Unique subfields of neuroscience exhibit more diverse language

Titipat Achakulvisut[*, 1], Daniel E. Acuña[2], Danielle S. Bassett[1, 3], Konrad P. Kording[1]

1 Department of Bioengineering, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
2 School of Information Studies, Syracuse University, Syracuse, New York, United States of America
3 Department of Electrical & Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

*** email: titipata@seas.upenn.edu**

Abstract

The fields within neuroscience differ in their heterogeneity and their variability of language usage. Some fields exhibit more diverse language while others write similarly to their neighboring fields. However, it is unclear how language diversity and similarity are intertwined. In this work, we propose that language differentiation of a sub-field promotes its language variability, a hallmark of exploration. We test this hypothesis using a dataset of recent abstracts from the Society for Neuroscience (SfN) conference. We find that a field that is more linguistically distant from its nearest field indeed has more within-field variability, even after controlling for field size and changes over time, suggesting more breadth of ideas.

Introduction

Neuroscience has matured rapidly since its beginnings in the 19th century [1] and it now consists of a diverse array of fields [2]. Each of these fields have their own, distinct set of phenomena to study, a common culture, and a shared language. For example in SfN, the field identified as *Molecular, Biochemical, and Genetic Techniques* develops tools to study the nervous system, whose structure has remained somewhat constant. But the field tends to utilize emerging techniques, e.g. sequencing, silencing, and most recently CRISPR/ Cas9 [3]. This technical evolution went along with linguistic variation which, in 2017, includes a great deal of words related to the immune system. But how does this technical evolution relate to the uniqueness of *Molecular, Biochemical, and Genetic Techniques* as field? Understanding this relationship is important for uncovering patterns of scientific exploration and maturity, exciting questions in science of science [4, 5].

The availability of increasingly large article databases has enabled various approaches to study the evolution of neuroscience. Previous researchers have studied a number of aspects in isolation: scientific communities within a specific country [6–9], particular sub-fields [10–14], specific diseases [15], or high impact neuroscience papers [16]. However, we still lack an understanding of language evolution in neuroscience. Recent datasets and Natural Language Processing (NLP) techniques [17] open the door to understand the underlying evolution of the language at larger scale.

Language has been used to understand scientific fields [18]. Natural language is more robust [19] than keywords [20–22] and immediate for analysis compared to citations [23–25]. In particular, language has been used to study the organization and evolution of science [18] and it is competitive for analysis compared to other more hard to get data such as citations. It is promising to track the evolution of language used in neuroscience.

One simple way of studying a field is by analyzing the variability of its vocabulary. This can be done by comparing word distributions within the field. Indeed, past research has found that as a field slowly evolves [18], its vocabulary use remains relatively distinct. While interdisciplinary research may blur the vocabulary differences between fields, specialization leads to ongoing differences. This raises the question about the variability of the language used in various fields within neuroscience.

Specifically, we are interested in fields which use more variability of language in comparison to others. One possibility may be that large subfields use more language variability; after all there are more scientists that can contribute to the language diversity. A second possibility may be that unique subfields (i.e., very different from their neighbors) may use more diverse language as they may have more space to explore into. Understanding the variability of vocabularies usage promises to shed light onto the language utilization and characteristics of discipline.

The Society for Neuroscience (SfN) conference is a representative place to understand the evolution of neuroscience. It is the biggest neuroscience conference with approximately 30,000 attendees. Each year, there are approximately 50,000 unique authors submitting about 12,000 abstracts. This is a remarkably large number if we consider that there are only about 35,000 papers per year in this discipline [22]. Posters being presented at SfN thus gives a good sample of the entire neuroscience discipline [26]. As such, the SfN conference may be an ideal place to analyze the variability of the language in its subfields.

Here, we wanted to quantify how neuroscience subfields diverge linguistically using abstracts from the SfN conference. We found that the language distance of the topics to its neighbor is a strong predictor of language variability of a field. Moreover, we found that the effects still persist in coarser scale such as a themes, and after controlling for field size. We argue that a field having a unique vocabulary may be a sign that it is distinct from neighboring areas, allowing it to have more breath of ideas.

