Published in: Eder, Maciej, Rybick, Jan (2016): Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 496-500.

Authors: Peter Fankhauser (IDS Mannheim), Jörg Knappen, Elke Teich (beide von der Universität des Saarlandes)

Topical Diversification Over Time In The Royal Society Corpus

Science gradually developed into an established sociocultural domain starting from the mid-17 th century onwards. In this process it became increasingly specialized and diversified. Here, we investigate a particular aspect of specialization on the basis of probabilistic topic models. As a corpus we use the Royal Society Corpus (Khamis et al. 2015), which covers the period from 1665 to 1869 and contains 9015 documents [1].

We follow the overall approach of applying topic models to diachronic corpora (Blei and Lafferty 2006, Hall et al. 2008, Griffiths and Steyvers 2004, McFarland et al. 2013, Newman and Block 2006, Yang et al. 2011) to map documents to topics. Probabilistic topic models (Steyvers and Griffiths 2007) have become a popular means to summarize and analyze the content of text corpora. The principle idea is to model the generation of documents with a randomized two-stage process: For every word wi in a document *d* select a topic zk from the document-topic distribution P(zk|d) and then select the word from the topic-word distribution P(wi|zk). Consequently, the document-word distribution is factored as follows:

P(wi|d)=∑kPwi|zkPzk|d.

This factorization effectively reduces the dimensionality of the model for documents, improving their interpretability: Whereas Pwi|d requires one dimension for each distinct word (10s of thousands) per document, Pzk|d only requires one dimension for each topic (typically in the range of 20-100). Topics are thus not given explicitly for each document, but constitute *latent* variables: A variety of approaches exist to estimate the document-topic and topic-word distributions from the *observable* document-word distributions. We use Gibbs-Sampling as implemented in Mallet (McCallum 2002).

For the preliminary analysis in this paper, we process documents as is, without segmenting them further into pages, only excluding stop words but not performing lemmatization or normalization in order to stay reasonably close to the original source. We experimented with the number of topics ranging between 20 and 30, reporting here results on 24 topics. Cursory analysis of multiple runs with different seeds (Steyvers and Griffiths 2007) shows that the resulting topics are rather stable.

Table 1 displays the top words for the topics with manually assigned labels and their overall percentage of occurrence. We can roughly distinguish four groups of topics; three non-thematic groups and one thematic. The first group comprises topics arising from documents in *Latin* and *French*, some of which are also translated into English. The second group *Formulae* and *Tables* relates to highly formalized modes of information presentation. The third group of topics is also clearly non-thematic but relates to general scientific processes: *Observation* and *Experiment* both contain rather general verbs and adjectives in addition to nouns. *Events* contains words describing remarkable events. *Headmatter* includes formulaic expressions typically

occurring at the beginning and end of documents that are letters. All topics in this group are relatively frequent. Finally, the topics in the fourth group (*Geography* through *Chemistry*), consisting mainly of nouns, indeed have a fairly clear thematic interpretation.

Label	Words	%
Latin	quae quam sed ab sit vero hoc ac sunt esse qui etiam autem pro erit inter quo aut sive	6.4
French	la le les des en du par dans qui il une qu pour ou ce sur ne au je	1.3
Formulae	cos equation sin equal series point equations number line terms form values curve	4.7
Tables	weight water oo oz parts gr grain io grains fat increase weights grs passed urine specific	1.7
Observation	great made make parts found body time small part water nature long good put find	10.4
Experiment	present general subject case results similar nature author state result cases fact	7.3
Events	great time account stone ground house fire letter place miles found side stones	5.9
Headmatter	years year author society age number time royal life great letter account part letters	5.4
Geography	water sea tide high found river coast north land tides miles height surface great level	3.1
Meteorology	day ditto rain wind cloudy weather fair clear april year days night march july june	3.2
Botany	leaves plant plants tree tab bark folio foliis trees seeds seed flowers species fruit leaf	2.9
Reproduction	cells animal blood fluid eggs membrane found egg part animals ova size young	3.0
Cells	fibres structure form surface portion cells anterior part section side posterior	2.7
Paleontology	part bone bones teeth surface upper side lower anterior length posterior tooth large	2.5
Physiology	blood heart muscles part animal nerves vessels left parts stomach bladder body	5.5
Galaxy	distance position stars star obs small hill double equatorial vf diff st magnitudes nebula	1.6
Terrestrial Magn.	observations needle ship magnetic direct force made variation observed north diurnal	2.6
Solar System	sun time observations moon made observed difference observation clock latitude	5.5
Thermodyn.	air water heat temperature experiments tube experiment glass made time mercury	4.2
Mechanical	made length weight end diameter iron instrument experiments	4.5

