Evolution, Cognition, and the History of Religion: a New Synthesis

Festschrift in Honour of Armin W. Geertz

Edited by

Anders Klostergaard Petersen
Ingvild Sælid Gilhus
Luther H. Martin
Jeppe Sinding Jensen
Jesper Sørensen

BRILL

LEIDEN | BOSTON

Contents

PART 3
Cognition

CHAPTER 14

Mind the Text: Traces of Mental States in Unstructured Historical Data

Kristoffer Laigaard Nielbo

We are currently witnessing a disruption in the empirical foundations of the social sciences and humanities (henceforth SSH). Digitization, that is, the proliferation in digitized and born-digital data, drives this disruption by bypassing traditional cultures and organizational patterns of SSH research (cf. Zeleny 2012). Traditional distinctions between qualitative-quantitative and interpretive-explanatory research are challenged by data mining of so-called unstructured data[1], which utilizes the full range of research positions in its collaborative workflow. A recent trend in software design and informatics focuses on the development of technologies that can adaptively integrate knowledge from users with expertise in traditional research domains (De Roure and Goble 2009; Rehm et al. 2014; Schobel et al. 2016). Although not the primary focus area, SSH is rich with domains that can benefit from technologies for extraction of high quality information from unstructured digital data. Historical research is a subarea of the SSH domains, spanning history proper, history of ideas, history of religion among others, that due to its heavy reliance on text as a source of information can be empowered by technologies for macro-scale analysis of texts.

Macro-scale analysis of texts refers to a range of techniques that targets large collections of data in natural language (Jockers 2013; Moretti 2013). The majority of these techniques are often referred to collectively as text mining or text analytics and are developed by computer science (specifically information retrieval, machine learning, natural language processing) and linguistics with the purpose of modeling language and extracting patterns from text-heavy data. Some of the large end-user communities of these techniques are business-related ('business analytics') and biomedical research ('bio-NLP'), but small enclaves of SSH (digital humanities and computational sociology) have also

1 Data that lack a machine readable model, such as natural language, and therefore demands preprocessing and for automated analysis. Unstructured data represent the vast majority of humanities data.

 | DOI:10.1163/9789004385375_016

appropriated these techniques. Specifically techniques for grouping and classifying texts based on their lexical and semantic features have become popular in SSH (Murdock et al. 2015; Roberts et al. 2014; Tangherlini and Leonard 2013; Underwood 2016).

While historical research in general can benefit from macro-scale analysis, approaches to history that are accustomed to formal description and quantification of their objects are particularly well-suited. This is not because their research objects are fundamentally different, but simply due to the way practices of operationalization and modeling have impacted theoretical constructs. Approaches that study psychological and cognitive constructs in historical data are an example of this. These approaches can utilize techniques for macro-scale analysis that have affinities with or even originate in psychological and cognitive theories. In order to illustrate this, the present study combines bayesian modeling, which has gained considerable popularity in the cognitive sciences (Griffiths et al. 2010), with dictionary-based sentiment analysis that is derived from psychological methods (Tausczik and Pennebaker 2010). This combination of techniques were selected in order to study cognitive and affective dynamics underlying cultural semantics as they are expressed in the texts by specific writers that share an interest in the similar cultural phenomena, namely mysticism and mental states, albeit from very different perspectives, one with the insiders' *emic* immersion and the other with the outsiders' *etic* characterization (Headland et al. 1990).

The target data set consisted of sample writings of Spanish mystic and Carmelite nun Teresa Ávila (1515–1582) and American-Danish anthropologist and cognitive scientist Armin W. Geertz (1948–). Both cases can be characterized as highly productive writers who have worked on topics related to mysticism and mental states while subjected to varying levels of cognitive-affective stress. Teresa Ávila (TÁ) represents *emic* semantics with her confessional approach to the mystical experience, while Armin W. Geertz (AWG) represents the *etic* semantics in his scientific approach to mysticism and religious practice. In what follows, we present two studies that compare semantics across the two domains (study 1) and subsequently model the cognitive-affective development of the careers of the two cases (study 2).

