Analysis of Cognitive Surprise in Scientific Writing by Discipline

Benjamin Goldstein

**ABSTRACT**

The history of science represents a dominant force in the evolution of human culture. Many of the transitions between epochs that historians use to catalog modern human history, i.e. the first and second industrial revolutions and enlightenment, have been driven by innovations in scientific methodology and practices. Understanding the evolutionary pathway of scientific writing can serve as a useful map back to the trajectory of scientific culture, and greater culture more generally. Digital archives of scientific texts combined with developments in methods of contemporary computational linguistics provide an accessible means of analyzing cultural evolution through the lens of quantitative diachronic change. The digitization of these resources also facilitates the inclusion of relevant metrics and metadata in analysis. In this study, the evolutionary pathway of scientific writing is studied through the lens of cognitive surprise, in the Bayesian sense. This method is applied to the investigation of the *Royal Society Corpus* (RSC), a digital archive of scientific texts from 1665-1920. The texts are divided into groups by topics (generated by reliable clustering models designed and run by the corpus’ management team), and this data is then compared against historical trends in the field. The results show a degree of support for decreased surprise following the generation of a newly coherent theory that unites a field, and increased surprise following discovery that broadens the field as well as a boom resulting in elevated linguistic diversity. The results also show some support for the overall trend that fields characterized by theoretical consolidation tend to decrease in entropy over the full period, while fields characterized by repeated broadening via new discovery display the opposite long-term behavior. This study generates quantitative support for the correlation between the trajectory of a field’s linguistics and that of the field’s history.

**INTRODUCTION**

Patterns in cultural evolution are crucial to understanding the history and nature of mankind. Modern computational methods and digital resources allow for the quantitative investigation of specific areas of cultural evolution, and comprise the foundation for a budding analytical science of human cultural change. One such subsection of cultural evolution that lends itself to such analysis is science. Science is clearly one of the chief drivers of culture: think of the influence of the combustion engine on ease of movement, the steam engine on industrialization and urbanization, and communication technology on human interaction. The trajectory of scientific innovation is thus of utmost importance to the study of the arc of human culture. Scientific journal articles represent an extremely valuable record developments in science, and can be treated as a manifestation of the evolution of science itself. Scientific writings, therefore, represent an enormously useful tool for the understanding of scientific progression.

Quantitative linguistic data represents an effective window into human culture. It is crucial to combine effective computational methods with ample historical data as well as substantive context. Many early studies in the development of this area analyze massive, heterogeneous selections of texts (largely based on arbitrary traits such as availability on google books) and apply a piecemeal system of contextual and linguistic findings to extract cultural information (Michel et al. 2011). Later studies establish the importance of using a coherently collected corpus of texts, and avoid picking and choosing words to analyze when attempting to gather big-picture cultural information (Sun et al. 2020).

The RSC is a selection of texts from the *Philosophical Transactions of Royal Society* (PTRS), the world’s oldest scientific journal dating back to 1665. The corpus ranges from 1665-1920, and represents a steady flow of scientific publications of various topics, making it a fit candidate for computational linguistic analysis. The period covered by the corpus spans the scientific revolution, first and second industrial revolutions, and the formation of modern scientific and mathematical methodology. The European Language Resources Association (ELRA) manages the corpus (Kermes et al. 2016), and in the latest version includes topic model (Blei et al. 2003) data using MALLET (McCallum et al. 2002) in order to categorize the texts by scientific discipline.

Scientific writings are widely accepted as exhibiting increased professionalization and specialization over the course of the period covered by the corpus. Such observations stem from qualitative observations and/or crude quantitative analysis, such as simple increases or decreases in the frequency of a given lexical phenomenon. More advanced statistical and computational methods allow for the generation of more complex metrics, which better reflect semantic meaning. Another important byproduct of increased usage of digital linguistic methods is higher quality metadata, which allows for better handling of text type diversity. The corpus provides a consistent stream of texts spanning over 250 years with high quality metadata all linked to PTRS, which would be difficult to accomplish from a general source like Google Books as opposed to the journal itself.

There is precedent for topic-wise analysis on scientific writings. Some studies use divergence metrics to investigate the similarities and dissimilarities from pre-categorized corpora (Dias et. al 2018). Topic modeling methods such as Latent Dirichlet Allocation (LDA) are well founded in information science, and are used by the ELRA to categorize the corpus (Kermes et al. 2016). Comparative analysis of methods in topic modeling are outside the scope of this study, but the results of

In addition to compiling thousands of texts from the PTRS with high quality metadata, the RSC also presents those texts in a machine readable format. Texts are tokenized (split into linguistic tokens), and then each token is tagged with its lemma and part of speech. The lemma of a word is the form under which it is registered in a dictionary (i.e. the lemma of most English verbs is the infinitive form). The RSC stores this data in various formats, but for this study the XML (extensible markup language) form of the texts is used. XML is readily parsed by industry standard parsers– XML is extremely commonly used to encode data in web based systems, so ample resources are available to manage it.

This study uses established methods in computational linguistics as well as the digital resources described above to address these issues and evaluate the degree of quantitative support for bigger picture trends. The principal method used in this study is the Kullback-Liebler divergence (KLD), a statistical measure derived from information theory that is widely accepted as a means of measuring cognitive surprise between linguistic corpora. Tracking this metric can provide insight as to the long term trend of cognitive surprise, as well as how a field’s short-term behaviors map to historical events.

**BACKGROUND**

The overall trend of scientific writing is toward increased specialization and professionalization (Sun et al. 2020). Specialization is operationally defined as an increased need for genre divergence, including the introduction of new disciplines. Similarly, professionalization is defined as an increased efficiency of communication within scientific writings. Notably, while professionalization can be easily tracked within a topic, specialization seems to mingle with the boundary of the topic itself. For the purposes of this study, periods of increased surprise are classified as periods of divergence, while those of decreased surprise are classified as periods of convergence. This distinction is once again supported by Sun et al. (2020) as well as in the essence of KLD as a Bayesian surprise metric.

