Computational Historiography: Data Mining in a Century of Classics Journals

David Mimno

More than a century of modern Classical scholarship has created a vast archive of journal publications that is now becoming available online. Most of this work currently receives little, if any, attention. The collection is too large to be read by any single person and mostly not of sufficient interest to warrant traditional close reading. This paper presents computational methods for identifying patterns and testing hypotheses about Classics as a field. Such tools can help organize large collections, introduce younger scholars to the history of the field, and act as a "survey," identifying anomalies that can be explored using more traditional methods.

Categories and Subject Descriptors: H.4.0 [**Information Systems Applications**]: General

General Terms:

Additional Key Words and Phrases:

1. INTRODUCTION

Humanities scholarship has traditionally focused on the careful, detailed reading of small numbers of high-value texts. Over the past five to ten years, large-scale digitization projects have vastly increased the quantity of cultural heritage material, to scales well beyond the amount of text any single scholar could meaningfully process. This development raises a vital question: how, if at all, should the work of humanistic scholarship adapt to the presence of orders of magnitude more potential source material? From the perspective of traditional scholarship, little has changed: the fact that most of recorded human intellectual output is now accessible does not increase the ability of a scholar to read it. Clearly, any fundamental advances must come from the fact that this material is now available for computational processing. A million-book library can now be provided as input to computer programs. This paper explores the question of what those programs might do, taking inspiration from established work in statistical data analysis, information retrieval, and text mining, and how the output of such computational methods complements traditional scholarship.

1.1 Related Work

Moretti describes a similar concept of "distant reading," distinct from yet complementary to traditional close reading [Moretti 2005]. The quantitative study of collections of research literature is established in the sciences, but has seen relatively

little application in the humanities. The concept of algorithmic historiography is introduced by Garfield, Pudovkin and Istomin [Garfield et al. 2003] in the context of citation databases such as ISI's Web of Science. Hall, Jurafsky and Manning use textual analysis tools such as topic models to study collections of relatively recent papers in computational linguistics [Hall et al. 2008]. Blei and Lafferty apply topic modeling to 100 years of articles from a single journal, *Science* [Blei and Lafferty 2007]. There is also substantial work in creating "maps" of science [Börner et al. 2003; Klavans and Boyack 2006; 2009; Small 1999]. In this paper I identify some of the issues that make textual analysis for computational historiography in humanities collections particularly difficult and show through case studies how these issues might be addressed. It should be noted that the goal of this paper is not specifically to learn about the history of classical scholarship, but to demonstrate mature technologies and methodologies that can be used now by scholars to identify their own hypotheses and explore how the evidence in the collection relates to those questions.

1.2 Description of the Collection

The collection consists of text generated by optical character recognition from 24 journals in classical philology and archaeology, generously provided for research purposes by JStor. As the OCR output consisted of the entire journal, including letters, book reviews and indices, we began by identifying pages containing research articles, as identified in the metadata XML files provided by JStor.

Information about the journals is provided in Table I, in approximately the order of their inclusion in JStor. This table highlights some of the difficulties in analyzing the collection.

One of the key challenges is working with humanities scholarship in contrast with scientific research is the presence of multiple languages. Science tends to be published predominantly in English, and attention tends to be focused on recent work. The first three columns after the publication title describe the linguistic heterogeneity of the JStor classics corpus. Each page in the collection was classified as one of English, German, Latin or undetermined by counting the number of occurrences of the ten to fifteen most common words in each language. Although we found this method extremely robust, the results of this analysis are nevertheless noisy, for example due to extensive use of quotations. For many pages the predominant language could not be safely identified due to small numbers of words. Language classification does, however, provide a good high-level view of the contents of the collection. Most journals are predominantly English, but several contain substantial text in German (Hermes, Historia, and Mnemosyne). One journal (Mnemosyne) also contains significant numbers of pages written in Latin, largely in the 19th century.

Although it would certainly be possible to simply remove all text in German, doing so would remove a substantial amount of the corpus from the start, thus limiting the evidence that can be used to examine hypotheses about the history of classics over the past 150 years. In addition, the difference in content between English- and German-language scholarship is in itself an important and interesting question. Fortunately, the methods used in this paper are linguistically relatively simple, using "bag of words" representations that ignore word order and thus do

I: Description of publications. The left columns show the proportion of pages likely to be predominantly in one of three languages, not including pages in other languages and pages that could not be automatically classified with high confidence. The plot on the right shows number of pages published per year, with gray vertical lines at 1900, 1950, and 2000.

Title	% Eng	% Ger	% Lat	Pages	Pages per year (1850–2006)
Am. J. Arch.	82.9	1.2	0.2	94881	
Am. J. Philol.	89.6	0.8	0.9	74847	
Clas. J.	95.1	0.0	0.5	34412	
Clas. Philol.	89.2	0.9	0.7	52137	
Clas. Quar.	92.1	0.2	0.6	46216	
Clas. Rev.	92.0	0.1	1.3	42937	
Gr. and Rome	96.6	0.0	0.4	15802	
Hesperia	82.0	0.2	0.3	63725	
Phoenix	91.6	0.6	3.3	14677	
TAPA	89.5	0.8	1.4	6038	
T&P APA	90.4	0.3	0.9	49577	
TAPA	94.7	0.2	0.2	17552	
Harv. Stud.	88.6	1.2	3.6	25944	
Brittania	71.9	0.1	0.1	1517	
J. Rom. Stud.	82.5	0.3	0.8	45684	
Hesp. Suppl.	71.8	0.8	0.2	13361	
Ath. Ag.	64.9	0.0	0.0	24411	
Clas. World	93.6	0.2	0.2	18294	
Hermes	15.8	70.2	3.2	126983	
Historia	58.7	27.7	3.8	56478	
Mnem.	40.2	10.4	37.1	85086	
Phron.	85.4	5.1	2.6	22490	
Clas. Weekly	90.6	0.2	0.6	17739	

not require consideration of complex syntactic relationships. It is therefore valuable to consider simple, effective methods that can provide sufficient semantic alignment to support low-level word counting algorithms without needing to resort to complicated machine translation methods.

A second challenge is the temporal heterogeneity of the corpus. The figures to the right in Table I show the number of pages published per year. Although the collection goes back to the mid-19th century, most journals do not begin publishing until the 20th century, with the largest concentration of work in the second half of the 20th century. Journals also change their rate of publication. The most dramatic examples include Hermes, which ceases publication entirely in the late 1940s. More gradual changes can be seen for example in the American Journal of Archaeology, which significantly increased its volume of publication in the last quarter of the 20th century. Finally, some publications (Hesperia Supplements, the Athenian Agora) are actually monograph series, and therefore show extremely irregular and "bursty" publication patterns. These patterns of publication volume are critical to take into account. One of the goals of this paper is to identify changes in the nature of classical scholarship over the period covered by this collection. It is therefore necessary to use methods that distinguish changes in the underlying intellectual environment of the field from simple variations in the editorial agenda of the publications that happen to have been included in this collection.