Study 1: *Emic* vs. *Etic* Semantics

In SSH, the distinction between *emic* and *etic* research is typically used to differentiate between research based on either an insiders' or an outsiders' conceptual system (Headland et al. 1990). A ritual practice, such as prayer, can be described in terms of the participants' beliefs and values in order to understand its culture-specific features (Headly and Parkin 2000) or it can be described in terms of scientific categories, such as physiology and psychology, in order to identify more general features (Schjødt et al. 2009). Although TÁ's writings are not research in the strict sense, they do describe TÁ's mystical experiences and mental development in great detail. AWG, on have other hand, employs scientific categories in his application of both ethnohermeneutics and cognitive science to religious behavior. In study 1 we compare *emic* vs. *etic* semantics by contrasting writings of TÁ and AWG in terms of their lexical density and semantics. Initially, one might expect *emic* semantics to be more self-contained relying only on its own conceptual system, while *etic* semantics combine technical concepts with references to native concepts. If this is the case, we can expect that *etic* semantics is more lexically dense, but less coherent in terms of lexical meaning because of the interspersion of concepts from both domains.

Methods

Data

The target data set consisted to two samples from TÁ and AWG's collected writings. At the outset, the TÁ sample consisted of 690 documents published from 1546–1594 (one posthumously) translated to English. Availability, OCR quality and language constraints (English only) limited AWG's writings to a sample size of 44 documents published from 1982–2014. For lexical density comparison, all documents were used (N = 734), but each document was length normalized into slices of 250 words (median document length was approximately 500 words) and lexical density was taken as the average of a document's slices. For lexical semantics comparison, the data set was balanced with 44 documents from each writer (N = 88) that were randomly selected in TÁ's case. Before model training and evaluation, the data set was subjected to preprocessing, specifically, case folding, removal of punctuation and numerals, stemming, and pruning for removal of frequent/infrequent words.[2]

Lexical Density

Word-level Shannon Entropy (*H*) was used to estimate lexical density. *H* is a central measure in information theory that models information contained in a document in bits, that is, as the average number of binary digits (0/1) needed to encode each word (Shannon 1948). As a measure of information, *H*

2 All code used in the article is available at: https://github.com/digitaltxtlab/mindthetext.

can be used to model the lexical density of a text (Thoiron 1986; Zhang 2016). Although lexical density is closely related to how predictable a given word is in a document, it does not sufficiently capture semantics.

Lexical Semantics

In order to model and compare lexical semantics, several machine leaning techniques were applied to the data. First, a non-parametric Bayesian model, specifically a latent Dirichlet allocation model, was trained on the data to represent each document as a distribution over ten coherent lexical topics. Second, each topic was categorized as either being either TÁ or AWG dominant by mapping the topic distribution for each document onto the writer information. Third, a series of binary naive Bayes classifiers (one for each topic) was trained and the obtained error rate used as an inverse measure of relative importance. Finally, TÁ/AWG dominant topics were compared according to their importance in classification.

Results

Estimation of H showed that AWG encoded more information in his documents ($M = 6.66$ *bits*, $SD = 0.17$) than TÁ ($M = 6.36$ *bits*, $SD = 0.04$), and this mean difference was statistically significant: $t(732) = 3.19$, $p = 0.002$, 95% CI [0.17, 0.49], $d = 2.47$.[3] Differences in H therefore indicate that *etic* semantics is indeed more lexically dense than *emic* semantics.

