Triangulation of History Using Textual Data

Kenneth D. Aiello, *Arizona State University*
Michael Simeone, *Arizona State University*

Abstract: This essay describes essential considerations and select methods in computational text analysis for use in the study of history, specifically the history of science and biomedicine. It explores specific approaches that can be used for understanding conceptual change over time in a large corpus of documents. By way of example, using a corpus of 27,977 articles collected on the microbiome, the essay studies the general microbiome discourse for the years from 2001 to 2010, examines the usage and the sense of the word "human" from 2001 to 2010, and highlights shifts in the microbiome discourse from 2001 to 2010.

TEXT ANALYSIS

Computational text analysis offers insights into the social, linguistic, and historical context of historical events.[1] Active participation by historians in text analysis enriches the growing community of scientists, researchers, and subject experts who are engaged in novel historical research projects. Previous historical projects embracing text analysis have looked at language and culture at multiple dimensions and scales, discovering novel trends in the growth and transformation of research and quantitative results on the history and evolution of science.[2] By engaging with computational tools like text analysis, historians are better able to develop, guide, and take ownership of such tools and approaches. Moreover, historians have a role and a stake in computational and digital methods, as today's information and data will become tomorrow's artifacts and sources of historical analysis.[3]

Text analysis approaches are specifically important to the study of history because understanding how language and words are used provides insight into social context during historical moments, how groups communicate, and the mutual influence of language and culture. Previous studies

Kenneth D. Aiello is a Postdoctoral Research Scholar with the Global Biosocial Complexity Initiative at Arizona State University. His research focuses on social science, history, language, and knowledge. Global Biosocial Complexity Initiative at ASU, Engineering Center—A Building (ECA), 1031 South Palm Walk, Tempe, Arizona 85281-2701, USA; kaiello@asu.edu.
Michael Simeone is an Assistant Research Professor in Biosocial Complexity at Arizona State University. He serves as the director of Data Science and Analytics for ASU Libraries. His research includes multidisciplinary data science, digital culture and technology, and data visualization. 1012 North 85th Place, Scottsdale, Arizona 85257m USA; michael.simeone@asu.edu.

[1] Jürgen Renn and Manfred D. Laubichler, "Extended Evolution and the History of Knowledge," in *Integrated History and Philosophy of Science: Problems, Perspectives, and Case Studies*, ed. Friedrich Stadler (Dordrecht: Springer, 2017), pp. 109–125.

[2] Kenneth D. Aiello, "Systematic Analysis of the Factors Contributing to the Variation and Change of the Microbiome" (Ph.D. diss., Arizona State Univ., 2018).

[3] Jane Maienschein, "Why Collaborate?" *Journal of the History of Biology*, 1993, *26*:167–183; and Bruno J. Strasser, "Data-Driven Sciences: From Wonder Cabinets to Electronic Databases," in "Data-Driven Research in the Biological and Biomedical Sciences," special issue, *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 2012, *43*:85–87, https://doi.org/10.1016/j.shpsc.2011.10.009.

Isis, volume 110, number 3. 0021-1753/2019/0110-0007$10.00.

have shown how social changes had linguistic consequences, looking at which words were used by different social and cultural groups. Further, language studies have provided evidence of changing group dynamics over time, showcasing changes in group solidarity and how groups create and reinforce social categories.[4] Across many scientific domains, there is a large body of empirical evidence, discovered through the study of words, discourse, and language, linking sociocultural phenomena to historical change. When examining twentieth-century histories of science and biomedicine, which partially consist of a published record of hundreds of thousands (millions) of documents, text analysis can be especially useful for understanding key dynamics at a broad scale.[5]