Previous studies analyzing linguistic change have tracked the frequency of very particular linguistic phenomena over time (Biber and Gray 2016). Word frequency alone is context independent, and cannot directly provide reliable insight. In this study, the words that drive linguistic change are identified using frequency changes, but that constitutes only the preprocessing phase. The relative entropy is then calculated between all decade pairs with sufficient data, providing a larger scale look at the linguistic landscape than that which can be offered by simple lemma-wise frequency change analysis.

The metric of relative entropy is used to track cognitive surprise between decades within each topic of text contained in the RSC. The KLD (Kullback and Liebler 1951) measures the amount of additional information needed to represent one probability distribution over another. In the linguistic context, this justifies why the metric can successfully quantify the amount of cognitive surprise between decades, as well as the previously established divergence/convergence distinction. The KLD between two discrete probability distributions p and q is calculated using the following equation:

Here, x is a discrete variable– continuous distributions require a formula that includes integration rather than summation. When p is a prior distribution and q is a posterior distribution, the KLD is becomes a metric of the gain or loss of information between the two periods. In the linguistic context, this represents the cognitive surprise encountered by a reader of texts from the post period having become accustomed to texts from the pre period.

Several recent studies exist that use KLD as a metric for linguistic change. KLD appears in Sun et. al (2020) in conjunction with word concreteness and imageability to create a wider approach for linguistic analysis in a case study on the RSC. An earlier study (Degaetano-Ortlieb and Teich 2018) uses KLD to generally measure the trajectory of the RSC, establishing the statistical methods used in Sun et. al as well as this study. KLD has also been applied to datasets other than the RSC, i.e. measuring rhetorical changes in revolutionary speeches and analyzing the trajectory of Darwin’s career through his writings.

**MATERIALS & METHODS**

The PTRS is one of the oldest scientific journals available, with available texts ranging from 1665-1920. The steady stream of publications over this period represents a unique resource for linguistic analysis. Scientific journals are the best kept records of scientific development, and are thus worthy of investigation.

The RSC tabulates all of the published texts in the PTRS in a variety of digital formats. Private versions contain additional texts over a slightly wider timeframe, but the publicly available version still boasts over 17,000 texts with over 75,000,000 total tokens. The RSC represents these texts in XML format, facilitating the storage of text-level and token-level metadata. Tokens are annotated by their part of speech and lemma. As previously mentioned, the lemma of a word is the standard form, i.e. “surround”, “surrounding”, and “surrounded”, all share the lemma “surround” even though they are different tokens. The dataset is available in several other formats as well, including TEI, TCF, VRT, and plaintext. Notably, the XML dataset is available in a compressed archive such that each text receives its own file, massively easing the memory demand of tree-like parsing methods.

As mentioned earlier, the RSC also provides high quality metadata for all of its texts. This data is available in tsv (tab separated value) format. This format yields a spreadsheet-like structure, facilitating lookups of other attributes by text id (which is available in the XML). The metadata used for this study includes text id, publication decade, primary topic, and text type. The text id is used for lookup purposes, the publication decade is used to group articles by time period, and the primary topic is used to group texts by topic. The journal includes writings that are not full research papers that may not provide as robust results to this study i.e. obituaries, abstracts, and book reviews. These writings are excluded using the text type column of the metadata file..

Notably, the topic metadata is different from other columns. It is not simply a description of some fact about the text, like author, publication date, or even text type, which can be discerned fairly easily and concretely. The topic of a text is generated using a clustering-based topic model, which uses machine learning to analyze how the texts logically coalesce into 30 different disciplines. The integrity of these topic models is outside the scope of this study, but it has been mentioned that LDA is a widely accepted means of topic clustering in the field of information science. Using the topic models as a means of grouping texts also restricts analysis of diversification of the field of science as a whole, as this would likely occur across disciplines rather than within a particular one or few.

Not all of the results of the topic model lend themselves to the goal of this study: mapping the KLD curves back to the history of the field. For instance, certain topics generated by the model are too vague to investigate any historical context, i.e. “formulae”, “tables”, and “measurement”. Another feature of the topic model that inhibits this sort of analysis is multiple clusters within a topic. For instance, biology is separated into three separate topics: Biology 1, Biology 2, and Biology 3. The limited available documentation shows a different set of most common tokens (only the five most common tokens for each topic were documented) for each of these subclusters. A basic eye test can determine, for instance, that Biology 3 is more geared toward microbiology while Biology 1 is closer to a bigger picture evolutionary field, but this is insufficient for rigorous quantitative analysis. These topics, therefore, are only analyzed for big picture trends and not included in the portion of the study that attempts to map the history of the field onto the entropy curve. Other topics do not have enough valid data once processed through the methods described in the subsequent sections, so those are excluded as well. Further analysis could study alternative methods of topic modeling and/or more carefully discern which of the numbered topics map cleanly to subfields, though this documentation did not seem to be publicly available in my investigation. Subsequent inquiries could also group the topics at higher levels in the tree, i.e. life sciences, observation/reporting, chemistry, and physics. This seemed to be overly broad for this study, but it is worth noting that the tree-like structure of the topic model’s result allows for this option.