Eng.	brass part point line	
Electromagn.	force electricity current wire action body power direction fluid motion surface effect	3.8
Optics	light rays glass eye red colours spectrum colour surface lines angle white blue object	3.7
Metallurgy	water acid salt grains quantity iron found solution colour substance experiments gold	4.8
Chemistry	acid water solution gas oxygen hydrogen carbonic cent action obtained salt potash	3.4

Table 1: Top words and percentages for topics

To investigate topical trends in the corpus we follow the approach in (Hall et al. 2008), by averaging the document-topic distributions for each year y:

$$P(z_k|y)=1/n\sum_{d_j\in y}P(z_k|d_j)$$

with n the number of documents in a year. *Figure 1* shows a selection of five topics with the most pronounced change over time. Interestingly, some of the major changes occur for non-thematic topics: The topic *Observation* declines sharply from over 30% to less than 1 %. The topics *Experiment* and *Formulae* on the other hand increase starting around 1750. This indicates a substantial paradigm shift over time. Indeed, as Gleick (2010) vividly describes, the early stages of the Royal Society were largely devoted to observing and reporting about natural phenomena. The non-thematic topic *Latin* reaches its peak in the early 18 th century, and the thematic topics *Cells* and *Chemistry* show a clear increase with the beginning of the 19 th century.