The contemporary turn toward "big data"—data analytics, machine learning, and large-scale datasets—has generated an overabundance of new data sources and resources, offering both challenges and novel avenues of investigation for historians and for science in general.[6] But when it comes to the study of large-scale text archives, historians are essential experts because of the deficiencies in the ways that machines read documents. Specifically, the context and semantics of terms and the significance of broader patterns in documents cannot be resolved computationally. Historians and historical approaches are critical as large-scale text archives continue to present an opportunity for new research in history.[7] As of now, there are millions of books accessible within the Google Books databases, millions of citations and abstracts in the National Library of Medicine, and an ever-growing collection of open-access documents. These collections and others offer historians the opportunity to analyze and document history with a depth and breadth never seen before.[8]

We see text analysis not as a substitute for traditional historical inquiry but as a complementary approach for understanding history across multiple scales and dimensions. Our contention is that the most challenging part of text analysis is triangulating results: merging accurate interpretations from historiographical methods with domain expertise. But in order to do this, a better understanding of text analysis and some of its fundamental mechanics is necessary. This essay describes key techniques and considerations for mixed-methods research in the history of science and medicine that uses documents as part of a broader practice of gaining historical understanding. We outline some approaches to working with data, analyzing texts, and interpreting results, followed by a representative example that studies the emergence of recent science pertaining to the human microbiome.

Data Collection and Cleaning

It is crucial to emphasize how impactful data collection, cleaning, and curation is during any text analysis. Errors in textual data or associated metadata can lead to inaccurate results; they can occur both in small analyses of single texts and in analyses of huge corpora of thousands of texts. Errors in textual data analysis may arise during data collection when transferring the data from one source to another, when downloading data from an electronic database or repository, or

[4] Ronald Wardhaugh and Janet M. Fuller, *An Introduction to Sociolinguistics*, 7th ed. (Chichester, West Sussex: Wiley-Blackwell, 2014), p. 10; William Labov, *The Social Stratification of English in New York City*, 2nd ed. (Cambridge: Cambridge Univ. Press, 1966), p. 298; and Lesley Milroy, *Language and Social Networks* (New York: Wiley, 1987), p. 72.

[5] danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication, and Society*, 2012, *15*:662–679, https://doi.org/10.1080/1369118X.2012.678878.

[6] Pascal Hitzler and Krzysztof Janowicz, "Linked Data, Big Data, and the Fourth Paradigm," *Semantic Web*, 2013, *4*:233–235.

[7] boyd and Crawford, "Critical Questions for Big Data" (cit. n. 5).

[8] Erez Aiden and Jean-Baptiste Michel, *Uncharted: Big Data as a Lens on Human Culture*, 1st ed. (New York: Riverhead, 2013), pp. 14, 207; and Manfred D. Laubichler, Jane Maienschein, and Jürgen Renn, "Computational Perspectives in the History of Science: To the Memory of Peter Damerow," *Isis*, 2013, *104*:119–130.

when merging data from different sources. Common errors in data collection may result in missing data fields, inaccurate data, or improperly formatted data. In the best-case scenarios, the error(s) can be identified by reading the text or through the analysis of text or metadata using descriptive statistics.

In many cases common errors in the data can be easily fixed by means of comparison between the source text and the new location. Often, and especially when dealing with large corpora, there can be hundreds or even thousands of errors stemming from the transfer of data from one source to another (or from Windows OS to Mac OS), complications caused by italicized words, hyphenated words, words joined incorrectly because space has been removed, pictures or images, and non-English words. Therefore, becoming familiar with the text by reading samples is a crucial first step. Systematic data collection, too, can help reduce textual errors. However, this essay is not intended as a guide to advanced text analysis, and there are multiple books and courses on data mining, collection, cleaning, and curation. We will not go into detail on those methods here but instead will point the reader to the Quartz guide to bad data.[9] The reality of text analysis is that most textual data is messy and hard to control. Cleaning and curation of the data is often the most important part of any study or experiment in the analysis of a text or collections of texts.