The ten categorized (Class) topics from the LDA model are shown in Table 14.1. The TÁ class topics reflects a personal relationship to Christian entities such as God and Jesus (authority, affection), while the AWG class are dominated by scientific approaches to religion (neuroscience, cognition). Topics 3 and 9 are qualitatively different. Topic 3 reflects AWG's continued fieldwork among Hopi Indians, which exhibits more of an *emic* flavor. For TÁ on the other hand, Topic 9 relates to the formal and stylistic aspects of writing letters (in a religious context) and therefore has some aspects that are less dependent on *emic* content. Finally, the binary classifier's obtained error rate indicated that TÁ class topics ($M_{err} = 0.05$, $SD_{err} = 0.02$) were better at discriminating between TÁ and AWG than the AWG class' topics ($M_{err} = 0.14$, $SD_{err} = 0.07$). Topics related to TÁ are, in other words, more distinct ('on' for TÁ and 'off' for AWG documents) and hence important for classification. Figure 14.1 show that in the AWG class it is topic 3 (Hopi indians) with most *emic* content that performs best, while for TÁ it is topic 9 with least *emic* content that performs worst.

3 The result was reproduced for word slices with length 100–1,000 word.

TABLE 14.1 Labeled (second column) and categorized (third column) topics from the LDA model trained on TÁ and AWG samples. Labels were assigned qualitatively based on the most frequent and distinctive words in each topic and categorization based on topic distribution for each document.

Topic ID	Label	Class
1.	Authority relations	TÁ
2.	Affective relations	TÁ
3.	Hopi indians	AWG
4.	Sibling relations	TÁ
5.	Religion and neuroscience	AWG
6.	God and soul	TÁ
7.	Ritual and cognition	AWG
8.	Monastics	TÁ
9.	Religious writings	TÁ
10.	Scientific study of religion	AWG

Discussion

In the comparison between *emic* and *etic* semantics, study 1 confirmed that documents dominated by *etic* semantics are indeed more lexically dense than documents based on *emic* semantics. Although generalization of this finding is limited to the two writers, it does support the claim that *emic* semantics relies on a limited conceptual system that, in contrast to *etic* semantics, limits combinations of domains and is more self-contained. This claim is corroborated by the classifier results, which indicated that *emic* topics had more discriminatory power than *etic* topics. The 'borderline' topics 3 and 9, which were the most and the least *emic* in opposite classes, further supported this interpretation because they were the best and worst, respectively, at discriminating between TÁ and AWG. Although study 1 shows semantic differences between insider/outsider perspectives on mysticism and mental content, it neglects the temporal dimension and treats each writer as a coherent whole. Writers however follow a developmental trajectories where previous writings, mental states and external events have differential effects.

Study 2: Cognitive-affective Trajectories

With macro-scale analysis it becomes possible to summarize numerous documents in terms of a few abstract properties. In study 1 each documents was

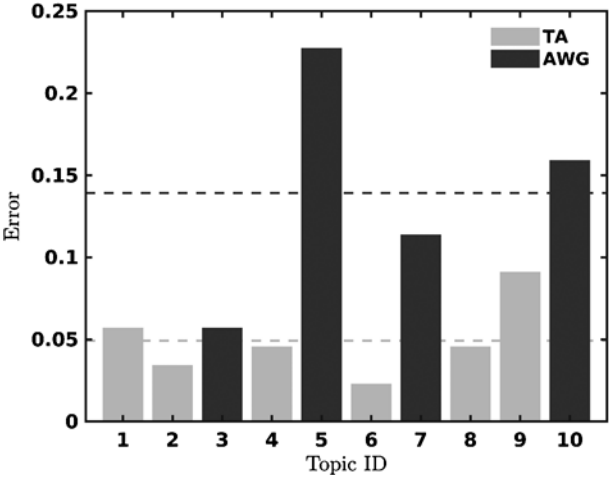

FIGURE 14.1 Obtained error rates for each topic used as an independent feature for classifying documents as TÁ or AWG. Topics categorized as TÁ dominant are shown in light gray and AWG dominant in dark gray. Vertical dotted lines represent the mean error for each category.

summarized in terms of its entropy and the writers' average lexical density were estimated and compared. Whenever a data set contains temporal coordinates (e.g., metadata about its objects' time of origin) macro-scale analysis renders analysis of the dynamics between documents possible and, by extension, the time dependent generative process underlying the data set. In study 2 we will model two such processes and compare their dependencies within each writer. The first generative process is creative state which by cognitive scientists has been describes as 'variation on a theme' (Hofstadter 2008). The second process is affective tone, which has a long tradition in social and personality psychology. Together these two processes can draw the contours of TÁ and AWG cognitive-affective trajectories. Compared to the previous study, study 2 is exploratory and the goal, therefore, is to summarize and visualize the data set in terms of these trajectories.