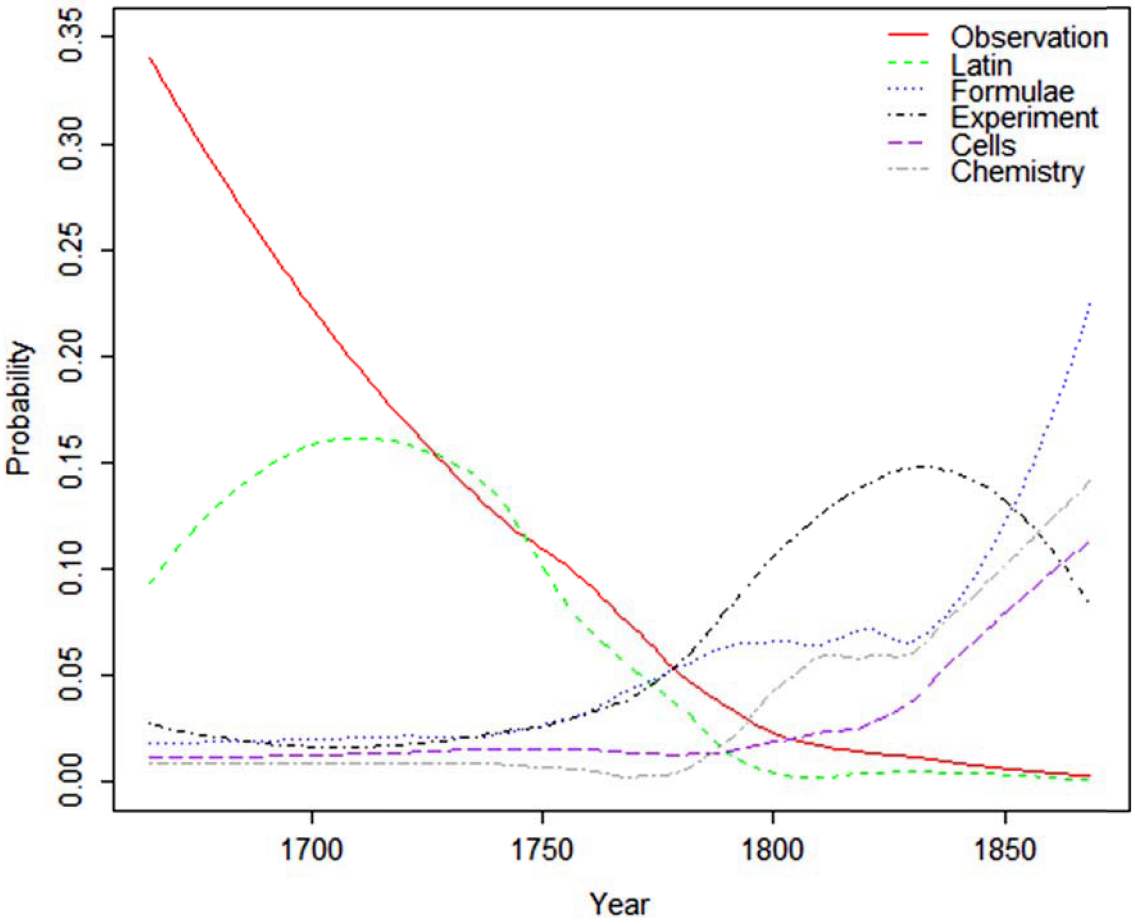

Figure 1: Major topical trends for selected topics

To gain a better understanding about the correlation of topics, we cluster them hierarchically on the basis of the Jensen-Shannon divergence between the topic-document distributions:

Pd|z=Pz|d/∑jPz|dj

Topics that typically co-occur in documents have similar topic-document distributions, and thus will be placed close in the tree.