Applications of Text Analysis

Text analysis can look at words, word combinations (two words, three words, etc.), phrases, discourses, or entire documents for trends that are directly related to meaning, semantics, or intent. Some of the appeal of text analysis is the flexibility of the approach, as shown by the range of earlier analyses using letters, words, multiword units, attributes of texts (dates, authors, identifiers), discourses, concepts, knowledge, and so on. In general, text analyses create or use a model of texts and language that highlights patterns and changes in the usage and meaning of words and phrases or semantic content.[10] By identifying and analyzing changes in semantic content, historians can ask questions related to words, concepts, language, and knowledge. Previous studies using text analysis have provided insight into the behavior and specific actions taken by individuals, social groups, national economies, and larger global socioeconomic structures.[11]

Prior to using text analysis as evidence in a historical argument, we suggest that the reader refer to William Labov's *Social Stratification of English in New York City*, Ronald Wardhaugh and Janet Fuller's *An Introduction to Sociolinguistics*, Roberto Franzosi's *From Words to Numbers*, Klaus Krippendorff's *Content Analysis: An Introduction to Its Methodology*, and Taylor Arnold and Lauren Tilton's *Humanities Data in R* as resources for understanding the range of approaches and effective uses of text analysis.[12] These texts helpfully span theory and mechanisms for using language as data. While we encourage readers to experiment with text analysis, as with any scientific experiment the rationale behind the decision to use this approach should be clear. To put it simply: the fact that you can run a text analysis does not necessarily mean that you should. In

---

[9] https://github.com/Quartz/bad-data-guide.

[10] Paul Baker, *Using Corpora in Discourse Analysis* (London: Black, 2006), p. 146; Kevin W. Boyack, Richard Klavans, and Katy Börner, "Mapping the Backbone of Science," *Scientometrics*, 2005, *64*:351–374, https://doi.org/10.1007/s11192-005-0255-6; and Chaomei Chen, "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature," *Journal of the American Society for Information Science and Technology*, 2006, 57:359–377, https://doi.org/10.1002/asi.20317.

[11] Boyack *et al.*, "Mapping the Backbone of Science."

[12] Labov, *Social Stratification of English in New York City* (cit. n 4); Wardhaugh and Fuller, *Introduction to Sociolinguistics* (cit. n. 4); Roberto Franzosi, *From Words to Numbers: Narrative, Data, and Social Science*, 1st ed. (Cambridge: Cambridge Univ. Press, 2004); Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 4th ed. (Los Angeles: SAGE, 2018); and Taylor Arnold and Lauren Tilton, *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text* (Dordrecht: Springer, 2015).

general, if you are analyzing a single document (not a book), are not interested in comparisons based on language use, or require primary data as the unit of analysis, other approaches might be more suitable. However, if you have textual data from an archive or repository and want to gain insight without reading every document, then text analysis makes sense.

Case Study: The Microbiome Corpus

To show the power of computational text analysis, the following sections will focus on methods and results from an analysis of scientific articles on the microbiome. We investigated the use of important themes, biomedical concepts, and words in a large-scale multidimensional corpus. Using a combination of computer-assisted and manual methods, articles with the word "microbiome" in the text were downloaded as PDFs from Web of Science, JSTOR, PubMed, and PubMed Central. After removing duplicates, cleaning, and manually curating the corpus, 27,977 publications were collected (see Figure 1). For this study, articles with "microbiome" in the text from the years 2001 to 2010 were used.

Measuring Change in Language

The rest of this essay highlights the specific computational text analysis approaches of frequency analysis, concordance, and keywords as they apply to measuring conceptual changes within the discourse of microbiome research. These techniques are seen broadly across work in linguistics