Methods

We collect abstracts from the Society for Neuroscience conference over the course of five years from 2013 to 2017. For each submission, authors are allowed to provide theme, topic and subtopic. For example, a poster can have subtopic "D.03.i" (Pain models: Behavior) which is in topic "D.03" (Somatosensation: Pain topic) within the theme "D" (Sensory Systems). For year 2013 to 2015, each submission has only one subtopic but for 2016 and 2017, each submission may have up to two subtopics. For our analysis, we always pick the first subtopic listed. From year 2013 to 2017, we found 9487, 13138, 12943, 12166 and 10877 submissions, respectively. The conference organizers create the topics from which the authors can choose from. There are 80, 85, 81, 84, and 86 topics for each year in 2013 until 2017, respectively. These topics belong to 8 main themes including Theme A: Development, Theme B: Neural Excitability, Synapses, and Glia, Theme C: Neurodegenerative Disorders and Injury, Theme D: Sensory System, Theme E: Motor System, Theme F: Integrative Physiology and Behavior, Theme G: Motivation and Emotion, Theme H: Cognition, and Theme I: Techniques. Below, we describe how we preprocess the abstracts and analyze their content using a natural language processing.

Text pre-processing

We clean each abstract by removing all HTML tags. Each abstract is tokenized to create a list of words. We apply lemmatization and lower case all words in the list. We remove English stop words, rare words that appear less than 4 times across documents, and words that appear more than 90 percent across all documents. We only keep the top 25,000 words by frequency. These are all standard steps for text analysis [27]. The same trained transformation is applied to all experiments conducted in this study.

Measuring language distance between neuroscience topics

We measure the difference of probability distribution of word usage between neuroscience topics. We define the word distribution of a topic i as $p_i(w) \equiv \mathbf{p}_i$. To calculate the distance matrix between word distributions between two given topics, we apply a generalized Jensen-Shannon Divergence measure [17] to compute the difference of word distributions. This is, given word distributions between two topics i and j as $\mathbf{p}_i$ and $\mathbf{p}_j$ respectively, we can measure the distance between two topic distributions as

$$D_{\text{lang}}(i,j) = \frac{2H_2((\mathbf{p}_i + \mathbf{p}_j)/2) - H_2(\mathbf{p}_i) - H_2(\mathbf{p}_j)}{\frac{1}{2}(2 - H_2(\mathbf{p}_i) - H_2(\mathbf{p}_j))}, \tag{1}$$

where

$$H_2(\mathbf{p}_i) = 1 - \sum_w p_i(w)^2, \tag{2}$$

which is a generalized entropy of order 2. The distance measures cross entropy between two word distributions. The denominator is applied to normalize the distance between 0 to 1. The measure is shown to be consistent even in heavy-tailed distributions [17]. For non heavy-tailed distributions, the measure is equivalent to cosine similarity. Sometimes we compare fields with different sizes, which may affect the distances. To control for this effect, in all of distance computations we repeat n times the following sampling scheme: a fixed set of n_d documents is sampled from topic i and topic j, and we then we measure the distance between them. We define *language variability* as the average distance of these n simulations. We use $n_d = 1$ and $n = 5000$ for all experiments.

Measuring expert distance between neuroscience topics

As mentioned before, a submission has an author-provided subtopic, and we will use this expert classification to check the consistency of language distances. We use hamming distance to measure expert distance between two topics. The expert distance between topic i and topic j is denoted as

$$D_{\text{expert}}(i,j) = \begin{cases} 0, & \text{same topic} \\ 1, & \text{same theme, different topic} \\ 2, & \text{different theme} \end{cases} \tag{3}$$

The correlation between pairs of distance matrices is calculated using standard Pearson correlation.