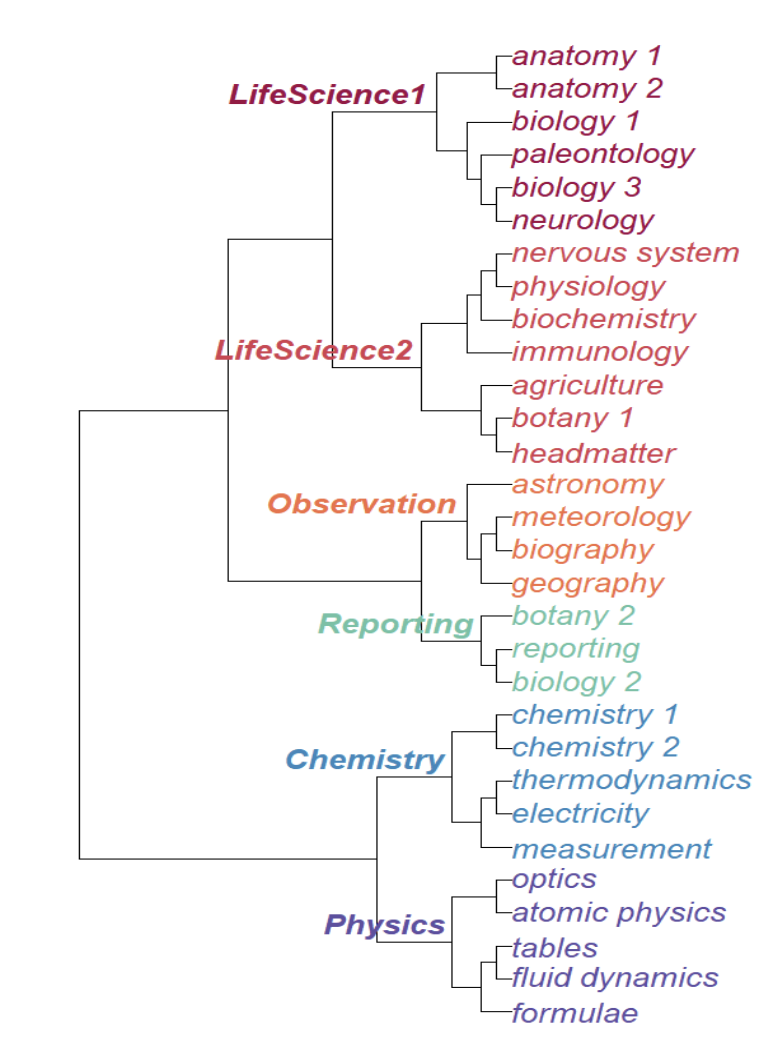


Fig. 1: Tree of results from the LDA Topic Model on RSC

The primary method used in this study is KLD, or relative entropy to measure the cognitive surprise a reader would encounter in a given decade compared to the prior decade when reading texts about a given topic. First, the metadata is loaded into a more convenient data structure facilitating lookups without needing to computationally scroll through thousands of rows of a spreadsheet. This data structure partitions texts by topic and decade, so lists of text ids for such groups are readily available. Next, BeautifulSoup XML/HTML parsing library is used to count the lemma frequency of all lemmas (case insensitively) of texts for each decade (pre period) and its subsequent decade (post period). A list of common words known to cause more noise than insight in linguistic analysis is excluded.[[1]](#footnote-1) Then, the words are filtered again looking for lemmas with a frequency increase of more than 100 appearances per million from the pre to the post period. This step is designed to identify the lemmas that most contribute to changes in the scientific landscape across that period. Finally, Welch’s t-test is run on the two frequency vectors and if a p-value smaller than 0.01 is achieved the data is deemed valid. If not, the increase threshold is decreased incrementally until statistically acceptable results are achieved. If no threshold achieves a p-value less than 0.05, the data is deemed invalid and no datapoint is said to exist for that topic in that decade. Finally, the frequency vectors are normalized and the relative entropy between them is calculated providing a set of data points relating decade to relative entropy, or cognitive surprise. Basic linear regression is then applied to these plots to detect overall trends, and qualitative historical analysis is applied to evaluate the correlation between this statistical metric and the real-life scientific landscape.

**RESULTS & ANALYSIS**

Preliminary analysis on the metadata of the corpus yields basic information about the prevalences of various topics in the corpus over time.