OCR quality in the collection is generally fair to good for English and Latin text. German text is handled more poorly: umlauts are largely missing or rendered as *ii* and the *ß* character is usually rendered as *B* or *B3*. These errors are relatively easy to spot and fix algorithmically, especially for very common words. Numbers are often poorly recognized, especially the digits 1 and 0 (often the letters *I* and *o*, respectively). Greek is in the worst condition, uniformly mangled to the point of unusability.

2. METHODS

Before considering specific case studies, it will be helpful to review some standard tools in statistical text mining.

2.1 Representations

In order for computational methods to be applied to text collections, it is first necessary to represent text in a way that is understandable to the computer. The fundamental unit of text is the word, which we here define as a sequence of (unicode) letter characters. It is important to distinguish two uses of *word*: a word *type* is a distinct sequence of characters, equivalent to a dictionary headword or lemma; while a word *token* is a specific instance of a word type in a document. For example, the string "dog cat dog" contains three tokens, but only two types (*dog* and *cat*).

After a document has been *tokenized*, dropping spaces and punctuation, the next choice that must be made is whether to maintain word order. For some applications, such as detecting phrases or training language models that predict the next word given previous words, ignoring word order would be foolish. There are a surprising number of applications, however, for which word order adds little or no additional information, such as searching and text classification. Using the example from earlier, the string "dog cat dog" would produce the bag of words representation $\{dog = 2, cat = 1\}$. The input strings "dog dog cat" and "cat dog dog" would result in the same representation. The set of distinct word types is called the vocabulary. If the vocabulary has size V, we can assign each type w a unique integer ID from 1 to V. Finally, we can construct a column vector of length V with the number of times each word type appears in a particular document in the position for that type's ID. Continuing the previous example, if the vocabulary is, in order, $\{mouse, dog, horse, cat, pig\}$, the vector representation of "dog cat dog" would be $(0, 2, 0, 1, 0)^T$.

2.2 Vocabulary curation

The vocabulary used in text mining applications is generally a subset of the set of all distinct tokens. Due to the power-law characteristics of natural language, a small set of very common word types (often called "stopwords") make up a large part of the tokens in the corpus, while the great majority of distinct types occur very infrequently. It is often useful to remove both very frequent and very infrequent words from the vocabulary. Common words generally provide little information, especially if word order is dropped, but can overwhelm more important semantic words in analyses. Removing them can reduce the size of the data that must be analyzed by up to half while improving semantic coherence. Infrequent words, while often having the most specific meanings, can be difficult to perform meaningful

inferences on due to small sample sizes. Removing them can substantially reduce the number of parameters that must be fit and thus decrease the complexity of statistical models. A good heuristic for identifying such words is to remove those that occur in more than 5-10% of documents (most common) and those that occur fewer than 5-10 times in the entire corpus (least common). In this collection, however, it is also important to consider multiple languages, so separate stoplists for German, Latin, and English were constructed and concatenated. Even though Greek is not useful for analysis, it was still necessary to consider it in curating the vocabulary due to extremely common word fragments such as *ev, av, ov*, corresponding to noun suffixes.

Although it is sometimes customary to apply stemming algorithms to words, I have opted not to use such methods. No stemmer is perfect, and as a result they can often add ambiguity to text. They also tend to produce output that does not look like words and can therefore be confusing. I find that minor variations in words (such as pluralization) can be effectively handled with more sophisticated contextual tools such as topic models, even for highly inflected languages such as Latin and Greek.

2.3 Smoothed Distributions

As shown before, a document can be tokenized and transformed into a vector of word type counts $\boldsymbol{d}$. It is often useful to transform that vector of positive integers into a probability distribution over the vocabulary, that is, a vector of positive real numbers that sum to 1.0. The resulting distribution can be thought of as the probability that the "next" word token written in the document is of a particular type. This process is the foundation of successful methods in statistical information retrieval [Ponte and Croft 1998]. The *maximum likelihood* distribution $p(w|d)$ can be obtained by dividing each element d_w in the vector by the length of the document $\sum_w d_w$. This distribution will, of course, have zero probability of emitting any word that did not occur in the original document. We may prefer to ensure that all words have at least some (possibly very small) probability, both for modeling reasons (we wish to be conservative and avoid overfitting) and for mathematical reasons (calculations can become unstable in the presence of zeros). Methods that redistribute probability mass to avoid such problems are referred to as *smoothing*. A common method is Dirichlet smoothing [Zhai and Lafferty 2001], in which a vector $\boldsymbol{\alpha}$ is added to $\boldsymbol{d}$ before the vector is scaled to sum to 1.0. The probability of a word type is then $p(w|\boldsymbol{d}, \boldsymbol{\alpha}) = (d_w + \alpha_w)/(\sum_{w'} d_{w'} + \alpha_{w'})$. The relative difference between the length of the document and the sum of the α_w parameters controls the "strength" of the smoothing: if this sum is 100 and the document contains thousands of word tokens, the smoothed distribution will be very close to the maximum likelihood distribution. If the document has five tokens, the smoothed distribution will be close to the *prior* distribution $p(w|\boldsymbol{\alpha}) = \alpha_w / \sum_{w'} \alpha_{w'}$.

2.4 Divergence

Now that we have defined documents as probability distributions, we can consider measuring the *distance* between groups of words. One standard metric is Kullback-Leibler divergence. The KL divergence between two probability distributions $p(X)$

and $q(X)$

$$D_{KL}(p||q) = \sum_x p(X) \log \left(\frac{p(x)}{q(x)} \right). \quad (1)$$

KL divergence is not symmetric (in general $D_{KL}(p||q) \neq D_{KL}(q||p)$). It can also return infinite values if there exists a value x such that $p(X) \neq 0$ and $q(X) = 0$ because of the division. A related quantity is Jensen-Shannon distance, which is the average of the KL divergence between p and the element-wise mean of the distributions p and q and the KL divergence between q and the same mean distribution:

$$D_{JS}(p||q) = \sum_x \frac{1}{2} p(X) \log \left(\frac{2p(x)}{p(x) + q(x)} \right) + \frac{1}{2} q(X) \log \left(\frac{2q(x)}{p(x) + q(x)} \right). \quad (2)$$

2.5 Topic Modeling

Individual words are informative, but often are not as meaningful as small groups of related words. In different contexts, the same word type could have very different connotations. Statistical topic models such as latent Dirichlet allocation [Blei et al. 2003] attempt to identify *groups* of words that tend to occur together, while allowing words to appear in multiple groups. Modern statistical topic models are related to latent semantic analysis [Deerwester et al. 1990] and probabilistic latent semantic analysis [Hofmann 1999].