Measuring of language variability in neuroscience topics

The language variability within the same year for topic i is denoted as

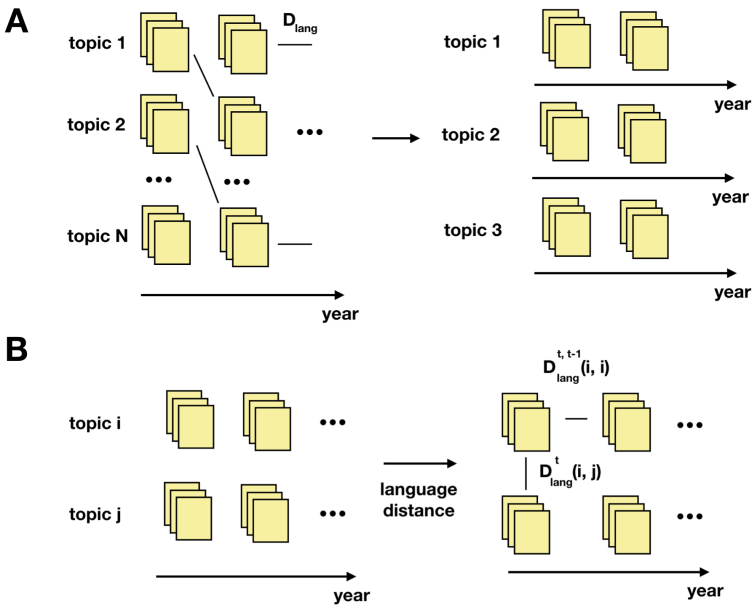

Figure 1. A. Tracking neuroscience topics over years. We track topics that remain unchanged across years at SfN. **B. Schematic of language distance calculation.** language distance between topics ($D^t_{\text{lang}}(i,i)$) and across topics ($D^t_{\text{lang}}(i,j)$) by year. It can also be calculated across years for the same topic ($D^{(t,t-1)}_{\text{lang}}(i,i)$).

$$\upsilon = D^t_{\text{lang}}(i,i). \tag{4}$$

We measure year-on-year language variability, υ_T, for each topic by calculating the average language variability between consecutive years

$$\upsilon_T = \frac{1}{T}\sum_{t=0}^{T} D^{(t,t-1)}_{\text{lang}}(i,i), \tag{5}$$

where t is time step of the measurement and T is total time step that we record.

Language distance to nearest topic can be measured by

$$\upsilon_{\text{neighbor}} = \min_{j\in N, j\neq i} D^t_{\text{lang}}(i,j), \tag{6}$$

where $D^t_{\text{lang}}(i,j)$ is language distance between topic i and j and N is the set of all topics. This measure can be interpreted as the uniqueness of the topic. The topic that lies further from its neighbor in language space means that they have distinct language usage.

Results

To ask how the uniqueness, theme, and size of a topic affects the language variability of a given topic, we use a large dataset of 58,611 abstracts from the SfN conference. For

each topic and year, we extract the probability distribution in terms of the vocabulary used in the abstracts. Calculating the difference between language probability distribution usage allows quantifying the usage of vocabulary of a given topic. Moreover, this will allow us to ask which aspects of topics correlate with its language variability.

Expert topic classification at SfN is highly correlated with topic language

For us to analyze topics, we first want to verify that topics are meaningfully defined. If they are, then human expert judgment, that ultimately gives rise to the classification at SfN, is correlated with the topic's text corpus. The distances implied by human classification (e.g. 2A) and the corresponding text distance (2B) are correlated (2C). We find high and consistent correlations between language distance and expert distance ($r \approx 0.7, p < 0.001$) (figure 2C). As a comparison, the correlation found between language and experts in Scopus is around 0.35 [18]. This means that the topics classification by SfN experts significantly correlates with language similarity.

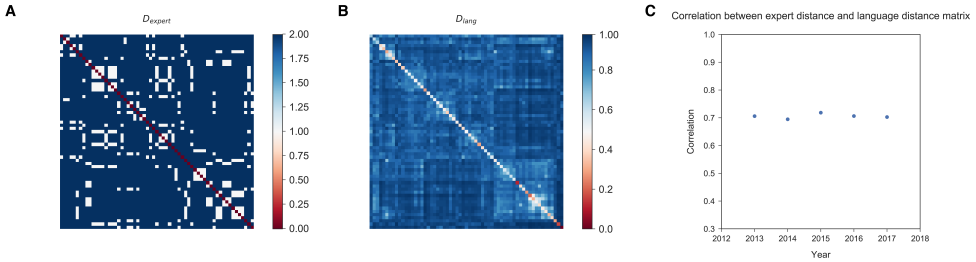

Figure 2. A. Expert distance matrix of topics in SfN year 2017. The distance matrix is calculated using hamming distance between expert classifications. **B. Language distance matrix of topics in SfN year 2017.** The distance matrix is calculated as the language distance between topics. **C. Correlation between expert and language distance matrix across years.** The correlation is consistent over years and has a value around 0.7. All correlations reported are significant with p-value less than 10^{-3}.