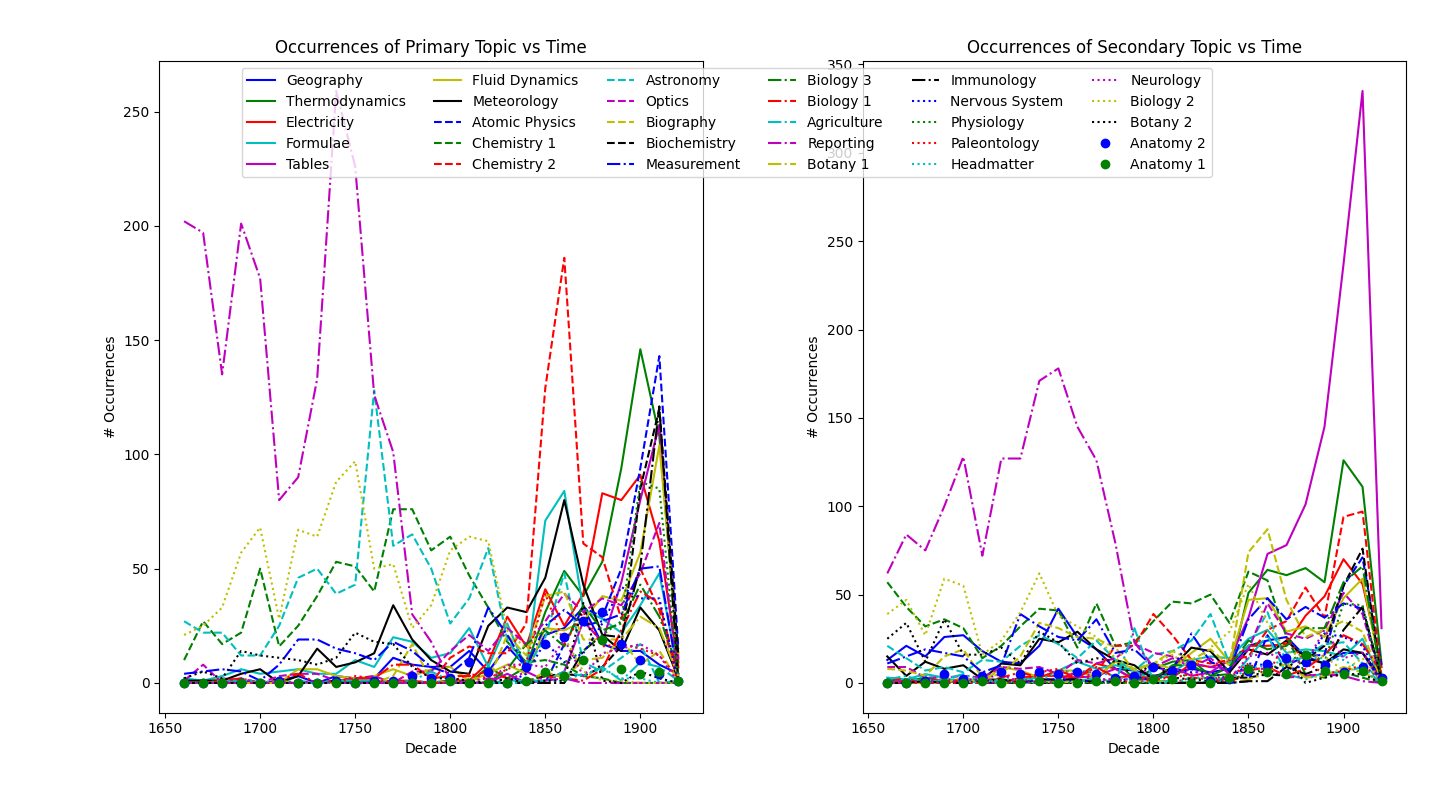


Fig. 2: Number of texts by given primary and secondary topic vs decade

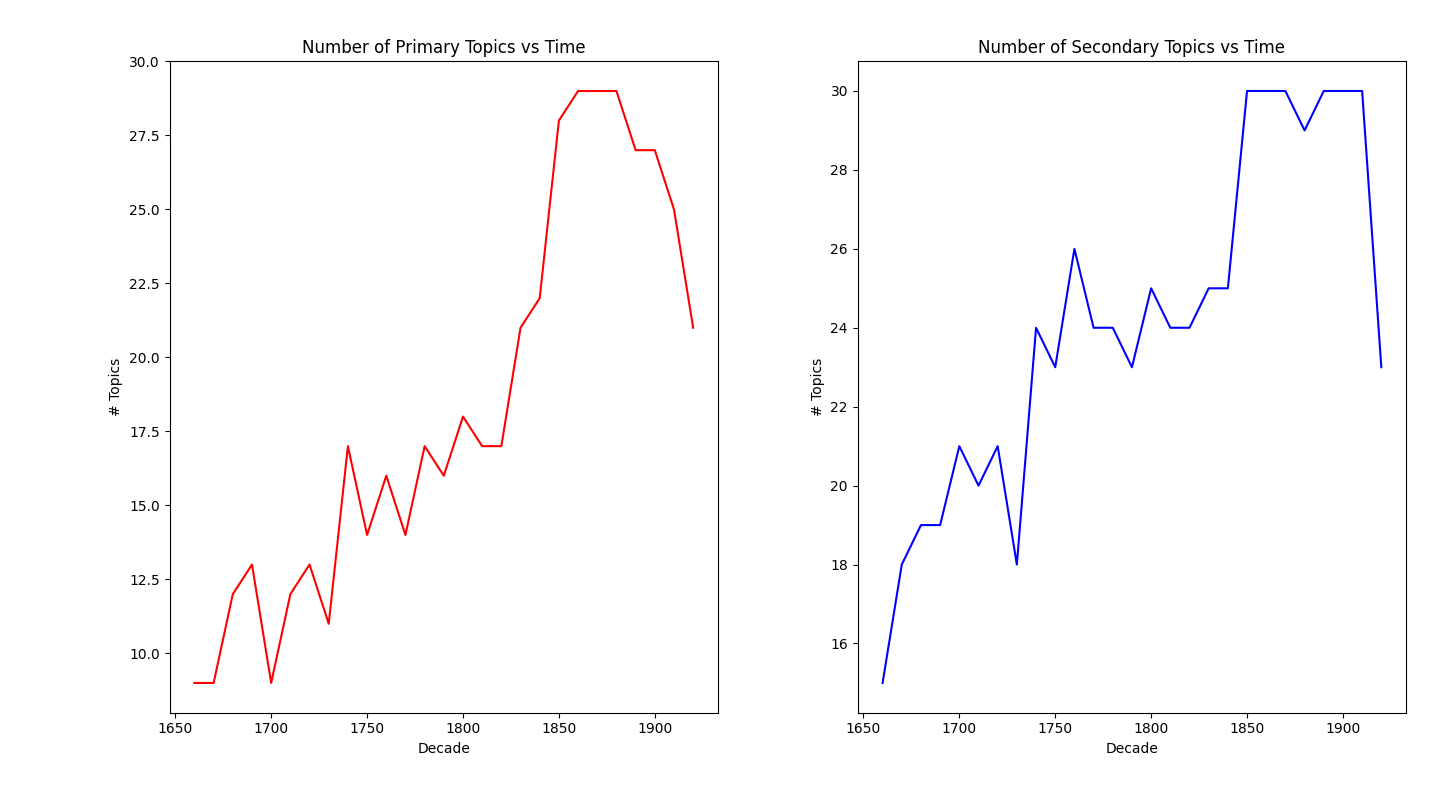


Fig. 3: Number of primary and secondary topics with at least one text by decade

These graphs corroborate the results of prior studies that implies increased specialization over time, as the topic landscape changes from a few dominant areas to a diverse landscape. Note that the decrease in the last datapoint can be ignored– the RSC stops at the year 1920 so only one year of texts are available for the final decade, hence the downturn. Also note a rapid decline in the number of texts with “reporting” as the primary topic rapidly during the Enlightenment period (mid 18th century), supporting an increased precision and usage of the scientific method. Conversely, there is a spike in the number of publications that have “tables” as a secondary topic, which supports increased use of quantitative data in publications.