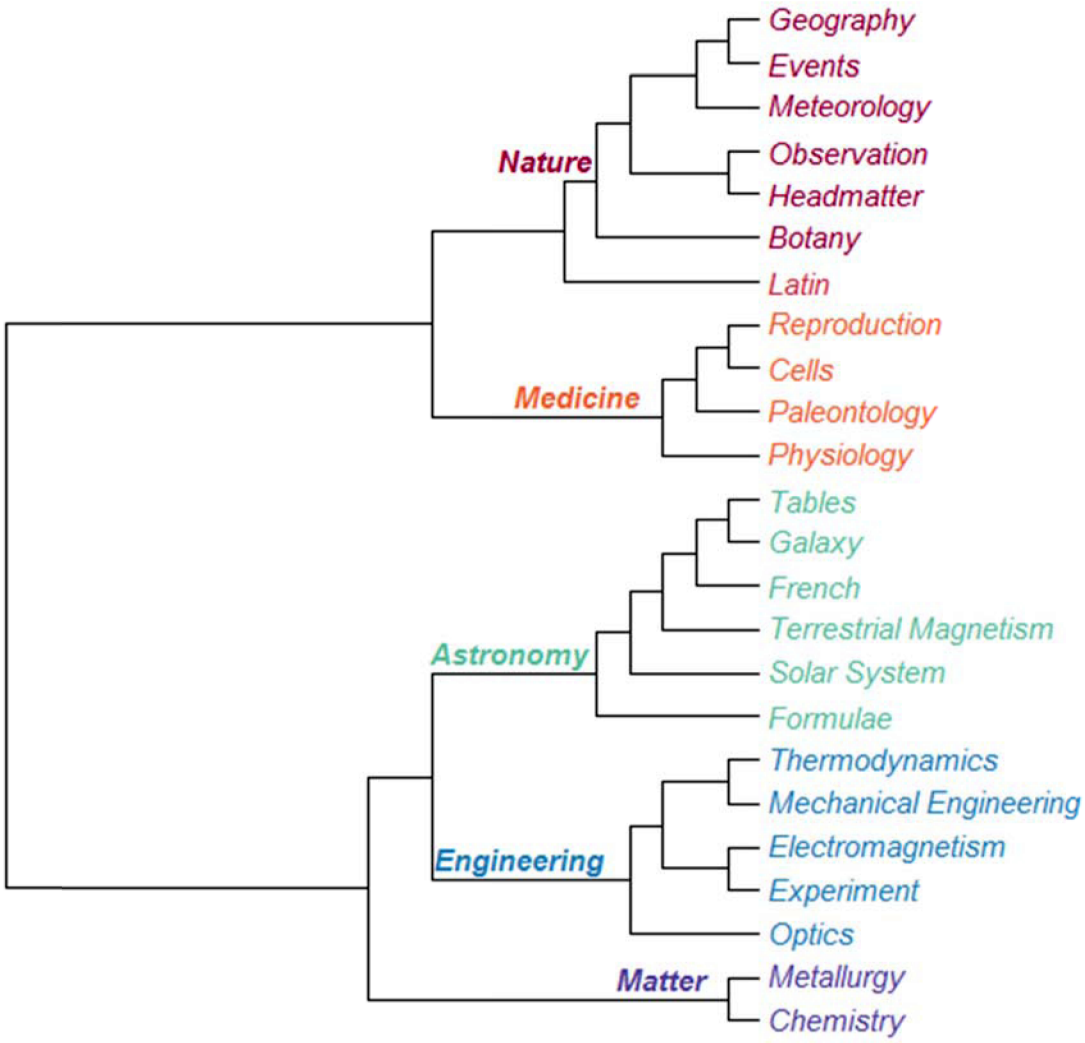

Figure 2: Hierarchical clustering of topics by their topic-document distribution

The resulting tree in *Figure 2* indeed identifies meaningful subgroups. Cutting the tree into six groups - *Nature, Latin, Medicine, Astronomy, Engineering, Matter* - allows us to investigate the overall topic distribution over time (*Figure 3* with *Latin* left out):