Topic modeling has recently been extended to corpora in multiple languages [Mimno et al. 2009]. These methods depend on the existence of an aligned training set consisting of sets of comparable documents. Pairs of comparable documents can actual direct translations, but can also be *topically* identical documents. Wikipedia articles with links between languages form a large and readily accessible corpus of comparable documents in a large number of languages. In the polylingual topic model, each topic is modeled as a set of distributions, one for each language's vocabulary. Each tuple of comparable documents (that is, one comparable document from each language) is modeled as having been generated by a single shared topic distribution, with the observed words being drawn from the hidden topics' distribution over the appropriate language.

3. CASE STUDY: VARIABILITY IN PHILOLOGY AND ARCHAEOLOGY

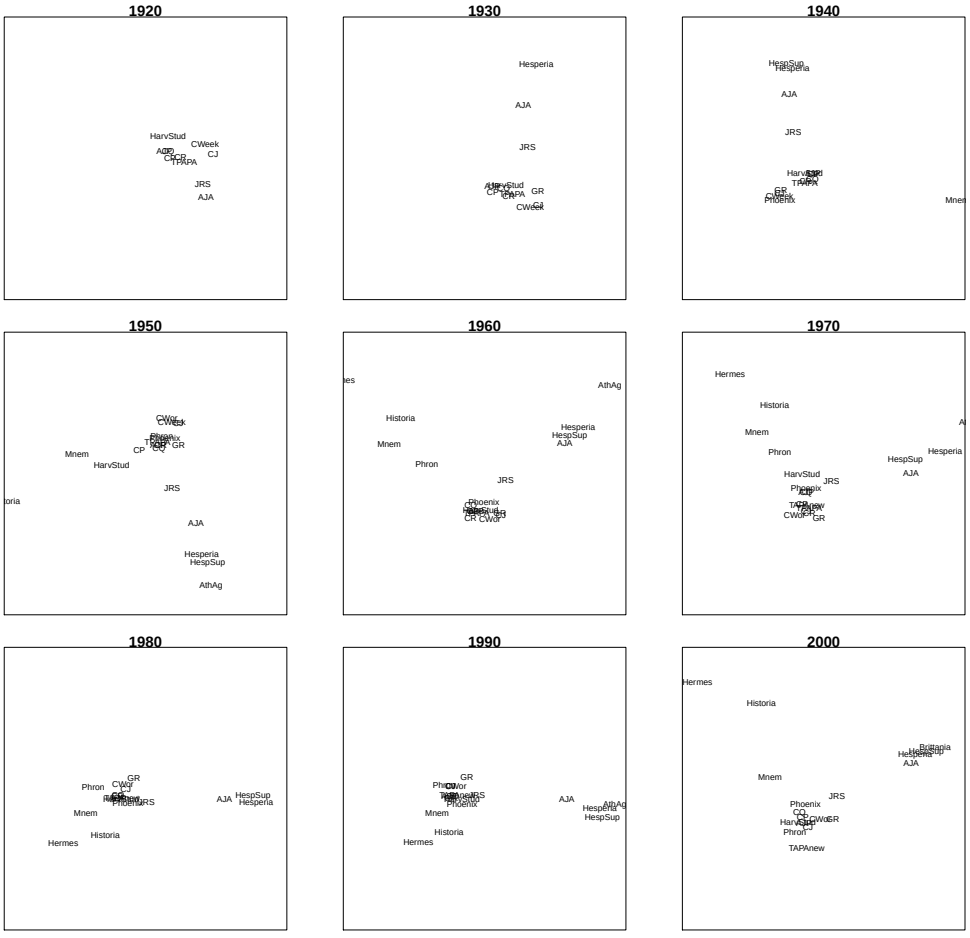
Classical studies is generally divided into two main branches, philology, or the study of texts including narrative histories, poetry and drama, and philosophy; and archaeology, the study of material culture. In this section I explore whether this view is backed up by the available data, how, if at all, it has changed over the last hundred years, and, more specifically, whether the *variability* of these fields has changed. My overall strategy is to start with the most general measurements and then move in to more specific analysis. Finally I will consider an analysis of the relative density of topics across languages, between scholarship in English and in German.

3.1 Journal similarity over time

To provide a quick overview of the relative similarities of journals over time, I divided the entire corpus of research articles into sections by journal and decade

and built bag-of-words representations. That is, I have one word count vector that combines all the articles published in AJA in the 1930's, another for all AJA articles published in the 1940's, and so forth. I next calculated the Jensen-Shannon distance between each pair of journals within a given decade. This procedure produces a matrix of dissimilarities, which can be projected into a 2-dimensional plane for visualization.[1] This procedure can be thought of as placing connections of varying elasticity between each point and then laying them out on a flat surface such that every connection is as relaxed as possible. Results are shown for the most recent nine decades in Table II. The x and y axes are not meaningful (only the distance between points is relevant), but the scale is the same for each plot.

II: Distances between journal-decade word vectors using Jensen-Shannon distance, projected to the 2D plane using multidimensional scaling.

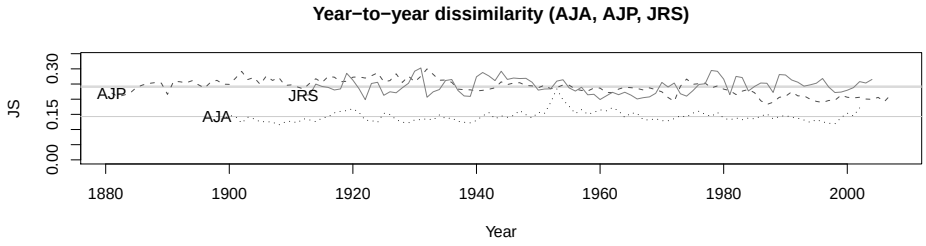


[1] using the R multidimensional scaling function `cmdscale`

There are three main clusters: the dense core of philology journals, the archaeology journals (AJA, Hesperia), and journals that are largely not English (Mnemosyne, Historia, and Hermes). In the earliest time period, archaeology (AJA) is less distinct from the philology journals, but the gap widens quickly and remains relatively constant. The Journal of Roman Studies (JRS) is perhaps the most interdisciplinary, sitting between the philology cluster and AJA for most of the 20th century. In the 1970's it appears to lose much of its archaeological content, but then seems to differentiate itself more in the most recent decade.