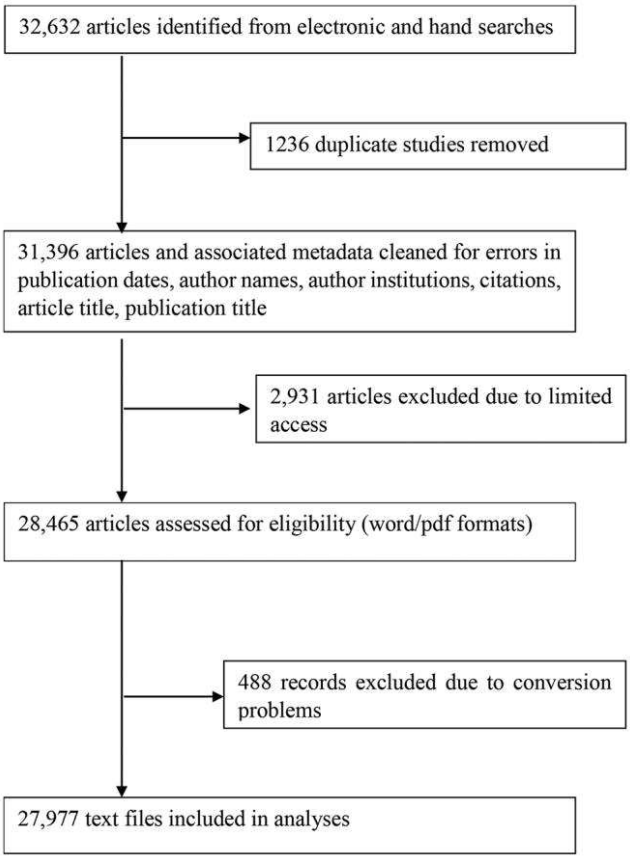

Figure 1. Result of article collection for the MB Corpus.

and computational social science, although the terminology sometimes varies (e.g., "frequency analysis" can appear as "wordlists" or "frequency profiles," and "keywords" can appear as "keyword in context").

We focus on these methods instead of on machine learning because historians are best positioned to provide insight into structured, supervised, hypothesis-driven results from computational text analysis. Generally, both approaches aim to understand language, meaning, knowledge, and the relationship between society and language, and detailed discussion of the differences are beyond the scope of this text. Briefly, we have chosen not to employ machine learning approaches (unsupervised learning), as these methods generally are used in cases where the outcome or response is unknown. These approaches are particularly useful when one is seeking to understand the relationship between variables or between observations, as in the clustering together of groups of things that do not have any predefined categories. A machine learning approach could be used, for example, to find unknown groups of words that cluster together or to find attributes of texts that are not visible through basic statistical analysis. Yet while machine learning approaches are useful in that they sometimes yield unpredictable or unforeseen results, the goal of this study is to combine qualitative, quantitative, and human insight to understand features of the texts. This account is a limited introduction to the differences between and the rationale for these approaches; interested readers should see Andreas Holzinger and Igor Jurisica's "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions" for a detailed discussion of the uses of machine learning in computational text analysis of biomedical articles.[13]

Our opinion is that the methods used here detail a hybrid approach that combines quantitative results with qualitative evidence, domain expertise, and close reading for validation. Historians play an important role in this approach, providing domain expertise, knowledge, and awareness of the historical context of language use. To illustrate, we will study knowledge related to the microbiome concept as the focus of this study. It is an apt case, as there is confusion about the historical evolution of the meaning of "microbiome," including the possibility that the concept encompasses multiple microbiomes—such as a core microbiome, a human microbiome, and an ecological microbiome—as opposed to one consensus interpretation. Some have traced the ongoing historical debate over the origin and source of "microbiome" to the question of whether Antoni van Leeuwenhoek or Joshua Lederberg coined the term. Beyond this discussion, microbiome language and knowledge make for an ideal text analysis case study because different narratives, interpretations, and meanings have been attributed to the microbiome vocabulary and ontology.[14] While there is ample evidence to allow for analysis of the microbiome in texts, there is still no consensus or agreed upon interpretation of the microbiome or other core concepts used with the microbiome. Along the same lines, other studies using similar approaches (supervised

---

[13] Andreas Holzinger and Igor Jurisica, "Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges*, ed. Holzinger and Jurisica (Berlin: Springer, 2014), pp. 1–18, https://doi.org/10.1007/978-3-662-43968-5_1. Regarding the benefits of machine learning see Gareth James *et al.*, *An Introduction to Statistical Learning: With Applications in R* (New York: Springer, 2013), p. 26. On the complexity of language see Clay Beckner *et al.*, "Language Is a Complex Adaptive System: Position Paper," *Language Learning*, 2009, 59(suppl.):1–26, https://doi.org/10.1111/j.1467-9922.2009.00533.x; and William A. Kretzschmar, *Language and Complex Systems*, 1st ed. (New York: Cambridge Univ. Press, 2015).