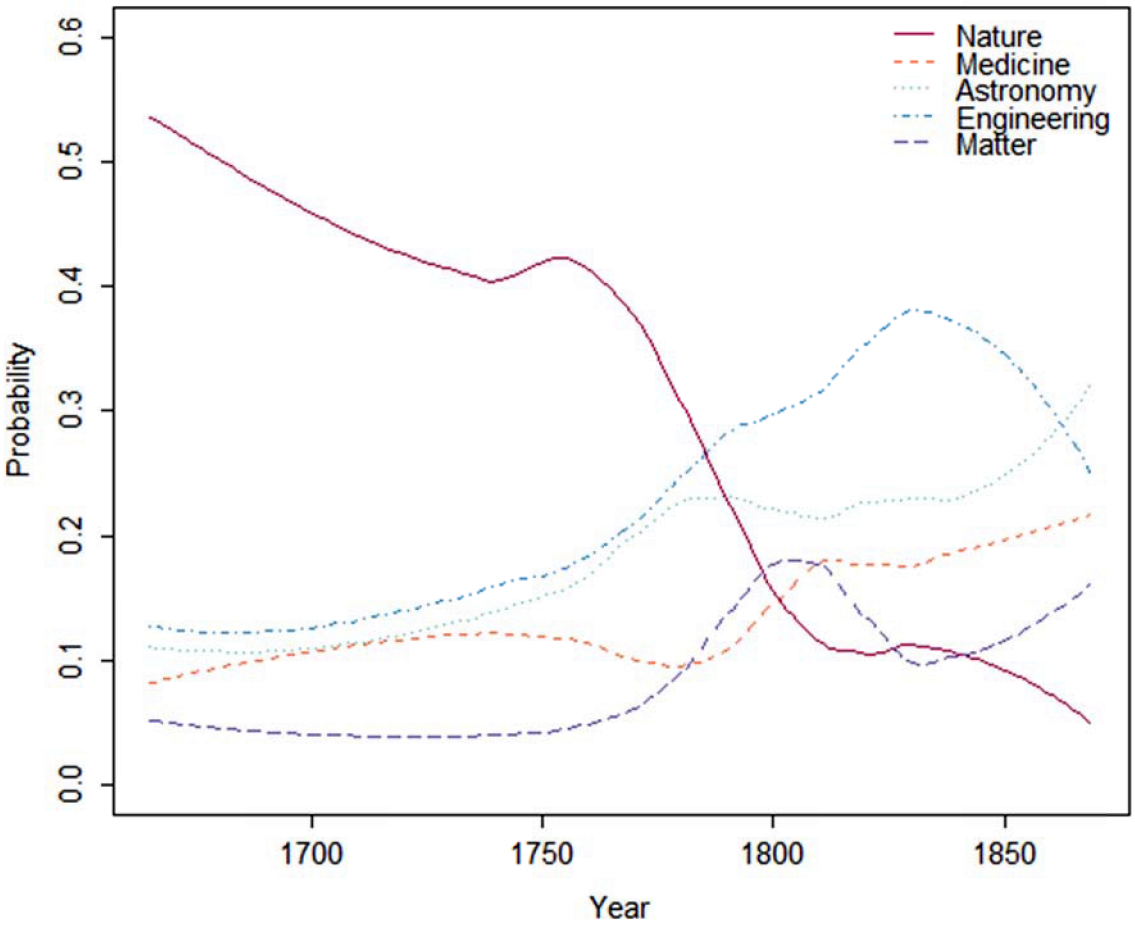

Figure 3: Distribution of topic groups over time

The topic group *Nature* comprising reports of all kinds of natural phenomena (Gleick 2010) clearly decreases over time, which is partially to be attributed to the strong decrease of the topic *Observation* in this group. The topic groups *Medicine* and *Astronomy* increase over time, whereas the topic groups *Engineering* and *Matter* also generally increase but with some intermediate peaks. Similar to the overall trends at the level of individual topics (*Figure 1*), the biggest overall change occurs in the 2 nd half of the 18 th century.

Looking at the individual trends together, *Figure 3* clearly indicates topical diversification: Until around 1770, the dominance of the topic group *Nature* leads to a highly skewed distribution of topic groups, whereas after 1770 topic groups are distributed much more evenly. The amount of skew can be characterized by the Shannon-Entropy:

$$H(P_y) = -\sum_k P_{z_k|y} \log_2 P_{z_k|y}$$

of the year-topic distributions P(zk|y) (Hall et al. 2010), with highly skewed distributions having low entropy. Indeed as can be seen in *Figure 4 (left),* the entropy (*ent*) increases fairly consistently during the 18 th century and levels out during the 19 th century, reflecting a general increase of topical diversity over time.

It is interesting to compare this with the mean entropy of the *individual* document-topic distributions (*ment*):

$$\mathrm{HmeanPy} := 1/n \sum_{dj \in y} H(\mathrm{Pdi})$$

with n the number of documents dj in year y. This measure decreases over time, i.e., while the overall topical diversity increases, the individual documents become more specific in terms of their topic distributions.

The difference between the entropy of year-topic distributions and mean entropy of individual document-topic distributions,

$$\mathrm{JSPy} = H(\mathrm{Py}) - \mathrm{Hmean}(\mathrm{Py})$$

is the Jensen-Shannon divergence, which is usually applied to two distributions, generalized to the ntopic distributions of all documents published in year *y*. The two opposing trends of these quantities lead to a constantly increasing Jensen-Shannon divergence, with a particularly sharp increase between 1750 and 1800. *Figure 4 (right)* depicts similar trends based on the 24 individual topic distributions. At this level, the year-topic entropy (*ent*) shows less of a clear trend, but mean entropy (*ment*) also clearly decreases, and consequently the Jensen-Shannon divergence clearly increases. Thus, at both levels of abstraction we can observe a clear diversification of the topics assigned to the individual documents. This strongly indicates a growing separation of individual scientific disciplines over time.

