLexicon* comes from three sources: The *AGFL Grammatica Latina* [2], a word form extraction from the *Perseus Hopper* [3], and a lexicon crawled from the *LemLat*-Service, using a corpus of Latin texts [4]. Additional resources were consulted for proper names of important persons and places.

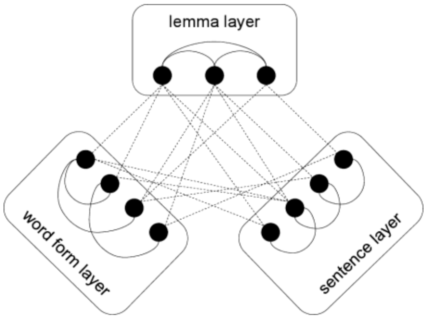

Figure 1: A three-level graph including a word form, lemma and sentence layer.

The resources were adapted to a consistent morphological format and the orthography was uniformed regarding the vocalic and consonantal usage of *i* and *u*, a subject even the most popular Latin lexica, the *Latin Dictionary* and the *Oxford Latin Dictionary* differ in.

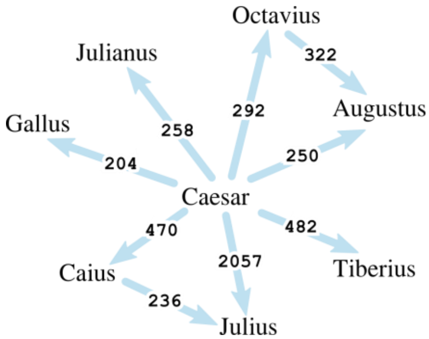

Figure 2: The lemma *Caesar* in the context of typical name-related usages (left neighbor).

Network Induction

In order to make the usage-based networking of word forms, lemmata and sentences accessible, we developed a framework for inducing so called *Three-Layer Networks* (L^{3}N) [10] from the PL (see <http://www.lexical-network.net>).

Generally speaking, an L^{3}N is a graph $G = (V,E)$ whose vertex set V is partitioned – by analogy to polypartite graphs – into non-empty disjunct subsets V_A, V_B, and V_C where the edge set E is additionally partitioned into six non-empty disjunct subsets E_{AB}, E_{AC}, E_{BC}, E_A, E_B and E_C so that any edge $\{x,y\} \in E_{XY}$ ends at vertices $x \in X$, $y \in Y$, $X \neq Y$, $X,Y \in \{A,B,C\}$, while any other edge $\{x,y\} \in E_X$ ends at vertices $x,y \in V_X$, $X \in \{A,B,C\}$. We speak of the subgraphs (V_A, E_A), (V_B, E_B) and (V_C, E_C) as (i) the *word form*, (ii) the *lemma* and (iii) the *sentence layer* of the L^3N. See Figure 1 for a visual depiction of such a three-level graph.

In order to span an L^3N by including word forms, lemmata and sentences, we need to account for significant co-occurrences on the lexical level as well as for sentence similarities on the syntactic level (of course, links of word forms or lemmata to sentences are trivially captured by their constituency relation). On the lexical level, we adopt the approach of Heyer. That is, we use *TinyCC 2.0* [11] to compute co-occurrence statistics: for two lexical items (i.e., lemmata or word forms) A and B that occur in a total of a and b sentences while they co-occur in k sentences, we compute their degree of lexical association by the following measure (n is the total number of sentences in the PL):

$$sig(A,B) = \begin{cases} \dfrac{\lambda - k * \ln\lambda + \ln k!}{\ln n} & : k \leq 10 \\[2ex] \dfrac{k(\ln k - \ln\lambda - 1)}{\ln n} & : k > 10 \end{cases}$$

$$\text{with } \lambda = \frac{ab}{n}$$

Table 3: neighbors of *Caesar* in the PL.

Left Neighbor	Right Neighbor
Julius (2056.95)	Augustus (950.05)
Tiberius (482.14)	Baronius (327.59)
Caius (470.19)	Egassius (209.9)
Octavianus (291.91)	Flavius (207.57)
Julianus (258.02)	creatus (111.2)
Augustus (249.99)	Dalmatas (105.42)

For any items A and B, for which $(k+1) / \lambda > 2.5$, we span an edge in the corresponding word form (or lemma) network and weight it by $sig(A,B)$.

Table 4 shows the number of vertices and edges in the resulting networks induced in this way. Note that the high order of the lemma network is due to the fact that not all word forms have been lemmatized. See Table 3 for the top 6 left and right neighbor collocations of *Caesar* as a word form. Figure 2 additionally shows the left neighbor collocations of *Caesar* as a graph.