Language variation of neuroscience topics within Society for Neuroscience Conference

We measure language variability for the 61 topics that remain unchanged through the years (see Methods, figure 3). Topics that have low language variability include Eye Movement, Spinal Cord Injury and Plasticity, Biological Rhythms and Sleep, and Stroke. These fields also have low language variability across consecutive years. This means that the language usage of the given topics is less variable compared to other fields. Topics that have high language variability include Molecular Biochemical and Genetic Techniques, Neurotransmitters and Signaling Molecules, Neurogenesis and Gliogenesis, and Data Analysis and Statistics. These fields have seen large changes over the last couple of years with the advent of techniques like optogenetics [28], CRISPR/Cas9, and data analysis and statistics. Clearly some fields are more variable linguistically than other fields.

Large fields may have more language variability as compared to smaller fields. To check this idea, we measure the correlation between size of each topic and language variation within the year, v. We found small significant correlation between these two values ($r = 0.228, p < 0.001$) but the effect size is small (Pratt's effect size = 0.046) [29].

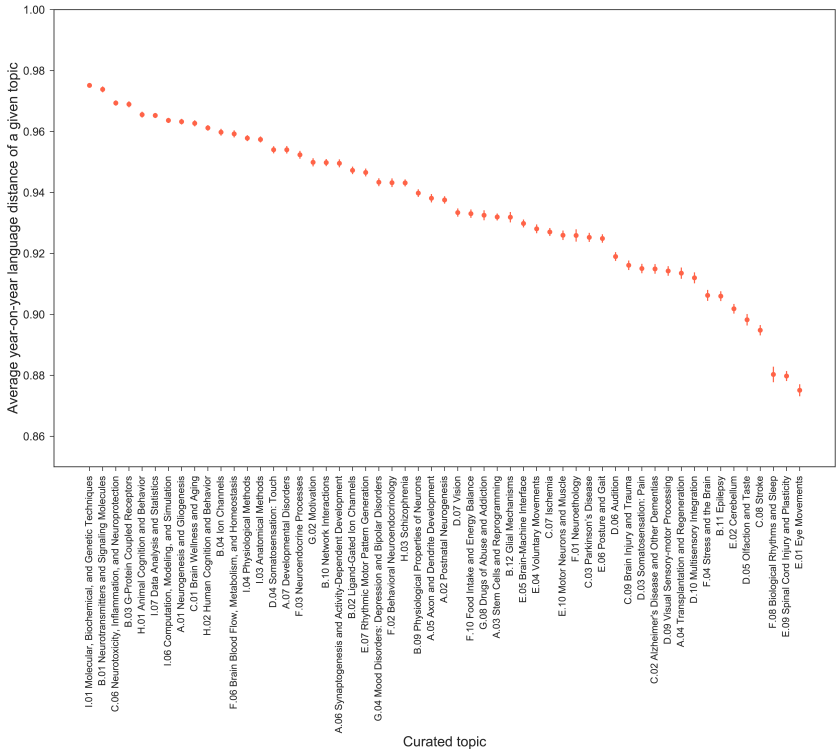

Figure 3. Average year-on-year language variability of a given topic. We plot a rank of year-on-year language variability of a given neuroscience topic from high to low language variation. The dots denote the average language distance over years and the error bara denote the standard error of the language distance. The measure is done by sampling documents from consecutive years and measuring language distance as described in the methods. This measure is independent of field size.

Topic uniqueness predicts language variability

Unique fields may express more language variation, and a simple proxy for uniqueness may be the vocabulary distance to the most similar topic. To check this idea, we assess the relationship between the within topic language variability, υ, the year-year topic variability, υ_T, and the language distance to nearest topic, $\upsilon_{\text{neighbor}}$. A Pearson correlation analysis reveals a significant correlation between υ_T and $\upsilon_{\text{neighbor}}$ ($r = 0.562, p < 0.001$, figure 4A, red dots). It thus does seem as if more unique fields have more language variation over consecutive years.