Methods

Data

The target data set was identical to the full 734 documents from study 1. Metadata was derived directly from the documents and 96 TÁ documents had to be removed because they lacked dates of origin resulting in a total sample size of 638 documents ($n_{TA} = 594$ and $n_{AWG} = 44$). For affective tone the documents were used in their original form, while for modeling of creative state, the documents were preprocessed as in study 1 part 2.

Creative State

Taking Hofstadter's 'variation on a theme' literal, creative state was modeled as the distance between each document represented as a probability distribution over lexical topics. A two-step process was used to construct a creative state time series for each writer. First, a latent Dirichlet model was trained in order to represent each writer's documents in terms of a set of coherent topics. To allow for more fine-grained representations, each model was trained on 50 topics. Second, the average Kullback-Leibler (KL) divergence was estimated for each document and its temporally prior documents. KL divergence, or relative entropy as it is also called, is widely used in information theory to measure the similarity between two probability distributions (Kullback and Leibler 1951). Creativity is, in this study, equated with novelty and the relative distance (dissimilarity) between two documents' topic distributions treated as the degree of topical novelty. The creativity scores were zero-centered by removing the mean from all values.

Affective Tone

To model the affective tone of each document, the Language Assessment by Mechanical Turk (labMT) sentiment dictionary was used (Dodds et al. 2011). A sentiment dictionary is a list of words and their rated affective valence. The labMT's dictionary performs extremely well when compared to other sentiment dictionaries and is designed to avoid several problems with dictionary-based sentiment analysis (Reagan et al. 2015). The dictionary consists of more than 10,000 frequent English words that are rated for positivity/negativity on a continuous scale. In study 2, the documents for each writer were time-ordered and the labMT dictionary was applied to each document resulting in two sentiment time series. Sentiment scores were also zero-centered by removing the mean from all values.

Results

Figure 14.2 (upper row) shows the annual averaged distance between time-ordered documents in the LDA-model. The temporal signature of creative state is different for the two writers. TÁ shows an early period of creativity with few documents and a relatively high level of novelty that starts to decrease around 1561–1562. This 'early style' signature is then followed by a flatline for the remainder of TÁ's active period of writing. AWG's time series indicates a similar early style followed by a flatline 1993–2000, but then a 'turn' in 2000–

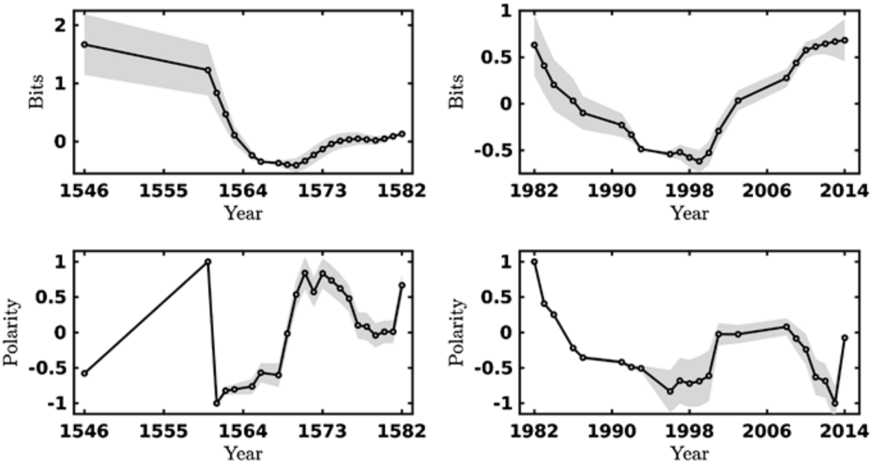

FIGURE 14.2 Upper row shows the cognitive trajectories (i.e., creative state measured as KL-divergence between topic distributions of documents) for TÁ (left) and AWG (right). In same order, the lower row show affective trajectories based on the labMT dictionary (sentiment vectors normalized to $\{-1\ 1\}$).