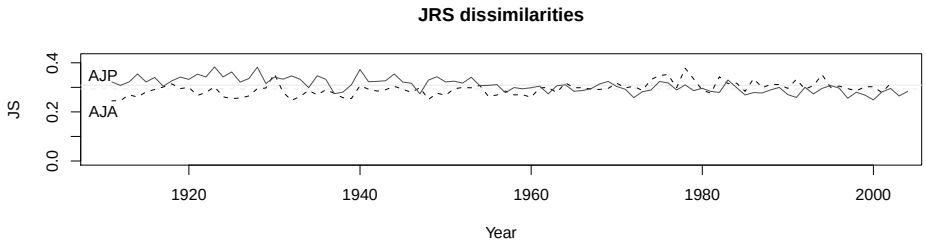
3.2 Variability of journal vocabularies

The previous plot shows the magnitude of the relative differences between word distributions of different journals. We can also look at the variability of each journal itself. Using the same methodology as the previous section, I defined word distributions by combining all the words published in a given journal in a given year and calculated the dissimilarity between distributions for each pair of subsequent years for each journal. The resulting time series of year-to-year dissimilarities is shown in Figure 1. Both AJP and JRS are consistently more variable than the archaeological journal AJA. The mean dissimilarity for each journal is shown with a horizontal gray line. The means for the AJP and JRS are indistinguishable. Interestingly, since the mid-1980's AJP has been in a period of historically low variability. Although it has never been as consistent year-to-year as the archaeology journal, it has become more less variable.



1: The Jensen-Shannon dissimilarity between word distributions for each available pair of subsequent years for three journals, for example $D_{JS}(JRS_{1951}||JRS_{1952})$. The mean dissimilarity is shown with a gray line.

This type of analysis can help to explore questions raised by the previous section. For example, is the Journal of Roman Studies becoming more philological? Figure 2 shows the dissimilarity between the word count vector for JRS in each year and the word count vectors for two prominent journals (AJA and AJP) in the same year. In the earlier half of the 20th century, JRS is consistently more similar to the archaeological journal (AJA) than the philological journal (AJP), although it is always fairly distinct from both. After about 1970, JRS is consistently closer to the philological journal than the archaeological journal, although the difference is not as great.



2: The Jensen-Shannon dissimilarity between word distributions for each year between JRS and two other journals, AJA and AJP, for example $D_{JS}(JRS_{1951}||AJP_{1951})$. The mean dissimilarity is shown with a gray line.

So far, this analysis has been entirely quantitative. We can also directly examine the word count vectors. It is typical, however, to weight word counts by some measure of the significance of words. The standard inverse document frequency (IDF) weighting is given by $IDF(w) = \log(D/DF(w))$, where D is the number of documents in the corpus and $DF(w)$ is the number of documents containing one or more tokens of type w. Table III shows the most frequent words by decade in JRS, weighted by the number of times the word occurs multiplied by the word's IDF weight. As we might expect, words for physical descriptions of places (*ft, south, wall, road, fort*) are more prominent in earlier years, while textual references (*op cit*) and historical words (*emperor, history, public*) are more common in recent decades.

3.3 Topic variation

Unfortunately, it is difficult to interpret trends by simply looking at the most distinctive words for an entire year. The top words for the 1970's, for example, seem to reflect an increase in publication in French, which does not necessarily imply any change in the semantic content of the journal. This particular issue (the presence of multiple languages) is not difficult to address in itself, but it is indicative of the larger problem: the overall word distribution of a given year is the result of a combination of many possibly independent trends. It would be preferable to distinguish the effect of individual trends so as to explore them individually, so that the relative proportion of French vs. English can be studied separately from the shift from physical descriptions of sites and artifacts to literary and historical analysis.

One method of disentangling the various trends in long-term collections of scholarly literature would be to ask trained researchers to tag the entire corpus with an agreed-upon set of labels. This approach, however, would be prohibitively expensive. In addition, getting scholars to come to consensus beforehand on what labels are interesting and by which criteria they should be applied is also likely to be difficult.

Statistical topic models provide a compromise between full manual tagging and naïve word counting. Topic models cannot match the understanding of a human annotator, but they can be applied to large quantities of text quickly (collections in the range of tens of millions of words can be processed in a few hours). In addition,

III: Most frequent words (with IDF weighting) by decade in JRS.

1911–1919	rome caesar cicero coins fig inscription lex law plate north province head senate year probably under war feet mommsen catiline obv south temple antioch
1920–1929	wall ft plate road britain coins east inscription fig north west stone province mr rome south date name legions hadrian side under galatia period
1930–1939	ft wall pl mr rome caesar coins britain date antony north east fig south legions period year trajan iv fort dio inscription augustus ditch
1940–1949	ft rome pl fig cil type coins mr wall north east cit jrs op ware road date west emperor arch vii senate south britain fort
1950–1959	ft south east fort wall north west jrs ditch fig mr road building street pl side rampart livy site cicero cil rome britain gate
1960–1969	ft rome jrs fort mr britain building lex ditch cil emperor dio vi polybius senate excavation cicero wall law iv fig information livy timber
1970–1979	la les vac rome cicero le emperor law cit lex iv senate op re vi cil augustus ils province imperial date dio pliny edict
1980–1989	imperial law rome emperor ch empire lex cit la senate augustus public city cicero op military under family land period le gibbon per case les
1990–1999	op cit rome empire la imperial augustus constantine gibbon history inscriptions city late iv ovid emperor propertius idem les per pay period le land portrait
2000–2004	op cit la rome imperial plancus world empire martial history population social en eds book pollio late mela caesar serapeum art cicero literary political

they are fully data-driven, and therefore not dependent on the perspectives and particular experiences of individual scholars.

In the context of the analysis of JRS, topic modeling can provide addition clarity relative to simple word counting. In this method, each word in each document is probabilistically assigned[2] to one of T word groups or "topics", such that each document contains relatively few topics and each topic contains relatively few distinct word types. In practice, assignments of topics that maximize these criteria are close to human understandings of the underlying concepts and linguistic categories in the corpus. Training is through iterative approximate inference, in this case Gibbs sampling as implemented in the MALLET toolkit.[3]

One common way to represent such topics is to count the number of words of each distinct type assigned to the topic and display a list of word types in descending order by count, with the most probable word first, and so forth. These lists of most probable words, along with time series plots of words assigned per year, are shown in Table IV for a selection of topics from a model with $T = 150$ trained on research articles from JRS. Topics are ordered by their mean year score: for each word token w assigned to topic t, let y_w be the year that word w was published. The mean year score is the mean of y_w for all words assigned to a given topic. The mean year for each topic is shown with a vertical blue line. The gray region is

[2]this assignment can be represented by a single topic or by a distribution over topics.
[3]`http://mallet.cs.umass.edu`

proportional to the number of words in each topic published in a given year. The scale of the y-axis is constant for all plots. Consistent with previous findings from word-distribution distance and decade-word-counts, topics with early mean year scores (towards the top of the table) tend to be words used in physical descriptions of sites and objects. Topics with the latest mean years tend to be more literary and historical, with poetry (e.g. Ovid and Horace) and social history showing the most substantial gains.