[14] John Huss, "Methodology and Ontology in Microbiome Research," *Biological Theory*, 2014, 9:392–400, https://doi.org/10.1007/s13752-014-0187-6; Susan L. Prescott, "History of Medicine: Origin of the Term Microbiome and Why It Matters," *Human Microbiome Journal*, 2017, 4:24–25, https://doi.org/10.1016/j.humic.2017.05.004; and Ashley Shade and Jo Handelsman, "Beyond the Venn Diagram: The Hunt for a Core Microbiome," *Environmental Microbiology*, 2012, 14:4–12, https://doi.org/10.1111/j.1462-2920.2011.02585.x.

learning) have provided insight into the trajectories of other biomedical concepts, characteristics of sociotechnical innovations, and the structure of scholarly networks.[15]

TEXT ANALYSIS OF THE MB CORPUS

The frequency of words and multiword units within texts is the basis for most computational text analysis. Word counts are used in everything from artificial intelligence text analysis and Google's unsupervised machine learning algorithms to manual human-conducted coding and text analysis.[16]

Analysis of the frequency of words provides information on the language used within a corpus and can serve as a basis for comparing individuals, social groups, institutions, and discourses. Understanding which specific words are used when and how often has provided insight into language and knowledge changes across social and temporal dimensions.[17] Using a corpus of 4.2 million words and a combination of text analysis approaches, historical differences in the usage of specific work-related words were found between males and females and between those under thirty-five versus those over thirty-five, as well as within categories based on socioeconomic class. Other studies, using the same corpus, have verified these results and created new experiments to test other novel language usage differences between males and females.[18] While not always the case, text analyses of word frequency can be reproducible by others and lead to interesting *post hoc* questions and analyses.

When analyzing for frequency differences it is helpful to divide a general corpus according to predefined categories of interest, like social characteristics or time slices.[19] Other analyses of word frequencies have confirmed differences in the usage of words related to differences in prestige, power, region, income, and social network relations. Previous studies have also shown how the discourses of research communities change; examples include changes in the focal questions asked, the units of analysis, and the core concepts.[20]

Differences in the focal point of microbiome discourse were discovered by conducting a frequency analysis on year slices of the MB Corpus (i.e., 2001, 2002, . . . , 2010). Each year slice of the MB Corpus is representative of the microbiome research community discourse for that year. The top ten words in the MB Corpus from 2001 to 2010 indicated that the highest-frequency

---

[15] Tudor M. Baetu, "Genes after the Human Genome Project," in "Data-Driven Research in the Biological and Biomedical Sciences," special issue, *Stud. Hist. Phil. Biol. Biomed. Sci.*, 2012, *43*:191–201, https://doi.org/10.1016/j.shpsc.2011.10.022; Mariana C. Arcaya, Alyssa L. Arcaya, and S. V. Subramanian, "Inequalities in Health: Definitions, Concepts, and Theories," *Global Health Action*, 2015, *8*, https://doi.org/10.3402/gha.v8.27106; and J. K. Chambers, Peter Trudgill, and Natalie Schilling, eds., *The Handbook of Language Variation and Change*, 2nd ed. (Malden, Mass.: Wiley-Blackwell, 2013).

[16] Aiden and Michel, *Uncharted* (cit. n. 8), p. 17.