2001 after which there is an rapid increase in novelty until the present. Finally, comparing the overall average distance between the two writers yields a significantly higher level of novelty for AWG ($M = 12.21$ *bits*, $SD = 1.37$) than for TÁ ($M = 7.97$ *bits*, $SD = 1.35$): $t(635) = 19.91$, $p < 0.0001$, 95% CI [3.82, 4.66], $d = 3.13$.

Time-ordered sentiment vectors show a slightly different pattern. For TÁ figure 14.2 (lower row) shows that her early period was characterized by great variation in few documents. Around 1561–1562 TÁ's documents have a predominantly negative tone, which subsequently show a positivity trend for the remainder of her career. AWG's sentiment vector is more similar to his creative state with an early decrease, a turn in 2000–2001, but then a positivity plateau followed by a decreasing trend. On average TÁ ($M_{labMT} = 0.35$, $SD = 0.22$) show a reliably stronger positivity bias than AWG ($M_{labMT} = 0.25$, $SD = 0.17$): $t(635) = 3.15$, $p < 0.01$, 95% CI [0.04, 0.17], $d = 0.55$.

Comparing creative state with affective tone shows opposite patterns of association for the writers. There is a inverse correlation between TÁ's creative state and affective tone ($r = -0.2$, $p < 0.0001$) indicating that higher levels of creativity tends to be associated with negativity and lower levels with positivity, AWG on the other hand displays a stronger and direct correlation between creative state and affective tone ($r = 0.32$, $p < 0.0001$), such that higher levels creativity is typically associated with higher levels of positivity and lower levels with negativity. Modeling predictive relations between creative state and affective tone using bivariate Granger causality (with a maximum number of lags at 5 documents) show a significant bidirectional relation for TÁ ($F_{cs \to at}(1, 591) = 5.02$, $p = 0.02$ and $F_{at \to cs}(5, 586) = 6.7$, $p < 0.0001$), which indicates an extraneous factor that is driving both variables. It is worth noticing that creative state seems further delayed in relation to affective tone. AWG shows no significant predictive relations ($F_{cs \to at}(1, 40) = 3.31$, $p = 0.08$ and $F_{at \to cs} < 1$) which is affected by the relatively small sample size.

Discussion

By modeling the individual cognitive-affective trajectories, it becomes possible to identify different career paths for the two writers. All things being equal, TÁ displays a typical early career pattern, where the majority of topical novelty is generated in the initial career phase. This creative state trajectory extends the Hardy-Wiener claim that there exists a negative relation between age and creativity, beyond the domain of mathematics.[4] Contrary to this, AWG's trajectory exhibits a turn behavior where an early career pattern is followed by a sudden change in topics and a massive increase in novelty. Looking at the underlying data, this specific turn reflects an orientation toward cognition, evolution and neuroscience in AWG's research. Comparisons between creative states also showed, not surprisingly, that turns will result in a higher degree of overall novelty in an authorship. Comparisons with affective tone gave indications of motivational mechanisms related to career paths. Although the analysis did not support a directional relationship between cognitive and affective variables, it is possible to argue that TÁ is an example of a suffering artist (i.e., a state incongruity writer), because her most creative states were associated with a negative tone. In contrast, AWG is a state congruity writer where creativity and affective tone behave similarly.

General Discussion

Macro-scale analysis offers entirely new ways of researching historical and contemporary textual sources. We can now pose questions to all the available data and answer with the formal rigor and capacity for generalization afford-

4 "I do not know of a single instance of a major mathematical advance initiated by a man past fifty" (Hardy 1967). Norbert Wiener further restricts the creativity cut-off to 30 years (Wiener 1964).

ed by mathematical and statistical modeling. These tools and techniques can empower domain experts to solve empirical problems at a different scale and pose more valid theoretical models. Novel technology is always associated with a risk and historical researchers, as any researchers, has to carefully access the potentials and pitfalls of macro-scale analysis.