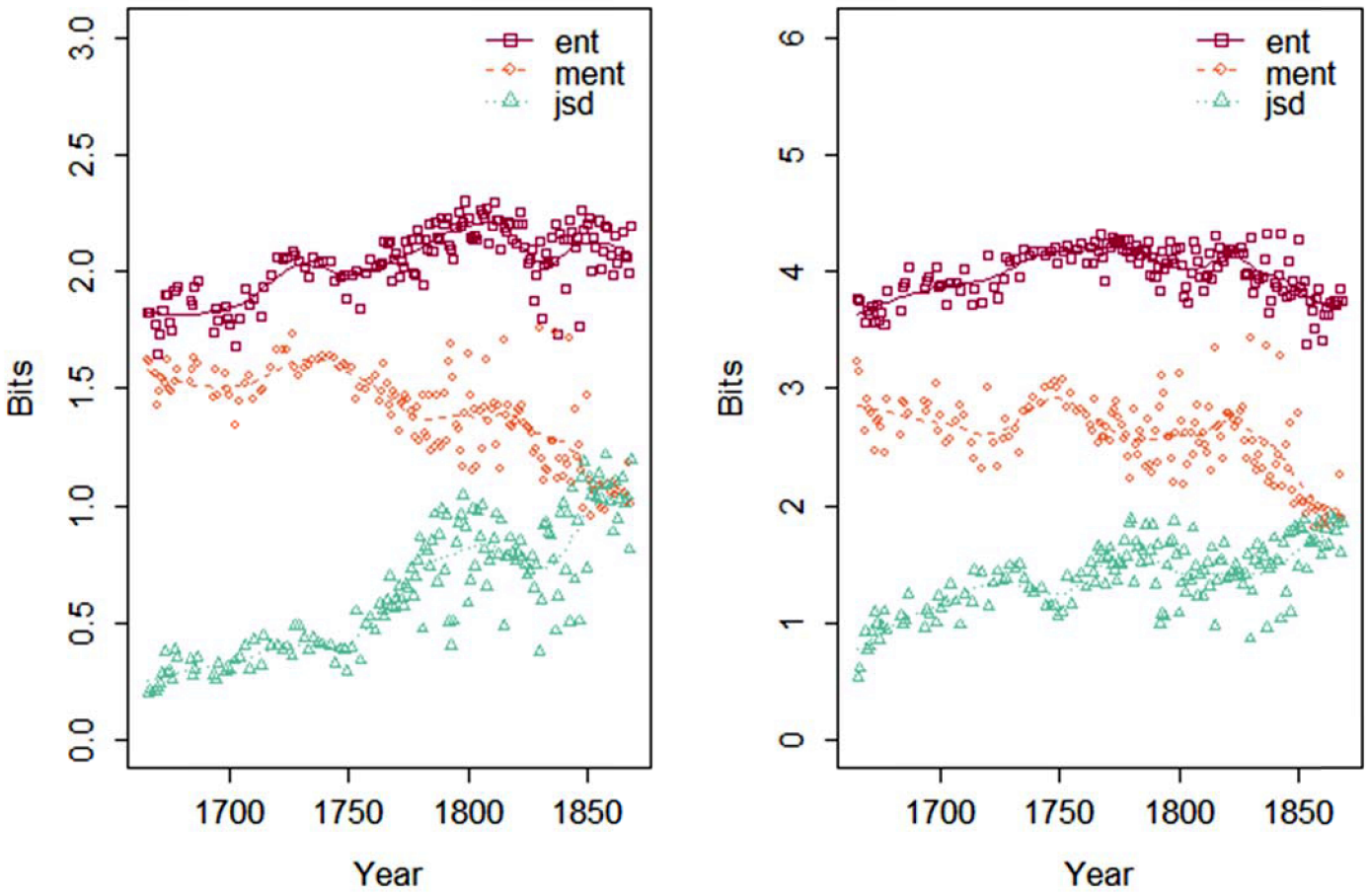

Figure 4: Entropy (ent), mean Entropy (ment), and Jensen-Shannon Divergence for topic groups (left) and individual topics (right)