[17] Michael Stubbs, *Discourse Analysis: The Sociolinguistic Analysis of Natural Language* (Chicago: Univ. Chicago Press, 1983); and Douglas Biber, "Stance in Spoken and Written University Registers," *Journal of English for Academic Purposes*, 2006, 5:7–116, https://doi.org/10.1016/j.jeap.2006.05.001.

[18] Paul Rayson and Roger Garside, "Comparing Corpora Using Frequency Profiling," in *Proceedings of the Workshop on Comparing Corpora*, Vol. 9 (Stroudsburg, Pa.: Association for Computational Linguistics, 2000), pp. 1–6; and Hans-Jörg Schmid, "Do Women and Men Really Live in Different Cultures? Evidence from the BNC," in *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, ed. Andrew Wilson, Rayson, and Tony McEnery (Frankfurt: Lang, 2003), pp. 185–222.

[19] Dell H. Hymes and John Joseph Gumperz, *Directions in Sociolinguistics: The Ethnography of Communication* (New York: Holt, Rinehart & Winston, 1972).

[20] For studies that note differences in word usage in the contexts of social factors see William Labov, *Principles of Linguistic Change*, Vol. 2: *Social Factors* (Malden, Mass.: Blackwell, 2001), p. 114; Chambers *et al.*, eds., *Handbook of Language Variation and Change* (cit. n. 15), pp. 537, 365; and James Milroy and Lesley Milroy, "Linguistic Change, Social Network, and Speaker Innovation," *Journal of Linguistics*, 1985, *21*:339–384, https://doi.org/10.1017/S0022226700010306. On changes in the discourse of research communities see Labov, *Principles of Linguistic Change*, Vol. 1: *Internal Features* (Oxford: Blackwell, 1994); and L. Milroy, "Language Ideologies and Linguistic Change," in *Sociolinguistic Variation: Critical Reflections*, ed. Ronald K. S. Macaulay and Carmen Fought (Oxford: Oxford Univ. Press, 2004), pp. 161–177.

Table 1. Most Frequent Words in MB Corpus from 2001 to 2010

N	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	the	the	of	the	the	the	the	the	the	the
2	of	of	the	of	of	of	of	of	of	of
3	and	and	a	in	and	and	and	and	and	and
4	a	to	and	and	in	in	in	in	in	in
5	to	in	to	a	a	a	a	a	a	a
6	in	a	in	gut	to	to	to	to	to	to
7	j	that	b	to	mice	b	for	for	for	for
8	that	for	j	bacteria	that	that	that	that	with	with
9	for	is	are	insect	with	for	is	with	that	that
10	is	are	m	that	from	with	with	is	by	by

words were function words, which is consistent with other studies on language (see Table 1).[21] Generally, function words like "the," "of," "a," "are," and so forth are considered unimportant and provide little insight, whereas content words like nouns, verbs, adjectives, and adverbs are usually found to be of interest. A stop list—a list of words not critical to the experiment—was used to remove the function words and highlight content words. Creating a stop list is an iterative process. For each analysis a separate stop list is created. The final list is determined by the researchers, based on when they believe that enough of the function words have been removed.

Most stopword lists or stop lists consist of words like "a," "as," "an," "of," "is," "but," "of," and "the"; they also incorporate words that, given the corpora or the experiment, should be ignored. Even before the stop list was applied in our study there are hints of differences in the discourse across time. Words like "bacteria," "gut," and "insect" were among the most frequent words for 2004 but not for any other year. Similarly, "mice" was a top word for 2005 but not anywhere else.