It could be that our findings simply reflect the different sizes of topics. To test this hypothesis, we perform a shuffle control where each poster is assigned to a random topic while keeping the topic size intact. We found small significant correlation in the shuffle control experiment ($r = 0.151, p = 0.023$, figure 4A, gray dots). This correlation is smaller than in the real data (one sided z-test $z(223) = -5.07, p_{\text{one-sided}} < 0.001$ using Fisher r-to-z transformation). Thus, the correlation between uniqueness and language innovation is not just a size artifact. Inspecting the distribution by eye suggests that the distribution may contain multiple linear trends. We thus compare our findings to those from a two-component mixture of linear regression models. However, we found that the data was better explained by only one trend ($BIC_{\text{mixture}} = 11.59$ vs. $BIC_{\text{single}} = 5.96$). A similar analysis was performed to establish a relationship between υ and $\upsilon_{\text{neighbor}}$,

and we found a consistent result ($r = 0.476, p < 0.001$, Figure 4B). Interestingly, the shuffle control correlation was significant ($r = 0.378, p < 0.001$) but smaller than in the real data (one sided z-test $z(223) = -1.31, p_{\text{one-sided}} = 0.095$ using Fisher r-to-z transformation) and the range of language distance to nearest topic was significantly different from the real data (Kolmogorov-Smirnov test $D = 0.780, p < 0.001$). This suggests that distance to other fields may encourage variability within the field itself.

Further, we wanted to control for other randomness that may be produced by random variations of language models. For this, we simulated a set of documents from different topics each sampled from a Dirichlet distribution with $\alpha = 0.001$. For each topic, we sampled a set of documents equal to the topic sizes seen in the data. In the simulation, we found no relationship between v and v_{neighbor} ($r = -0.0278, p = 0.831$). Therefore, the relationship that we did find in the data does not seem to be result of a random language distribution.

Consistent finding is found in coarser scale theme classification in neuroscience

We further analyzed whether the effects of nearest neighbor distance on language variability can be observed in larger-scale topics i.e. theme classification. For this, we performed a similar analysis to the one describe before but instead of considering variability within sub-topics we do so within themes in SfN. Themes are the highest level of organization of posters (see Materials and Methods) and therefore represent a coarse but meaningful organization of the posters. Consistent with the results described above, we did not find a strong correlation between the size of the theme and year-year language variability ($r = -0.023, p = 0.893$). Further, a Pearson correlation shows high correlation between distance to the nearest theme and year-year language distance within the theme ($r = 0.560, p < 0.001$). A shuffle control of the theme is conducted, removing this effect ($r = 0.014, p = 0.931$, figure 4A, gray dots). Finally, we find significant correlation between the variability within theme and distance between the theme and its neighbor ($r = 0.507, p < 0.001$). Similar to the finding in topic level, we still observe a significant relationship between them ($r = 0.331, p = 0.026$). However, distribution of the distance to nearest theme is different from real data (Kolmogorov-Smirnov test, $D = 0.467, p < 0.001$). These results suggest that the same trends found in the finer grain expert organization of topics is found at the coarser level classification.

We performed a similar verification of the theme effects on other correlations. We observed that there is correlation between the distance of theme to its neighbor and the language variability across years of the sub-topics of the theme. We found significant correlation between distance to the neighboring theme and year-year language distance of the topic ($r = 0.367, p < 0.001$). We do not find significant correlation in the shuffle control ($r = -0.041, p = 0.545$). This means that topics that belongs to the theme that is distant from its neighbor in language space also tends to have high language variability.

Discussion

In this study, we explore language variability of neuroscience sub-fields over five years of the Society for Neuroscience conference. We first show that language variation is highly correlated and consistent with expert classification. We found the correlation between expert classified topics and language distance between topics to be around 0.7 over the years. Compared to expert classification used in Scopus, our computed correlation is twice as high [18]. This means that topic classifications proposed by SfN organizers are