The principal KLD analysis yielded varied results across topics. The topics divided into subclusters exhibited wild fluctuations and little to know analyzable characteristics in their curves. Chemistry 1 and Chemistry 2 both exhibit slightly negative slopes with extremely weak, statistically insignificant correlations (r-value -0.13 and -0.04 respectively). Biology 3 exhibits a moderately strong upward correlation (r-value -0.33), Biology 2 exhibits wild fluctuation with no statistically significant overall correlation, and Biology 1 exhibits a strong upward correlation (r-value = 0.66).

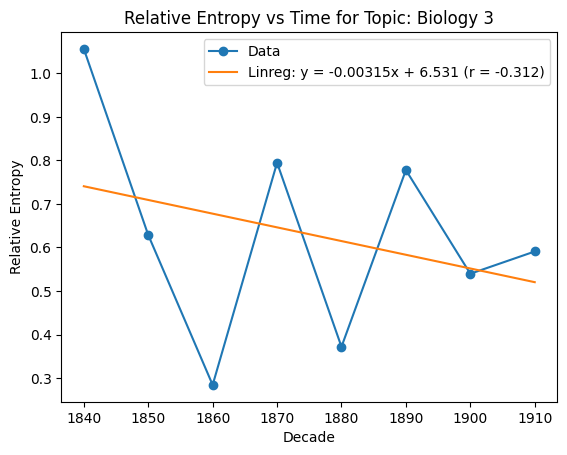
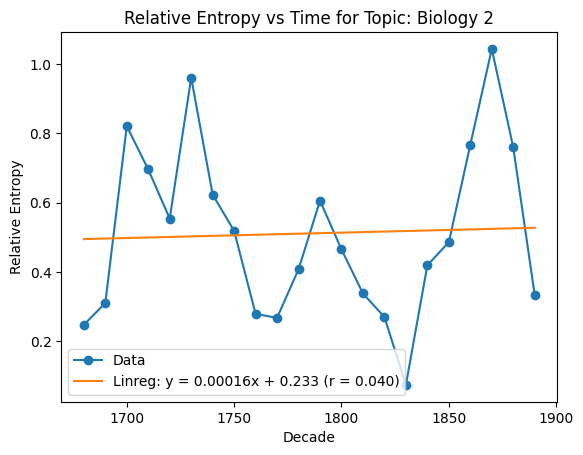
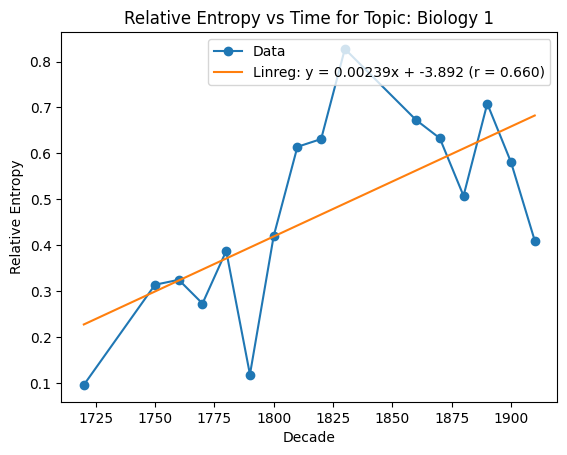


Fig. 4: KLD vs time plots for the 3 subdivisions of biology

This is consistent with the eye-test mentioned earlier (microbiology should converge into coherent theory while evolutionary biology should expand its vocabulary with a steady stream of discovery), but that is not stable grounds for any hard support of the divergent/convergent hypothesis. Botany and anatomy (the other two topics with numbered subdivisions) did not provide any statistical insight, as any overall trends in one group were counteracted by the other and no qualitative subdivisions were found when examining the most common lemmas.

All subsequent analysis will concern sharply defined topics– subdivided or vague topics are excluded. Strong support for convergence was shown in physics, where the period contained in the RSC represents one of robust theoretical development and novel agreement on a unified vocabulary.

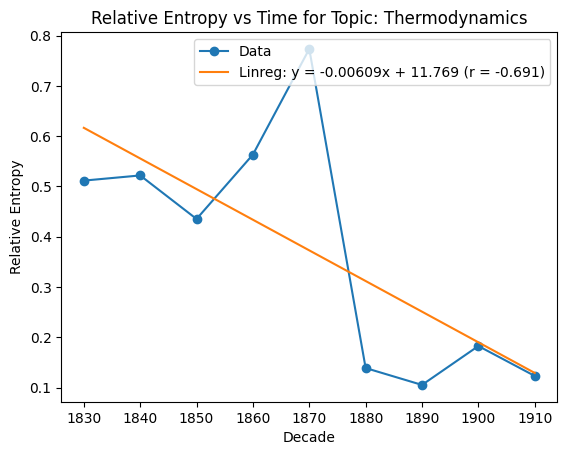
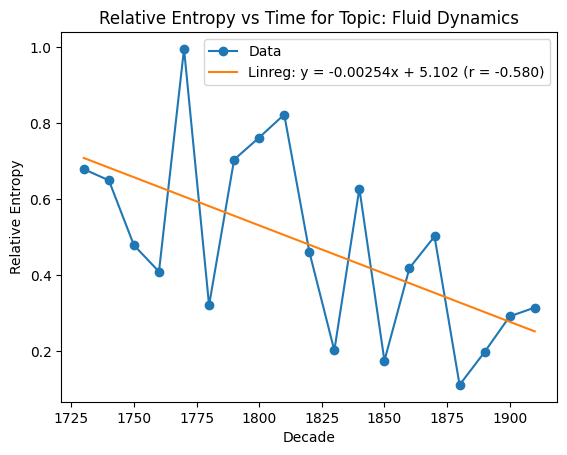
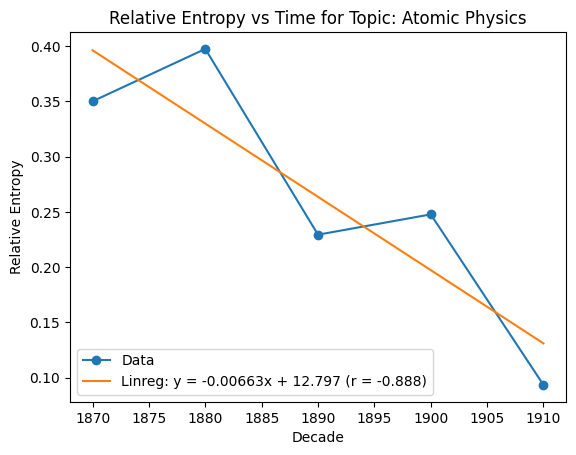