As an alternative perspective on topical entropy *Table 2* gives examples of authors with more than 20 papers. The first three authors have the lowest entropy. The dominating topics for Cayley and Owen clearly characterize their main theme of work. Conversely, Rev. John Swinton's top topic *Headmatter* (62%) does not really reflect the overall theme of his publications (Orientalism), but rather their style as letters to members of the Royal Society – the dominant form of publication in this period. The second three authors have the highest entropy, their three top topics together amount for less than 50% of their overall topic distribution. However, they do characterize the main line of work of the authors in question fairly well.

author	Papers	Ent	Ment	Jsd	Years	Top Topics
Arthur Cayley	30	1.26	1.12	0.14	1850-1866	Formulae
Richard Owen	26	1.83	1.58	0.25	1843-1869	Paleontology
John Swinton	35	2.50	2.05	0.45	1753-1774	Headmatter

John Davy	58	4.05	3.33	0.72	1800-1856	Experiment, Chemistry, Physiology
William Watson	39	4.03	3.09	0.95	1739-1778	Events, Observation, Botany
Edmond Halley	65	3.93	2.75	1.18	1683-1731	Solar System, Observation, Latin

Table 2: Authors with minimum entropy (top) and maximum entropy (bottom)

In this paper we have analyzed the progression of topics in a corpus of the Royal Society of London. Our main result is the observation that the overall mixture of topics becomes more diverse over time, while the topics of individual documents become more specialized. These two opposing trends lead to a topical fragmentation of scientific discourse, which can be quantified by means of the generalized Jensen-Shannon divergence between the topic distributions of individual documents per time period. We are currently working on consolidating our analysis, experimenting with documents segmented into pages, focusing the analysis on different text types, and more carefully evaluating the resulting topic models (McFarland et al. 2013).

Of course, topic models only provide one, rather broad perspective on diversification of domain specific language. We plan to apply our approach also to other levels of linguistic analysis, such as terminology or grammar.

Bibliography

1. **Blei, D. and Lafferty, J.D.** (2006). *Dynamic topic models*. ICML.
2. **Gleick, J.** (2010). At the Beginning: More Things in Heaven and Earth: Bryson, B. (Ed.), *Seeing Further. The Story of Science and The Royal Society*. Harper Press, pp. 17-36.
3. **Griffiths, T. L. and Steyvers M.** (2004). Finding scientific topics. *PNAS*, 101 Suppl **1**:5228–35.
4. **Hall, D., Jurafsky, D., and Manning, C.D.** (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363-71.
5. **Khamis, A., et al.** (2015). A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus. *Corpus Linguistics 2015*. Lancaster.
6. **McCallum, A. K.** (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.
7. **McFarland, D. A., et al.** (2013). Differentiating language usage through topic models. *Poetics*, **41**(6): 607-25. http://dx.doi.org/10.1016/j.poetic.2013.06.004.
8. **Newman, D. J. and Block, S.** (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inf. Sci. Technol.*, **57**(6): 753-67. DOI=http://dx.doi.org/10.1002/asi.v57:6
9. **Steyvers, M. and Griffiths, T.** (2007). Probabilistic topic models. Landauer, T., et al.(Eds.), *Handbook of Latent Semantic Analysis*. Hillsdale, NJ: Erlbaum.
10. **Yang, T., Torget, A. J., and Mihalcea, R.** (2011). Topic modeling on historical newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH

'11). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 96-104.

Notes

1.

Of these 205 years, 159 years actually contain documents (mean = 56.7, median=36, sd=61.6, min=12, max=444)