The next step of L^3N induction concerns the spanning of the sentence network. Our idea is to link those sentences that manifest alike conceptualizations, that is, sentences, which have many relevant lexical constituents in common. In order to implement this approach we utilize the *idf* (*inverse document frequency*) measure in conjunction with a multiset representation of sentences: Let S_1, S_2 be two sentences that are represented by multisets S_1, S_2. Then,

$$\sigma(S_1, S_2) = \frac{\sum_{x \in S_1 \cap S_2} i(x)}{\sum_{x \in S_1} i(x) + \sum_{x \in S_2} i(x) - \sum_{x \in S_1 \cap S_2} i(x)}$$

their mutual significance is computed by

Note that $S_1 \cap S_2$ is the multiset intersection. $i(x)$ is the *idf* of x in the corresponding reference corpus. 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 $\sigma(S_1,S_2) = 0$.

Table 4 shows the order and size of the resulting networks. All in all, we experiment with an L^3N network of an overall set of 9,282,855 vertices and 174,470,929 edges. To the best of our knowledge this is the first study that integrates three levels of linguistic resolution by historical networks of such a size.

Table 4: network types, their order and size.

Type	#Vertices	#Edges
word forms	967,565	85,360,958
lemmata	839,111	52,137,832
sentences	7,476,179	36,972,139

Experiments

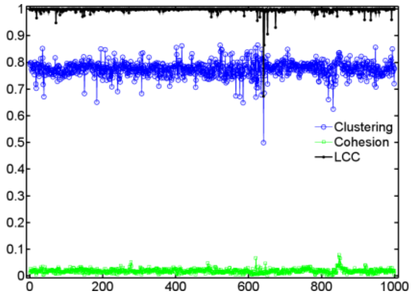

Figure 3: Cluster coefficient, fraction of vertices in the largest connected component and cohesion of 1,000 lemma networks derived from 1,000 texts from the PL.

We conduct several network experiments with the Latin L^3N induced according to the preceding section. As shown in Figure 3, we see, for example, that lemma networks basically retain a high cluster value irrespective of the document under consideration. The same is true for the cohesion of the networks, the size of their largest connected component and for the gamma coefficient of the power law fitted to their out-degree distribution. All in all this hints at linguistic networking according to scale-free networks, a pattern that is retained irrespective of the linguistic layer under consideration.

Conclusion

This paper presented an approach to making historical corpora accessible to network analysis. We have described several processing steps that are indispensable to reach this goal. This relates to the build-up of an appropriate lexicon as well as to the preprocessing of linguistic data so that network analyses come into reach that grasp more than a single linguistic layer. Based on that, we introduced the framework of multilevel networks to capture networking of lexical and syntactic units. Our analyses show a remarkably stable behavior of network characteristics over time. We also exemplified the local networking of single lexical units as a first attempt to capture relevant linguistic information of language change by example of the PL.

References and Notes

1. J.-P. Migne, *Patrologiae cursus completes: Series Latina* **1-221** (Cambridge: Chadwyck-Healey, 1844).

2. C. H. A. Koster, "Constructing a parser for Latin," *Proc. of the 6th CICLing*, Mexico City, Mexico (February 2005) pp. 48-59, <http://www.agfl.cs.ru.nl/lat/>.

3. D. A. Smith, J. A. Rydberg-Cox and G. R. Crane, "The Perseus Project: A Digital Library for the Humanities," *Literary and Linguistic Computing* **15**, No. 1 (2000), pp. 15-25, <http://www.perseus.tufts.edu/>.

4. M. Passarotti, "Development and perspectives of the Latin morphological analyser LEMLAT," *Linguistica Computazionale* **3** (2000), pp. 397-414, <http://webilc.ilc.cnr.it/~ruffolo/lemlat>.

5. L. Burnard, "New Tricks from an Old Dog: An Overview of TEI P5," *Digital Historical Corpora, Dagstuhl Seminar Proc.* (IBFI, Germany, 2006).

6. J. Klausenburger, *Grammaticalization. Studies in Latin and Romance Morphosyntax* (Amsterdam: John Benjamins, 2000).

7. Totals without redundancy.

8. C. Lehmann, "New Reflections on Grammaticalization and Lexicalization," *New Reflections on Grammaticalization* (Amsterdam: John Benjamins, 2002), pp. 1-18.

9. R. G. G. Coleman, "Poetic Diction, Poetic Discourse and the Poetic Register," *Aspects of the Language of Latin Poetry, Proceedings of the British academy*, **93** (1999), pp. 21-93.

10. A. Mehler, "Structural Similarities of Complex Networks", *Applied Artificial Intelligence* **22**, No. 7-8 (2008), pp. 619-683.

11. C. Biemann, U. Quasthoff, G. Heyer and F. Holz, "ASV Toolbox – A Modular Collection of Language Exploration Tools," *Proc. of the LREC* (2008).