Fig. 5: KLD vs time plots for sub-disciplines of physics

The period between 1665 and 1920 represents the creation of theories of classical mechanics, thermodynamics, and electrodynamics, and even the birth of modern physics. These plots show strong support for a linguistic convergence from scattered opinions and vocabulary onto concisely expressed theories and common language. Notably, the field of optics did not exhibit the same degree of support, overall declining but with a weaker correlation of r-value -0.227.

Medical science showed a similar trend to that of physical science: strong convergence over the period analyzed.

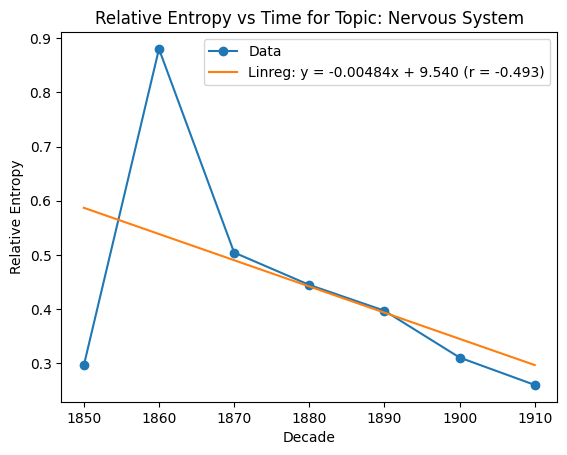
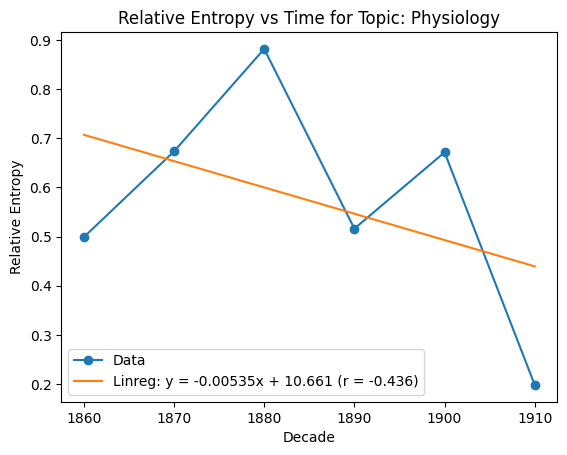
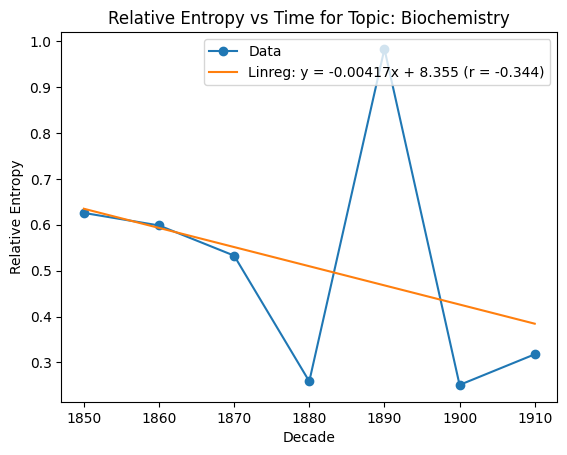


Fig. 6: KLD vs time plots for sub-disciplines of medical science

Though with slightly weaker correlation, these plots still support the same overall phenomenon of linguistic convergence as that previously shown for physics. At the start of the analysis period, the first discoveries of cell-level phenomena were occurring: Leeuwenhoek discovered blood cells in 1670 and first observed bacteria in 1683. By the end of the period, modern technologies such as Aspirin (1899), the electrocardiograph (1913), early vaccines (1870), and X-rays (1895) were quickly becoming mainstays of the medical world (Hajar 2015). Notably, immunology and headmatter are excluded due to lack of valid data, and the field of neurology exhibits extreme fluctuations with a slight increase (r-value = 0.11), largely attributable to its later development as a field in the mid-late 20th century.

Conversely, a less coherent group of topics exhibit moderate to strong upward trends in cognitive surprise, reflecting general divergence, the strongest of which are shown below.

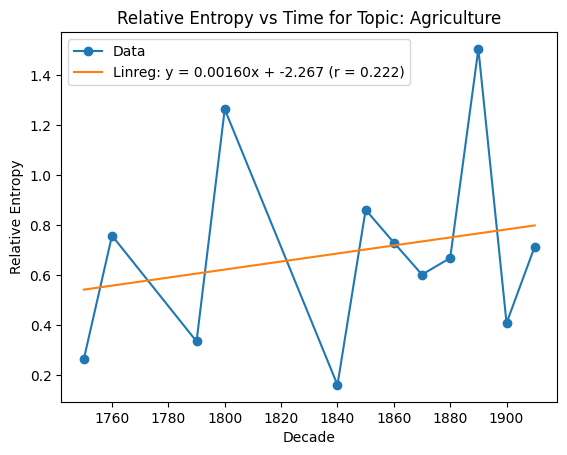
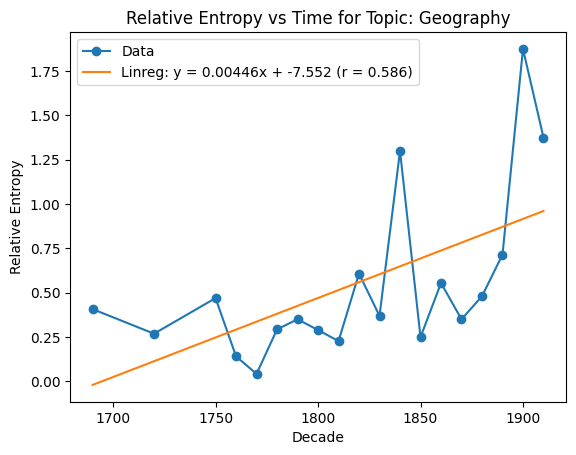
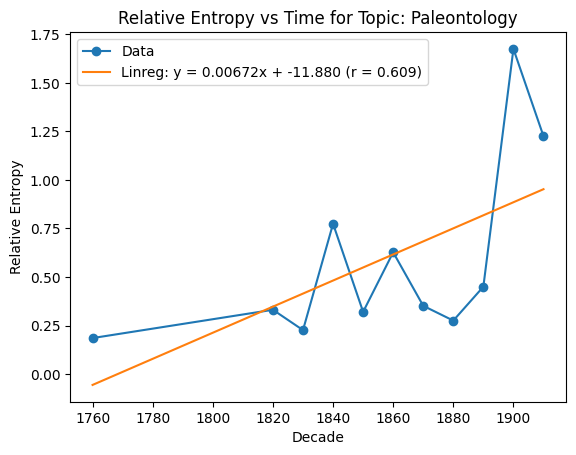