The two studies of TÁ and AWG illustrated how historical research can use macro-scale analysis to access mental content and development at the level of semantics for individual writers. In contrast to micro-analysis, the studies were carried out on the collected works of TÁ and a substantial sample of AWG's writings in English. While it would have taken substantial human-based information processing to read and analyze more than 700 documents, it is only a matter of minutes for silicon-based processing. The scale of macro-analysis allows historical research to go beyond qualitative assessment of few sources and transcend the cognitive limitations of the individual researchers. It further enables the use of data mining and machine learning techniques, because they often depend critically on large amounts of data. Although 700 documents comprise a small data set in machine learning terms, the studies illustrated how both unsupervised (i.e., the LDA model) and supervised (i.e., the naive Bayes classifier) machine learning, combined with information theory, can enrich our understanding of psychological systems (i.e., writers) with a history codified in text. Methods for comparison within and between historical systems, whether they constitute persons, groups or traditions, change with the availability of similarity measures and alignment techniques of (time-dependent) sequences. As shown in study 2, it becomes possible to identify couplings between relevant system properties (e.g., cognitive state and affective tone) and identify system-level differences (e.g., state congruity vs. state incongruity writers). These techniques are beginning to create models that can not only describe a system's behavior, as shown in study 1 and 2, but also generate it (e.g., Lipton et al. 2016).

Macro-scale analysis has many limitations, erroneous use of macro-scale analysis can have both negative scientific and ethical consequences, and it will not replace genuine domain expertise. Data availability is one such limitation that was exemplified by the size of AWG's sample. Although AWG has published more than 300 documents, only a limited number was freely available. It is an important point that the contemporary writer's data were hardest to gain access to. While historical data often needs digitization, availability of more contemporary data is often blocked by copyright and commercial interests. A limitation that is more pronounced for historical data is their quality. In study 2 it was necessary to remove almost 100 TÁ documents because they lacked proper dating. Optical character recognition, which is the process of converting images of text to machine encoded text, is often very error prone, especially for blackletter fonts.

Beyond availability and quality of data, there are several pertinent concerns related to biases and generalization. Even if one accepts the convenience sample of AWG, the study is still in many aspects limited by the target data set. While it is possible to say something about the individual writers, the study only have one instance (measured over a long time period) of *emic* or *etic* semantics respectively. While many documents give the illusion of a large sample, the actual size depends critically on the research question. Imagine that the target data set was used to argue that women exhibit lower lexical density than men using TÁ and AWG as class representatives. While the concrete analysis might allow for this interpretation, the samples do not. We can therefore only conclude that TÁ's writings are less lexically dense than AWG's. The potential sampling bias in macro-scale text analysis of historical systems multiply when one starts to consider the demographic biases in written, published and preserved textual sources. In comparison to randomized controlled trials, which is often considered the gold standard for psychological and cognitive research, it is apparent that variable control is severely limited in macro-scale text analysis and that *more data* do not always solve the problem (Lazer et al. 2014). While experimental research can control the temporal order of variables, macro-scale analysis must use the organization of study-external events and, at least to some extent, rely on *post hoc ergo propter hoc*. Furthermore, the textual effect of multiple contextual events will often have to be simplified considerably and can often be hard to estimate. The novelty signal of TÁ, for instance, might have been stronger than that of AWG if both were compared to a baseline of their respective historical contexts.

Independently of the methods and techniques applied to data, new knowledge in any domain ultimately rests on human interpretations and the value of these interpretations lie in their capacity to make others, as Andreas Buja have said, 'fruitfully think about an idea' (Hastie et al. 2009). As exemplified by AWG's developmental trajectory, it is necessary to be open to and combine multiple techniques and perspectives in order to make valuable interpretations of human behavior.

Bibliography

Dodds, Peter S., Harris, Kameron D., Kloumann, Isabell M., Bliss, Catherine A., and Danforth, Christopher M. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE* 6(12): e26752.