Although such topics are useful for browsing and gathering quick overviews of the contents and trends of a corpus, they can also be used to facilitate searching and the identification of articles for closer reading. Table V shows top-ranked articles for three topics, including the single earliest and latest topics by mean year. Articles are sorted by the number of words assigned to the topic in the article. Note that this ranking function is biased towards longer documents, but produces results that are consistent with the topic. In practice, a combination of many ranking functions should be used. The "early" topic, which includes words describing the plans of tombs and buildings, is generally represented by publications from the early 20th century, although there are highly ranked articles from as late as 2004 (not shown for space considerations). The middle topic is interesting in that it appears to be one of the last active topics involving physical descriptions. The articles that are most responsible for the prominence of this topic are a series of aerial surveys of Roman sites in Britain that begin in the 1940's and continue until the late 1970's. Finally the "late" topic, involving elegaic poetry by Ovid and Propertius, includes predominantly articles from the last decade, although significant work is included from as early as the 1970's.

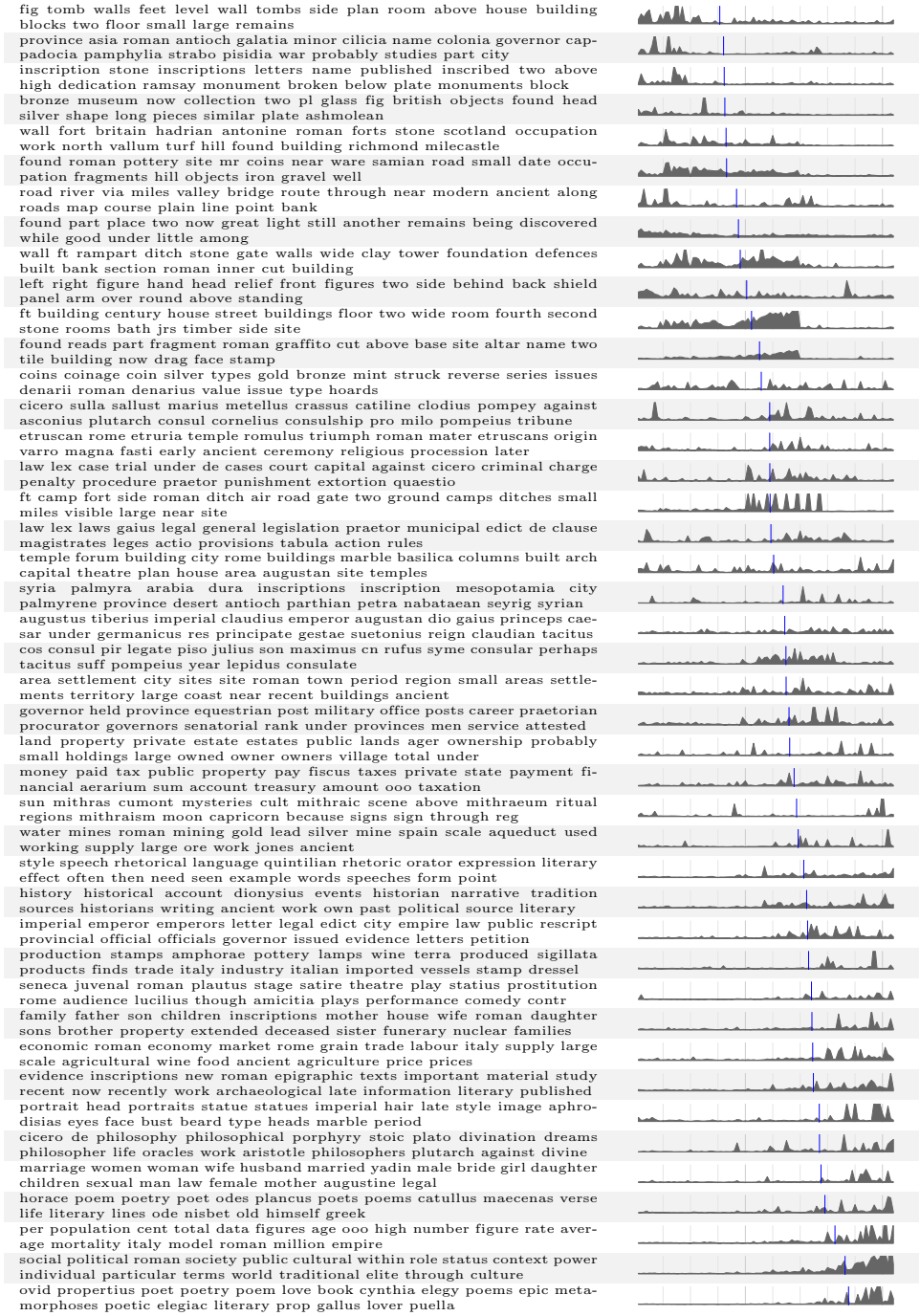
3.4 Topic modeling over multiple languages

One of the most difficult questions in the analysis of scholarship in the humanities is the comparison of work that is published in different languages. Scientific research, at least in the 20th century, is almost entirely in English, and thus much work in computational analysis of scholarship has not touched on this area. At the same time, it is common to hear general descriptions of national or language-specific "schools" in humanities research. Particular nationalities are associated with distinct interests and methodologies. Evaluating such claims through close reading requires not only a deep knowledge of many different subfields, but also strong linguistic ability in all relevant languages, sufficient to read large quantities of text.

As shown previously, topic modeling provides a useful method for identifying and disentangling trends in research literature. Standard topic models represent topics as a single distribution over a vocabulary. In previous work, it has been shown that topic models can effectively learn topics that are represented as *sets* of distributions, one over each of L vocabularies, corresponding to L different languages [Mimno et al. 2009].

Previous experiments with such multi-lingual topic models have depended on the existence of a training collection that contains sets of articles, one in each language, that are either parallel (exact transations) or comparable, for example the English and German Wikipedia articles for *Barack Obama*, which are not translations, but are about the same person. It was also shown that a collection of non-comparable

IV: Selected topics from JRS ($T = 150$), in order by mean publication year (blue vertical line), 1911–2004. Physical descriptions of sites dominate the earlier 20th century, while recent decades have focused on social and economic history.



V: Articles ranked by the number of words in each article assigned to topic t, for three topics.

(a) **fig tomb walls feet level wall tombs side plan room above house building blocks two floor**

3602	Gino Rosi. Sepulchral Architecture as Illustrated by the Rock Facades of Central Etruria: Part I. (1925) pp. 1-59.
2975	Gino Rosi. Sepulchral Architecture as Illustrated by the Rock Facades of Central Etruria. Part II. (1927) pp. 59-96.
2514	Thomas Ashby, G. McN. Rushforth. Roman Malta. (1915) pp. 23-80.
2456	S. Rowland Pierce. The Mausoleum of Hadrian and the Pons Aelius. (1925) pp. 75-103.
1872	Esther Boise van Deman. The Sullan Forum. (1922) pp. 1-31.
1700	T. Ashby, R. A. L. Fell. The Via Flaminia. (1921) pp. 125-190.
1411	Thomas Ashby. Recent Discoveries at Ostia. (1912) pp. 153-194.
1226	O. L. Richmond. The Augustan Palatium. (1914) pp. 193-226.
1060	Thomas Ashby. The Bodleian MS. of Pirro Ligorio. (1919) pp. 170-201.
975	A. W. Van Buren. The Technique of Stucco Ceilings at Pompeii. (1924) pp. 112-122.
894	G. Lugli. Recent Archaeological Discoveries in Rome and Italy. (1946) pp. 1-17.