Frequency Analysis after Stopwords Were Removed

Comparing the frequency of the top ten words from 2001 to 2010 for each year in the corpus after stopwords were removed shows how the frequencies of the top ten words changed over time and suggests shifts in the discourse of the microbiome research community (see Table 2). Ten years is a small window; but given that "microbiome" didn't occur in multiple articles until 2001, this narrow window highlights noticeable differences. As expected, words like "bacteria," "microbial," "microbes," and "microbiota" remained high-frequency words over time. Our results confirm the microbiome as a concept influenced by biology and microbiology, as previous studies, historians, and experts suggested.[22]

Conversely, there is no consensus scholarly argument on the focal point of microbiome research, with some arguing that the microbiome is an ecological space—that is, a "micro"-biome—and others arguing that the microbiome concept is specific to humans.[23] The results

[21] Baker, *Using Corpora in Discourse Analysis* (cit. n. 10), p. 130; and William A. Kretzschmar, Jr., *The Linguistics of Speech* (Cambridge: Cambridge Univ. Press, 2009), https://doi.org/10.1017/CBO9780511576782, p. 143.

[22] Julian R. Marchesi and Jacques Ravel, "The Vocabulary of Microbiome Research: A Proposal," *Microbiome*, 2015, 3(31), https://doi.org/10.1186/s40168-015-0094-5.

[23] Martin Blaser *et al.*, "The Microbiome Explored: Recent Insights and Future Challenges," *Nature Reviews: Microbiology*, 2013, *11*:213–217, https://doi.org/10.1038/nrmicro2973; and Beth Mole, "Microbiome Research Goes without a Home," *Nature News*, 2013, *500*(7460):16, https://doi.org/10.1038/500016a.

Table 2. Most Frequent Words in MB Corpus from 2001 to 2010 after Stopwords Removed

N	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	host	microbial	microbiota	gut	mice	genes	genes	human	human	human
2	bacteria	human	intestinal	bacteria	cells	mice	gut	microbial	microbial	mice
3	bacterial	microbes	members	insect	microbiota	expression	genome	bacterial	gut	cells
4	use	disease	type	microbiota	intestinal	host	gene	gut	bacterial	bacterial
5	science	use	human	microbial	gut	gut	human	bacteria	analysis	analysis
6	microbial	host	genome	insects	villus	microbial	analysis	sequences	bacteria	microbial
7	infection	bacteria	microbial	species	animals	cells	species	analysis	microbiota	using
8	disease	new	two	bacterial	genes	gene	microbial	gene	using	gene
9	immune	one	homologs	plant	bacteria	human	sequences	host	gene	data
10	population	expression	bacteroides	bacteria	after	protein	data	using	genes	cell

from the frequency analysis also point to this ambiguity, with the text analysis showing high in-text frequencies for a range of different animals like "insects" and "mice" earlier in the corpus and then later a drop-off in these concepts with a shift to "human" in MB 2008 and MB 2009. That the words "cells" and "cell" were among the top words in MB 2010 provides additional evidence of a transition in microbiome research and possible points of divergence, as they do not occur as top words in any other year. Another noticeable difference is that "genes" was the most used word in 2006 and 2007 but did not appear nearly as often pre-2006 and post-2007. These increases and decreases might be attributed to changes in what the units of analysis were ("animals," "human," "bacteria," "bacteroides," "genes," "insects," "mice"), the scope of the analyses ("human," "gut," "species," "sequences"), and the research focus of the microbiome community ("genes," "genome," "gut," "human," "intestinal"). Our results hint at a possible spectrum of interpretations for the microbiome and variation in the pivotal concepts related to it. While these *post hoc* questions are beyond the scope of this analysis, they provide interesting avenues for future research.

To gain a more in-depth understanding of one of the concepts used most frequently in the MB Corpus, we investigated how "human" was being used. Specifically, was the word being used as an adjective, as in "human liver" or "human body," or was it being used as a noun, as in "being human"? To answer these questions, we turned to the textual context and performed a manual reading of the usage of "human" in the texts.