Griffiths, Thomas L., Chater, Nick, Kemp, Charles, Perfors, Amy, and Tenenbaum, Joshua B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8): 357–364.

Hardy, Godfrey H. (1967). *A Mathematician's Apology*. Cambridge: Cambridge University Press.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed. 2009; corr. 7th printing 2013). New York: Springer.

Headland, Thomas N., Pike, Kenneth, and Harris, Marvin (eds.) (1990). *Emics and Etics: The Insider/Outsider Debate*. Newbury Park: SAGE Publications, Inc.

Headly, Steven C., and Parkin, David J. (2000). *Islamic Prayer Across the Indian Ocean: Inside and Outside the Mosque*. Richmond, Surrey: Routledge.

Hofstadter, Douglas (2008). *Metamagical Themas: Questing for the Essence of Mind And Pattern*. New York: Basic Books.

Jockers, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Kullback, Solomon, and Leibler, Richard A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1): 79–86.

Lazer, David, Kennedy, Ryan, King, Gary, and Vespignani, Alessandro (2014). The parable of Google flu: traps in big data analysis. *Science* 343(6176), 1203–1205.

Lipton, Zachary C., Vikram, Sharad, and McAuley, Julian (2016). Generative Concatenative Nets Jointly Learn to Write and Classify Reviews. https://zdoc.site/queue/generative-concatenative-nets-jointly-learn-to-write-and-cla.html.

Moretti, Franco (2013). *Distant Reading*. London and New York: Verso.

Murdock, Jamie, Allen, Colin, and DeDeo, Simon (2015). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. Preprint. arXiv:1509.07175.

Reagan, Andrew J., Tivnan, Brian F., Williams, Jake Ryland, Danforth, Christopher M., and Dodds, Peter S. (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. Preprint. arXiv:1512.00531.

Rehm, Sven-Volker, Reschenhofer, Thomas, and Shumaiev, Klym (2014). IS Design Principles for Empowering Domain Experts in Innovation: Findings From Three Case Studies. Retrieved from http://aisel.aisnet.org/icis2014/proceedings/ProjectManagement/9/.

Roberts, Margaret E., Stewart, Brandon M., Tingley, Dustin, Lucas, Christopher, Leder-Luis, Jetson, Gadarian, Shana Kushner, and Rand, David G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58(4): 1064–1082.

Roure, David De, and Goble, Carole (2009). Software design for empowering scientists. *IEEE Software* 26(1): 88–95.

Schobel, Johannes, Pryss, Rüdiger, Schickler, Marc, Ruf-Leuschner, Manfred, Elbert, Thomas, and Reichert, M. (2016). End-user programming of mobile services: empowering domain experts to implement mobile data collection applications. In *2016 IEEE International Conference on Mobile Services (MS)*, 1–8. IEEE.

Schjødt, Uffe, Stødkilde-Jorgensen, Hans, Geertz, Armin W., and Roepstorff, Andreas (2009). Highly religious participants recruit areas of social cognition in personal prayer. *Social Cognitive and Affective Neuroscience* 4(2): 199–207.

Shannon, Claude E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1): 3–55.

Tangherlini, Timothy R., and Leonard, Peter (2013). Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics* 41(6): 725–749.

Tausczik, Yla R., and Pennebaker, James W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1): 24–54.

Thoiron, Philippe (1986). Diversity Index and Entropy as Measures of Lexical Richness. *Computers and the Humanities* 20(3): 197–202.

Underwood, Ted (2016). The Life Cycles of Genres. Journal of Culture Analytics. Retrieved from http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

Wiener, Norbert (1964). *I Am Mathematician*. Cambridge: MIT Press.

Zeleny, Milan (2012). High technology and barriers to innovation: from globalization to relocalization. *International Journal of Information Technology and Decision Making* 11(2): 441–456.

Zhang, Yanhui (2016). Entropic evolution of lexical richness of homogeneous texts over time: A dynamic complexity perspective. *Journal of Language Modelling* 3(2): 569–599.