Fig. 7: KLD vs time plots for the 3 strongest upward-trending topics

Though these topics are seemingly disparate, the topics they represent are all extremely broad and lack what could be described as a central theory: there are no fundamental trios of laws for paleontology, geography, or agriculture the way there are for thermodynamics and Newtonian mechanics. These fields seemingly will expand endlessly as new species are discovered in fossils, more climatological and geological phenomena are analyzed (even today, data in those fields is scanty), and novel methods and species-level biology of agriculture is introduced. This overall shows support for the correlation between cognitive surprise and general expansion or diversification within a field without the establishment of an underlying backbone theory or vocabulary. Notably the upward trend in agriculture is relatively weak, and other fields (for instance meteorology, r-value -0.13 and astronomy, r-value -0.15) that intersect physical science more directly exhibit weak convergence rather than the divergent behavior described above.

Overall, topic-wise analysis over the full period of the RSC shows strong support for convergence in physical and medical science via the development of central theory and vocabulary. Conversely, there is support for divergence in fields that exhibit rapid diversification without the creation of a linguistic and theoretical base. Some fields that fall at the intersection of those two designations show noisy fluctuations in entropy, as would be expected given the competition between the two aforementioned trends.

The results also show varied support for the mapping of historical events to features of the KLD curve. One of the fields that most strongly supports this hypothesis is thermodynamics.

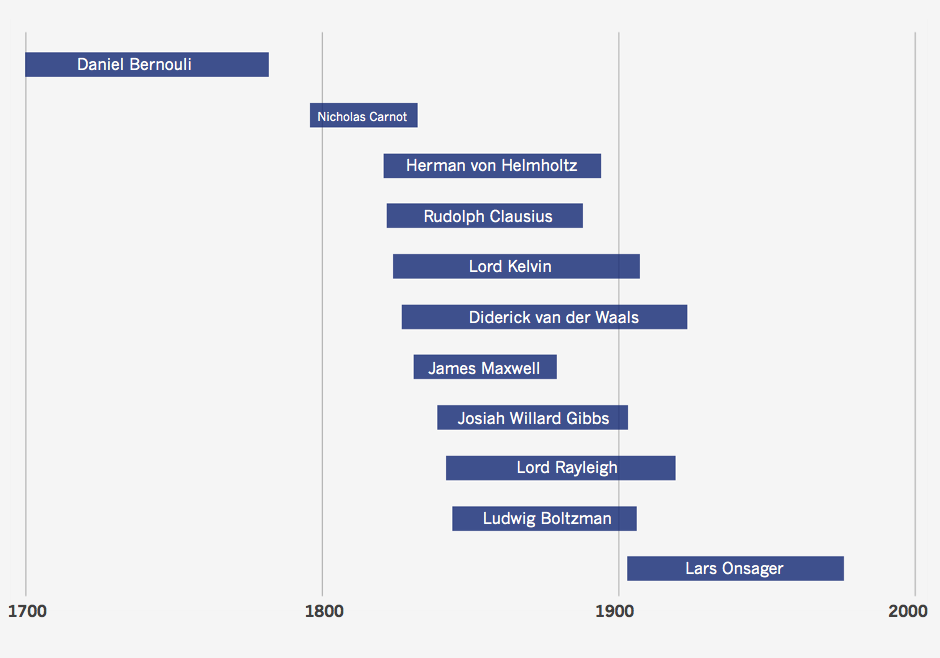
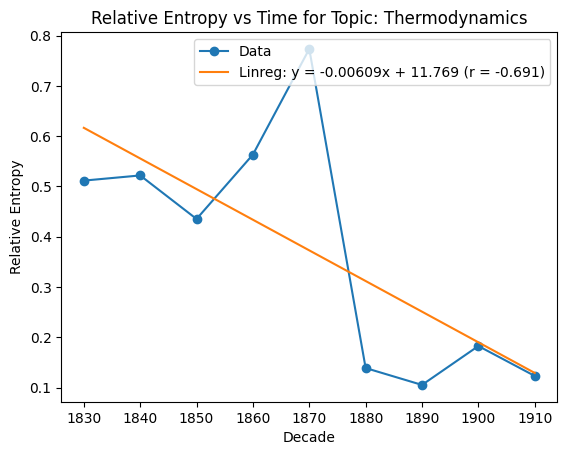


Fig. 8: KLD vs time for thermodynamics (left) and lifetimes of influential scientists (right)[[2]](#footnote-2)

During the mid 19th century, the fundamental laws of thermodynamics were quickly and coherently developed by an array of brilliant scientists, turning thermodynamics from a vague field into a clear set of rules for physical and chemical processes. Notably, such theories drove innovations in engine technology crucial to the Second Industrial Revolution of the late 19th and earliest 20th centuries. Linguistically, these developments quickly introduced a new set of words and laws into the literature, resulting in a boom in cognitive surprise during that period. Once these laws were accepted, however, a reader of thermodynamics would have come accustomed to seeing words like enthalpy, entropy (in the thermodynamic sense, not the linguistic one), and free energy drastically reducing their degree of cognitive surprise heading into the early 20th century. The strong degree to which the KLD curve matches this qualitative description of this history of thermodynamics shows support for the relationship between the history of the field and the linguistics of its writings.