(b) **ft camp fort side roman ditch air road gate two ground camps ditches small miles visible**

5888	J. K. St Joseph. Air Reconnaissance in Roman Britain, 1973-76. (1977) pp. 125-161.
5469	J. K. St Joseph. Air Reconnaissance in Britain, 1969-72. (1973) pp. 214-246.
4545	J. K. St Joseph. Air Reconnaissance in Britain, 1965-68. (1969) pp. 104-128.
3573	J. K. St. Joseph. Air Reconnaissance in Britain, 1961-64. (1965) pp. 74-89.
3476	J. K. St. Joseph. Air Reconnaissance in Britain, 1958-1960. (1961) pp. 119-135.
3409	J. K. St. Joseph. Air Reconnaissance in Britain, 1955-7. (1958) pp. 86-101.
3365	J. K. St. Joseph. Air Reconnaissance of Southern Britain. (1953) pp. 81-97.
3342	J. K. St. Joseph. Air Reconnaissance of North Britain. (1951) pp. 52-65.
1916	J. K. St. Joseph. Air Reconnaissance in Britain, 1951-5. (1955) pp. 82-91.
759	I. A. Richmond. Recent Discoveries in Roman Britain from the Air and in the Field. (1943) pp. 45-54.
621	D. R. Wilson, R. P. Wright. Roman Britain in 1968: I. Sites Explored: II. Inscriptions. (1969) pp. 198-246.

(c) **ovid propertius poet poetry poem love book cynthia elegy poems epic metamorphoses**

3215	K. Sara Myers. The Poet and the Procuress: The Lena in Latin Love Elegy. (1996) pp. 1-21.
3199	Jeri Blair Debrohun. Redressing Elegy's Puella: Propertius IV and the Rhetoric of Fashion. (1994) pp. 41-63.
2320	Llewelyn Morgan. Child's Play: Ovid and His Critics. (2003) pp. 66-91.
2195	Maria Wyke. Written Women: Propertius' Scripta Puella. (1987) pp. 47-61.
2106	Sara Myers. The Metamorphosis of a Poet: Recent Work on Ovid. (1999) pp. 190-204.
2011	Bruce Gibson. Ovid on Reading: Reading Ovid. Reception in Ovid Tristia II. (1999) pp. 19-37.
1846	R. O. A. M. Lyne. Propertius 2.10 and 11 and the Structure of Books '2A' and '2B'. (1998) pp. 21-36.
1704	A. R. Sharrock. Womanufacture. (1991) pp. 36-49.
1614	Monica R. Gale. Propertius 2.7: Militia Amoris and the Ironies of Elegy. (1997) pp. 77-91.
1467	Michael Comber. A Book Made New: Reading Propertius Reading Pound. A Study in Reception. (1998) pp. 37-55.
1102	R. D. Anderson, P. J. Parsons, R. G. M. Nisbet. Elegiacs by Gallus from Qasr Ibrim. (1979) pp. 125-155.

documents, that is, documents without parallel or comparable documents in any other language, can still be used in aligned topic models so long as the collection is combined with a sufficiently large "glue" corpus of topically similar comparable documents.

In this section I train a bilingual topic model in English and German using aligned WIkipedia articles and apply it to articles from the American Journal of Philology (AJP) and Hermes, two philological journals with long histories and relatively constant rates of publication, except for a gap in Hermes following WWII. Using this model I then plot the relative concentrations of particular topics in English and German scholarship.

3.4.1 *Collecting comparable articles from Wikipedia.* Wikipedia makes sense as an initial source for comparable documents. Although individual factual statements in the online encyclopedia should be taken with skepticism, we can be confident that it is sufficient for training coarse bag-of-words representations. Wikipedia's primary attraction, however, is its ready availability for computational use, as the entire collection can be downloaded openly.[4] It is also relatively easy to identify articles in different languages that link to each other as being "about" the same entity, using cross-language links. The difficulty then becomes identifying pairs of English and German articles that are relevant to the philological corpus.

A first attempt was based on categories assigned to articles. I began with a small set of "seed" articles, and then alternated between gathering the most common categories assigned to those articles and gathering articles assigned to those categories. This method resulted in large numbers of relevant articles, but also easily "escaped" into related but irrelevant categories, such as descriptions of geographic regions and pages for modern Greek sports teams. This collection produced topic models with large numbers of topics that were not useful.

A second corpus was constructed by directly measuring the similarity of the words in documents to the "target" corpus of scholarly articles. For each word w, I measured the pointwise KL divergence between its probability in the target corpus (AJP or Hermes) and the "background" corpus consisting of all of Wikipedia: $p(w \mid \mathcal{C}_{target}) \log \frac{P(w|\mathcal{C}_{target})}{P(w|\mathcal{C}_{backgr})}$. For each Wikipedia article I then calculated the sum over all words in the article of that word's pointwise KL divergence and ranked articles by that score and selected all articles above a certain threshold. This procedure resulted in a much more consistently relevant corpus that matches the target corpus. The main source of irrelevant articles was lists of species in particular taxonomic groups by their scientific names, which contain Latin words and names that are also found in history, mythology and literature. Removing articles that contained more than 90% capitalized words was effective in eliminating most of these lists.

In order to evaluate the fit of the proposed comparable corpus to the target collection, we can measure the most common words in the JStor articles that do not occur, or that occur rarely, in the Wikipedia collection. For the German articles, the most common of these words are almost entirely OCR errors due to dropped or poorly handled umlauts (*dafi, konnen, fuir, wahrend, zunachst, fiber, namlich,*

[4] `http://download.wikimedia.org/`

uiber, lsst). These words are generally not useful for modeling anyway. The most prominent content word that is missed is *Wilamowitz*, a noted scholar. On the English side, the most common missing words are Latin (*modo, saepe, ipsa, illi*), scholarly abbreviations (*suppl, verg, cols, prop, plaut*), and author names (*fraenkel, meritt, lindsay, duckworth*).

This corpus resulted in topic models that were much more focused on classical philology than the earlier category-based corpus. Unfortunately, it was not large enough relative to the target corpus of scholarly articles to produce aligned topics over the entire collection: topics were individually coherent within each language, but did not match up between languages. As a result, rather than training the model over all documents, using both the comparable corpus and the unaligned corpus of articles, I trained the model only on the comparable corpus and then inferred topics for each article independently given the trained topic-word distributions. This procedure has been shown to be highly accurate as long as the "testing" corpus has a similar vocabulary to the training corpus [Yao et al. 2009], which is true in this case by construction. Again, I used the topic inference code in the Mallet toolkit.