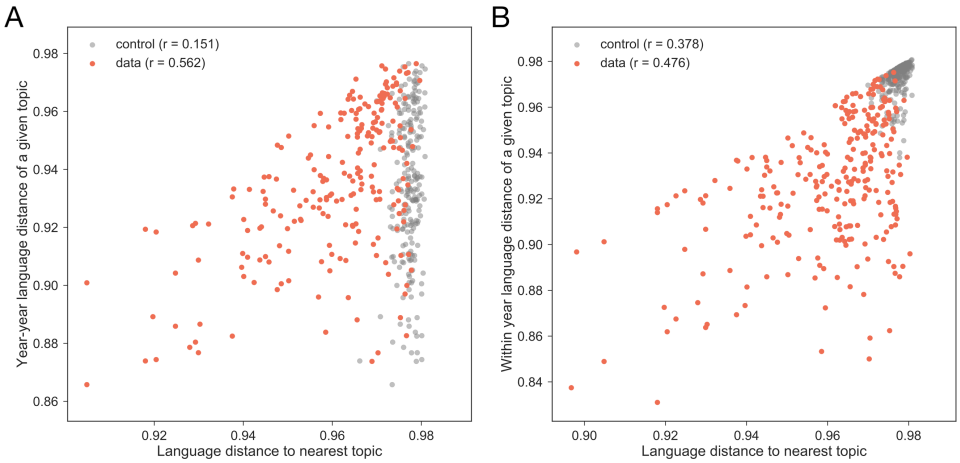

Figure 4. A. Relationship between year-on-year language distance of a given theme and language distance to its nearest neighbor. Year-on-year language variability within a given topic and language distance to nearest neighbor topic is shown (red dots). A shuffle control is also shown where topics with the same size are assigned randomly to the posters (red dots). We found a significant correlation between language distance to its nearest neighbor and language variability over consecutive years ($r = 0.562, p < 0.001$). The shuffle control reveals small significant correlation between two quantities ($r = 0.151, p = 0.023$, gray dots). **B. Relationship between within year language distance of a given topic and language distance to nearest neighbor.** We show language distance within year versus the language variability to the nearest neighbor. We found a significant relationship between them ($r = 0.476, p < 0.001$, red dots). For the control, we also find significant but smaller correlation ($r = 0.378, p < 0.001$, gray dots). However, distribution of language distance sto nearest neighbor topic is significantly different from what we would find in the data (Kolmogorov-Smirnov test, $D = 0.780, p < 0.001$).

a good signal to capture language variability. This suggests that measuring variability using language is a robust framework for studying neuroscience fields.

Additionally, we rank neuroscience topics by language variability. Example topics with low language variability includes Eye Movement, Vestibular System, Spinal Cord Injury and Plasticity, and Stroke. Example topics with high language variability includes Molecular Biochemical and Genetic Techniques, Neurotransmitters and Signaling Molecules, Neurogenesis and Gliogenesis and Data Analysis and Statistics. We observe that topics that change less tend to revolve around specific systems and functions. On the other hand, topics that change more revolve around system-level analysis, high level functions, or new techniques. Moreover, we control that this correlation is not driven by topic size. We only observe small a correlation between topic size and language variability. To the extent of our knowledge, no other research group have found that language varies in this way in neuroscience. More importantly, we show that topics with language that are far away from their neighboring fields have large language variability within topic. Again, this effect remains after controlling for random properties of the language such as size and word distribution. Taken together, these results suggest that topics that are further away from other topics or more unique topic of study seem to expand linguistically. We interpret this as a principled way of exploring scientific topics.

We have made our conclusions about neuroscience based on a large conference over a

relatively reduced time frame. It is possible that the topics presented at conferences significantly differ with those published in journals [22]. Also, the time scale of changes and exploration in neuroscience are probably much longer than the five years we analyzed. For example, topics that are older have a better-established language base but are still exploring a great deal. However, we think it is likely that scientists use the same language variability when writing journal abstracts as those published in conference abstracts. Also, because we found a high correlation between within-year language variability and year-on-year linguistic variation of a given topic, the effect of time on variability is somewhat already predicted by the language variability within the topic. Future studies will aim at expanding the source and time frame of publications to verify these hypotheses, especially whether the age of topic produces some of our main results.

Language is a limited source of information for studying the organization and changes of knowledge. For example, when low language variability is encountered, it is possible that other sources of variability persist, such as citations, keywords, or funding. Therefore, we need to further understand other internal data indicative of variability as well as external signals. However, as other researchers have found [18], there is surprising consistency between high language variability and expert classification variability. We will explore this issue further in future studies.

Conclusion

In this study, we show that text analysis can reveal patterns of language exploration and differentiation across fields of neuroscience. Fields that are distinct from other fields in language space also increase their language variability. These results hold under a number of controls, such as field sizes, word richness, and time. In summary, our results show that language can be effectively used to quantify scientific evolution.

Supplementary

This section contains mathematical definitions used in the manuscript.

Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) [30] is used for model selection in mixture of linear regression in our study. BIC can be calculated as follows

$$\mathrm{BIC} = \log(n) \times k - 2\log(\hat{L}) \tag{7}$$

where n is number of data points, k is number of parameters estimated by the model and $\hat{L}$ is maximized value of log-likelihood of the model for the given data. The lower BIC means the better fit with fewer number of parameters. We can conclude that one model has strong preference against another if ΔBIC is greater than 6.