It should be noted that the nature of the history of thermodynamics– a sharply defined period of rapid development to transform a vague field into concise, concrete laws– is seldom encountered in other areas examined by this study. Note fluid dynamics, for instance, shown in Fig. 5. The plot shows consistent but noisy decline over the entirety of the analysis period, generally corresponding to increased convergence but with little room to map history to the curve. Subsequent analysis shows more limited support, but still a general correlation, between historical events and curve features in other fields.

The history of biochemistry also shows support for the theory that a boom in theoretical development followed by convergent theory should result in a spike followed by a decline in cognitive surprise. As shown in Fig. 6, biochemistry exhibits a spike in cognitive surprise in the 1890s. This is the decade of the first biochemical process successfully run outside of a living cell (1896). Though much of the convergent theory is developed outside the window of texts analyzed in this study, the years following that build on the newly developed germ theory of disease (~1870) and the aforementioned experiment (Talabis 2016). In an example of the relationship between historical events and an upward-correlated KLD plot, the divergence in paleontology in the second half of the 19th century mirrors a massive uptick in fossil discoveries. Significant neanderthal discoveries occurred in 1848, 1856, and 1886. The first fossil evidence of early Homo sapiens was discovered in 1868, and a Homo erectus skull cap was unearthed in 1891 (Dorey 2019). These represent some of the most significant fossil discoveries for the study of human ancestry, and correspond to a massive boom in cognitive surprise in the paleontology plot of Fig. 7.

It is worth reiterating that many fields analyzed in this study, for a variety of reasons, do not exhibit the relationship between KLD and historical events seen in the fields mentioned above. Even some of the fields that represent some of the strongest overall trends with respect to correlation are too noisy to map curve features back to historical events, i.e. fluid dynamics. Some fields, particularly upward trending ones, are just too nonlinear in their histories to determine which events would propagate throughout the literature and cause linguistic change, i.e. geography. When noise is low and history is well defined, though, there is support that curve features do map at least loosely to historical events, or alternatively stated, that developments within a field propagate throughout the literature in the form of diachronic linguistic change which manifests in KLD changes.

**CONCLUSION**

Preliminary analysis of RSC metadata corroborates previous studies’ results that the corpus diversifies in terms of the total number of topics present in a given decade as well as the distribution of texts/topic in a given decade. Quantitative analysis of the RSC using statistical preprocessing and KLD shows strong support for convergence in fields that consolidate around central theories and/or vocabularies, and conversely divergence in fields that expand via discovery based diversification without such an established center. Some fields are overwhelmed by noise and yield no statistically significant result, and others likely fall somewhere between the two categories defined above resulting in weak correlations in either direction. Only a few fields are conducive to a mapping of historical events to KLD curves. Such analysis demands relatively low noise, a well defined history and boundary of the field, and a noticeable overall trend. The few fields that do exhibit these traits show support for the phenomenon of increased cognitive surprise during periods of rapid development. In cases where such development represents theoretical convergence, the entropy rapidly falls off resulting in a spike. In cases where such development represents diversification, the cognitive surprise remains high in a plateau or stable increase. The support for generalizing those feature-history relationships is non-negligible but limited, and is likely to be corroborated by any further inquiry that would mitigate noise. Such inquiries would also likely show support for the convergent-divergent distinction mentioned earlier. This study generates insight with a novel combination of breadth and depth with regards to the relationship between historical events and diachronic linguistic change. Overall, the study establishes support for a topic-specific relationship between cognitive surprise and historical events, and supports a high degree of confidence that additional statistical analysis designed to reduce noise and increase resolution will further bolster those positions.

**REFERENCES**

1. Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. Science, 331(6014), 176–182.
2. Sun, K., Liu, H. & Xiong, W. The evolutionary pattern of language in scientific writings: A case study of Philosophical Transactions of Royal Society (1665–1869). Scientometrics 126, 1695–1724 (2021). <https://doi.org/10.1007/s11192-020-03816-8>
3. Kermes, Elke. "The Royal Society Corpus: From Uncharted Data to Corpus." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA),
4. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993–1022.
5. Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.
6. Dias Laércio, Gerlach Martin, Scharloth Joachim and Altmann Eduardo G.. 2018 Using text analysis to quantify the similarity and evolution of scientific disciplines. R. Soc. open sci. 5171545. 171545. <http://doi.org/10.1098/rsos.171545>
7. Biber, D., & Gray, B. (2016). Grammatical complexity in academic english: Linguistic change in writing. Cambridge: Cambridge University Press.
8. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86.
9. Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (pp. 22–33).
10. Talabis, W. (2016, August 28). Timeline on history of Biochemistry. Academia.edu. ​​<https://www.academia.edu/28065020/TIMELINE_ON_HISTORY_OF_BIOCHEMISTRY>
11. Dorey, F. (2019, December 23). A timeline of fossil discoveries. The Australian Museum. <https://australian.museum/learn/science/human-evolution/a-timeline-of-fossil-discoveries>
12. Hajar R. (2015). History of medicine timeline. Heart views : the official journal of the Gulf Heart Association, 16(1), 43–45. <https://doi.org/10.4103/1995-705x.153008>

1. https://www.textfixer.com/tutorials/common-english-words.txt [↑](#footnote-ref-1)
2. http://theinformationdiet.blogspot.com/2012/08/timeline-of-thermodynamics.html. [↑](#footnote-ref-2)