The output of this topic inference procedure is a distribution over topics for each new document d from the corpus of articles. To verify that the inferred topic distributions are consistent with the original training corpus, it is possible to construct ranked lists of words from the word vectors of the documents $\boldsymbol{w}_d$ and the topic distributions, $P(t|d)$, by multiplying the number of times a word occurs in a document by the probability of the topic in the document.

$$score(w,t) = \sum_d count(w,d)P(t|d) \tag{3}$$

If the inference procedure is correctly assigning documents to topics, ranking words by their score in a particular topic should produce a list of words that is similar to the original topic word distributions. Results are shown in Table VI. The words are similar between the original topics and the inferred topics, with the exception that specific author names (Catullus, Horace, Thucydides, Livy) tend to appear more prominently in the list of words derived from scholarly articles than in the topics derived from Wikipedia. Note that in the last topic, the model has successfully identified articles with the words *romer* and *romischen*, even though these words lack the umlaut seen in the original topics (e.g. *römer, römische*), and are therefore, as far as the algorithm knows, entirely unrelated words.

3.4.2 *Comparing German and English philology.* The resulting topic concentrations over the 20th century are shown in Table VII for topics selected from a model with $T = 150$. The first column contains the most probable words in descending order in the topic's distribution over English words. The second column shows the most probable words in descending order in the topic's distribution over German words. The most probable words are frequently either direct translations (with small morphological variations) or even identical strings (*ovid, cicero*). The plot on the right shows the relative number of words in each topic published per year in English (above the axis, in dark gray) and German (below the axis, in lighter gray). Note that some plots are truncated in order to maintain a constant scale for the

VI: Comparison between original topics trained on comparable Wikipedia documents and "topic-weighted" words derived from document-topic distributions inferred for AJP and Hermes articles based on that topic model.

Original topics	Words weighted by inferred topics
poet poems poetry poem poets epic lyric literary life verse collection scholars composed famous epigram included literature horace lines	horace poem poet poetry catullus poems lines odes ovid love poets epic book literary poetic life elegy satire vergil lyric
dichter gedicht gedichte dichtung waren gilt jh später dessen lyrik form versen behandelt deutsche sänger epigramm gedichten vorbild elegie	dichter horaz gedicht vergil properz verse dichtung catull gedichte kallimachos form dichters elegie ovid tibull liebe vers buch epigramm epos
athens athenian bce athenians greece tyrant general peloponnesian city archon plutarch opposed pericles democracy cleisthenes solon thucydides statesman alcibiades	athens athenian thucydides athenians political democracy war pericles city year herodotus democratic lysias evidence plutarch sparta men aristophanes history years
athen athener athenischen herrschaft athens seines peloponnesischen perikles demokratie tyrann korinth attischen politiker attika stammte zahlreiche stadt tode kleisthenes	athen thukydides athener athens athenischen demokratie perikles plutarch stadt rede attischen herodot alkibiades sparta poseidonios hippias geschichte herrschaft politik themistokles
love sexual desire beauty well because aphrodite character relationship eros another whose describes trying part different door parents sex	love lover catullus poem eros apuleius desire man poet erotic passion beauty nature sexual another socrates life rather poetry lines
liebe gilt menschen beziehung sexuelle zurück immer eher sehen geprägt personifikation hauptsächlich zuneigung hervor freundschaft eros versucht bezeichnen sexuellen	liebe apuleius lust amor menschen eros narcissus gen sokrates beziehung leidenschaft ovid platon dichter properz sehen leben echo rede kein
rome romans city etruscan people jupiter livy italy early forum probably via seems old became oscan juno italic latium	rome livy city augustus romans jupiter early italy aeneas war varro history temple etruscan book forum evidence people vergil villa
rom römische stadt roms wurden außerhalb jupiter etruskischen villa livius via forum römern römer lateinisch später kapitol italien etrusker	rom livius romischen stadt roms varro latinischen romer polybios darstellung silius geschichte namen jahr augustus cicero daf dabei wahrend cil

y-axis. There are several clear differences between topic distributions. In its earliest years, AJP contains significant text on historical linguistics and Indo-European studies. This topic disappears by the early 20th century, and is never represented significantly in the German literature. German articles have substantially more text on rhetoric and law over the entire time period. English philological scholarship has seen a substantial increase in attention to gender and sexuality, which has largely not been mirrored in this selection of German scholarship from Hermes. Several historical topics (Persia, Athens, Sparta, Imperial Rome) show a similar pattern: German scholarship in Hermes is highly variable from year to year, but has a steady overall trend, while interest in history — particularly Athenian history — increases significantly in AJP, beginning around 1935. Both English and German articles show a strong but varying interest in poetry. In German there is a substantial increase after publication resumes following the war until around 1980, when both languages publish very little on poetry and poets. Poetry recov-

ers, however, becomes extremely common in the late 1980's and early 1990's, and is currently seeing substantial interest in both languages.

4. CONCLUSIONS

In this paper I have demonstrated that it is possible to use automated language processing methods to identify trends in large collections of complex secondary literature. The tools demonstrated use simple bag-of-words representations. Although such representations throw away substantial amounts of information contained in syntactic relationships, they are also simple, robust, and easy to extend in order to accommodate the special characteristics of the collection, such as multiple languages and difficult OCR.

In addition to the technical details, it is also important to consider why such tools are useful to scholars. The observations made using automated analysis are often not surprising to experienced scholars. and other experts in the particular field. It would, indeed, be ridiculous to propose that the "distant reading" analysis presented here can be substituted for decades of close reading. It would also, however, be foolish to dismiss the advantages of quantitative methods. There are at least three broad areas that can benefit all researchers:

(1) Overviews for students and cross-disciplinary researchers. Experienced researchers are generally familiar with the broad trends of their specific field of study. Even in a fairly narrow domain such as classics, however, there is too much material for even senior researchers to fully understand the entire space. Students, younger researchers and members of the public, in contrast, can benefit substantially from the additional contextual information provided by analyzing entire research collections and reach a level of understanding more quickly.

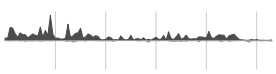
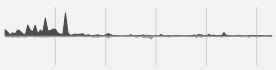
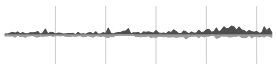
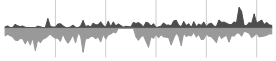
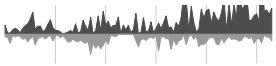
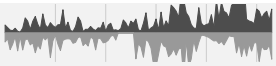
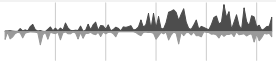
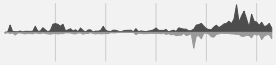
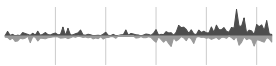
(2) Quantitative measurement of intuitions. The history of an intellectual domain is often phrased as a narrative: areas of study rise and fall, "schools" form and diverge, nations have particular characteristics. Using tools like those described in this paper, such stories can be framed as hypotheses and tested.