Acknowledgement

Titipat Achakulvisut was supported by the Royal Thai Government Scholarship grant #50AC002. Daniel E. Acuna was partially supported by the NSF award #1646763. We thank SfN and Coe-Truman Technologies for providing the abstract dataset for our analysis.

References

1. Wickens AP. A history of the brain: from stone age surgery to modern neuroscience. Psychology Press; 2014.
2. Shulman RG. Brain imaging: What it can (and cannot) tell us about consciousness. Oxford University Press; 2013.
3. Doudna JA, Charpentier E. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346(6213):1258096.
4. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science. 2018;359(6379):eaao0185.
5. Evans JA, Foster JG. Metaknowledge. Science. 2011;331(6018):721–725.
6. Bala A, Gupta B, et al. Mapping of Indian neuroscience research: A scientometric analysis of research output during 1999-2008. Neurology India. 2010;58(1):35.
7. Glänzel W, Danell R, Persson O. The decline of Swedish neuroscience: Decomposing a bibliometric national science indicator. Scientometrics. 2003;57(2):197–213.
8. Persson O, Danell R. Decomposing national trends in activity and impact. In: Handbook of quantitative science and technology research. Springer; 2004. p. 515–528.
9. Xu W, Chen YZ, Shen ZC. Neuroscience output of China: A MEDLINE-based bibliometric study. Scientometrics. 2003;57(3):399–409.
10. Robins RW, Gosling SD, Craik KH. An empirical analysis of trends in psychology. American Psychologist. 1999;54(2):117.
11. Li T, Ho YS, Li CY. Bibliometric analysis on global Parkinson's disease research trends during 1991–2006. Neuroscience letters. 2008;441(3):248–252.
12. Bruer JT. Can we talk? How the cognitive neuroscience of attention emerged from neurobiology and psychology, 1980–2005. Scientometrics. 2010;83(3):751–764.
13. Robert C, Wilson CS, Donnadieu S, Gaudy JF, Arreto CD. Evolution of the scientific literature on pain from 1976 to 2007. Pain Medicine. 2010;11(5):670–684.
14. Yeung AWK, Goto TK, Leung WK. A bibliometric review of research trends in neuroimaging. Current Science (00113891). 2017;112(4).
15. Bishop DV. Which neurodevelopmental disorders get researched and why? PLoS One. 2010;5(11):e15112.
16. Yeung AW, Goto TK, Leung WK. At the leading front of neuroscience: a bibliometric study of the 100 most-cited articles. Frontiers in human neuroscience. 2017;11:363.
17. Gerlach M, Font-Clos F, Altmann EG. Similarity of symbol frequency distributions with heavy tails. Physical Review X. 2016;6(2):021009.
18. Dias L, Gerlach M, Scharloth J, Altmann EG. Using text analysis to quantify the similarity and evolution of scientific disciplines. arXiv preprint arXiv:170608671. 2017;.

19. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. Journal of clinical epidemiology. 2010;63(11):1205–1215.

20. Chang YW, Huang MH, Lin CW. Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. Scientometrics. 2015;105(3):2071–2087.

21. Chavalarias D, Cointet JP. Phylomemetic patterns in science evolution—the rise and fall of scientific fields. PloS one. 2013;8(2):e54847.

22. Yeung AWK, Goto TK, Leung WK. The Changing Landscape of Neuroscience Research, 2006–2015: A Bibliometric Study. Frontiers in neuroscience. 2017;11.

23. Kuhn T, Perc M, Helbing D. Inheritance patterns in citation networks reveal scientific memes. Physical Review X. 2014;4(4):041036.

24. Herrera M, Roberts DC, Gulbahce N. Mapping the evolution of scientific fields. PloS one. 2010;5(5):e10355.

25. Porter A, Rafols I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics. 2009;81(3):719–745.

26. David SV, Hayden BY. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. PloS one. 2012;7(10):e46608.

27. Manning C, Raghavan P, Schütze H. Introduction to information retrieval/Christopher D. Cambridge University Press, Cambridge, England; 2009.

28. Deisseroth K. Optogenetics. Nature methods. 2011;8(1):26–29.

29. Yin P, Fan X. Estimating R 2 shrinkage in multiple regression: a comparison of different analytical methods. The Journal of Experimental Education. 2001;69(2):203–224.

30. Schwarz G, et al. Estimating the dimension of a model. The annals of statistics. 1978;6(2):461–464.