(3) Support for close reading. Perhaps most importantly, it is not necessary to see distant reading and close reading as opposing and mutually inconsistent methodologies. In order for close reading to be useful, it is necessary to know what one should read. Automated tools can assist in several ways. One use is as a survey tool, identifying anomalies and unexpected phenomena that warrant closer examination. Another use is in supporting more thorough searches for relevant material. Keyword searches and searches of particular journals can miss substantial quantities of material without giving any indication to the user. Models that provide richer representations of the ideas, topics, and vocabulary of a field can suggest broader ranges of potentially relevant articles. Finally, close reading depends on an implicit understanding of the context of a work.

Acknowledgments

This work was funded by the Andrew W. Mellon foundation. All data was provided by JStor.

VII: Aligned topics in English (AJP), in dark gray extending upwards, and German (Hermes), in light gray extending downwards. Lines are at 1920, 1940, 1960, 1980 and 2000.

case languages language example nominative english genitive means cases object dative subject grammatical accusative article finnish like noun feminine	sprachen sprache kasus beispiel fall akkusativ haus genitiv grammatik deutschen dativ latein beispiele kennt deutsch person ausdrücken präposition regel	
sanskrit self vedic mind buddhist three nature buddhism hindu type means through buddha reality meditation practice veda seeing compound	sanskrit drei vier buddhismus indischen lehre bedeutet prakriti purusha dukkha wesen hinduismus zustand buddhistischen buddha meditation yoga	
linguistics context words use specific type example particular meanings linguistic same terms different lexical analysis sound function literal within	bedeutung zwischen beispiel verschiedene linguistik also beispielsweise wort verschiedenen zeichen lassen wörtern kontext sprachwissenschaft englisch gegenstand siehe unterscheidung spricht	
rhetoric demosthenes cicero against speeches isocrates speech orator others style public along should teacher important private orators ten funeral	cicero rede rhetorik reden redner allem ersten demosthenes ciceros bedeutung tod prozess seinem letzte wobei damit rhetor auseinandersetzung meidias	
law legal court case criminal laws under right rule principle will act crime civil cases decision parties judge courts	recht gesetz wenn also lat law keine ohne wegen liegt tat non sog strafe gemäß deutschen grundsatz vertrag deutschland	
she daughter wife married mother sister marriage husband child children became queen gave bore woman named father birth herself	tochter mutter ihr frau schwester vater ihrem ihre ihrer ihres heiratete ehe gattin kind verheiratet ihren kinder ehefrau geboren	
women woman men man female young male considered famous seen than sexual feminine beautiful role make reports sex parallel	frauen männer gehört mädchen männlichen frau anderen personen jungen junge ihre weiblichen mann männern bestand tritt bringen weiblicher männliche	
love sexual desire beauty well because aphrodite character relationship eros another whose describes trying part different door parents sex	liebe gilt menschen beziehung sexuelle zurück immer eher sehen geprägt personifikation hauptsächlich zuneigung hervor freundschaft eros versucht bezeichnen sexuellen	
play comedy plays character probably sophocles tragedy out drama euripides dramatic aeschylus tragic lost comic performed took characters plautus	tragödie antike griech griechische haben griechisch komödie schon häufig euripides stück neben sophokles darstellung dichters theater allem erkennen dichter	
persian herodotus greeks asia cyrus persia minor against darius brother caria lydia persians sent under revolt empire xerxes part	könig während persischen herodot seinem später konnte sogar perser persische quellen aufstand feldzug dort tod kleinasien jahrhunderts karien dareios	
athens athenian bce athenians greece tyrant general peloponnesian city archon plutarch opposed pericles democracy cleisthenes solon thucydides statesman alcibiades	athen athener athenischen herrschaft athens seines peloponnesischen perikles demokratie tyrann korinth attischen politiker attika stammte zahlreiche stadt tode kleisthenes	
sparta spartan war states died men cities although sent greece spartans did cyprus kings fleet salamis evagoras battle allowed	sparta doch worden flotte dort spartanischen regierung jahr ganz haus blieb zypern ermordung galt leben spartaner könige nachdem eroberte	
emperor tiberius claudius augustus nero tacitus elder died second antonia before suetonius reign agrippina domitius younger drusus historians vespasian	jahr claudius tiberius ihn kaiser nero augustus kaisers ließ seinen ebenfalls nachfolger tod jahre ersten tacitus agrippina drusus herodes	
poet poems poetry poem poets epic lyric literary life verse collection scholars composed famous epigram included literatur e horace lines	dichter gedicht gedichte dichtung waren gilt jh später dessen lyrik form versen behandelt deutsche s änger epigramm gedichten vorbild elegie	
virgil aeneas aeneid whom named river hero ixion aegina upon founding against mount book turnus accounts pan aeacus off	ihn andere aeneas vergil könig aeneis da land peleus frau aphrodite verwandelt kentaur thetis schlacht hilfe eigenen pan aufgenommen	
story narrative events written literature novel stories literary fiction tale characters point character narrator audience plot prose described film	literatur erzählung geschichte handelt darin handlung darstellung griechische roman deren englischen beschreibt leser erzählt häufig prosa mythen erzählungen film	
ovid myth times origin justice even metamorphoses book appears female through others myths instead represented mistress power bird greece	genannt ovid name mythos verschiedenen griechisch lateinisch namens metamorphosen dargestellt gerechtigkeit darstellungen allem gebracht römische taucht entsprechend belegt recht	

REFERENCES

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Blei, D. M. and Lafferty, J. D. 2007. A correlated topic model of *Science*. *AAS 1,* 1, 17–35.

Börner, K., Chen, C., and Boyack, K. W. 2003. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 179–255.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. *JASIS 41,* 6, 391–407.

Garfield, E., Pudovkin, A., and Istomin, V. 2003. Why do we need algorithmic historiography? *JASIST 54,* 5, 400–412.

Hall, D., Jurafsky, D., and Manning, C. D. 2008. Studying the history of ideas using topic models. In *EMNLP*. 363–371.

Hofmann, T. 1999. Probilistic latent semantic analysis. In *UAI*.

Klavans, R. and Boyack, K. W. 2006. Identifying a better measure of relatedness for mapping science. *JASIST 57,* 2, 251–263.

Klavans, R. and Boyack, K. W. 2009. Toward a consensus map of science. *JASIST 60,* 3, 455–476.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., and McCallum, A. 2009. Polylingual topic models. In *EMNLP*.

Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. In *SIGIR*.

Small, H. 1999. Visualizing science by citation mapping. *JASIS 50,* 9, 799–813.

Yao, L., Mimno, D., and McCallum, A. 2009. Efficient methods for topic model inference on streaming document collections. In *KDD*.

Zhai, C. and Lafferty, J. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*.