Concordance of "Human"

A concordance analysis of the texts was performed to get a sense of how "human" was being used in the corpus and to provide context for the frequency-based results. The concordance highlighted word clusters and starting points for manual reading that offered more insight into the actual usage of "human." The concordance analysis yielded a table with all occurrences of "human" in the text along with the sentences the word was embedded in. For brevity, only 12 lines of the concordance analysis for "human" are shown; 20,920 lines were analyzed using concordance (see Figure 2). Each line in the concordance showed the occurrence of "human" and additional context words in the sentence. For the concordance analysis we used Wordsmith Tools.[24]

The concordance analysis also provided frequencies for three-word clusters containing "human" in the MB Corpus from 2001 to 2010. The clusters showed that "human" was used most often as an adjective to describe things related to humans: "of the human" (2,223 occurrences) and "in the human" (1,525 occurrences). The clusters also emphasize a connection between "human" and parts of the gastrointestinal system of the human body: "the human gut" (1,091 occurrences), "human distal gut" (356), and "the human intestine" (384). Other clusters show the scale of analysis, with frequent word clusters referring to the "human gut microbiota" (538) and "human intestinal microbiota" (169). Our results show that the word "human" was used as an adjective, generally referring to parts of the human gastrointestinal system, with an emphasis on microbes or microbiota. A close reading provided additional evidence that supported these interpretations, as seen in the following examples:

> To explore whether and how these principles apply to the gut microbiota and its microbiome, we have determined the complete genome sequences of two Bacteroidetes with highly divergent 16S rRNA phylotypes that are prominently represented in the distal

[24] Mike Scott, "Wordsmith Tools" (Stroud: Lexical Analysis Software, 2018), https://lexically.net/publications/citing_wordsmith.htm.

in a gutless marine worm [9], phosphorus-removing activated sludge [10], the ***human*** [11] and termite [12] gut and marine microbial [13,14] and viral [15] sample
deeply affects ***human*** physiology. To identify the genomic features common to all ***human*** gut microbiomes as well as those variable among them, we performed a large
, undernourishment based on PEM greatly increases susceptibility to major ***human*** infectious diseases in low-income countries, particularly in children [2
establishment of criteria for disease causation and further characterization of the ***human*** microbiome during states of health. These challenges and the goal of under
, no metabolic phenotype (metabotype) has ever been ascribed to an individual ***human*** dietary preference group. Variations in the basal metabotype have, however
are not precisely defined for microbes, we use them here to frame a view of ***human*** gut microbial ecology. When the sequences(n = 495 greater than 900 base
have integrated with insects. The consortium of microbes inhabiting the ***human*** gut (1014) is estimated to outnumber the somatic and germ cells of the body
present, due in part to the low reproductive intake ability, low nutrition and ***human*** activities that have destroyed their efficiency, natural habitat [5, 11, 1
capabilities among Lactobacillus and dobacterium strains isolated from the ***human*** intestinal ecosystem as the initial step for further investigation of the
strategies to change the representation and/or properties of M. smithii in the ***human*** gut microbiota. archaeal-bacterial mutualism comparative microbial genomic
characterization of these libraries was in good agreement with previous studies of ***human*** feces using molecular approaches (18). A two-step microplate screening met
microbial genomics functional genomics and metabolomics gnotobiotic mice ***human*** gut microbiota to the gut ecosystem. First, comparative metagenomic

Figure 2. Concordance of "human" in the MB Corpus.

> gut of healthy *humans*—Bacteroides vulgatus and Bacteroides distasonis (now also known as Parabacteroides distasonis [7]).[25]

> Triggered by the growing number of 16S ribosomal RNA (rRNA)-based approaches, insights in the evolutionary diversity of the *human* adult gut flora has changed drastically in recent years.[26]

> The *human* large bowel microbiota, consisting of ~10^{14} bacteria, contributes to *human* nutrition and health[1]. It is estimated that ~10% of caloric intake is derived from plant polysaccharides that are deconstructed exclusively by enzymes produced by the microbiota.[27]

Previous studies have combined frequency, clustering, and concordance analyses on the most frequently occurring word(s) to determine differences in word usage related to changes in discourse, provide details on how the meaning of words changed over time, and highlight semantic differences based on social or cultural factors. A further consideration is that these differences may highlight fluctuations in usage or be part of a larger trend in the overall variation of the discourse. Other studies using these approaches have shown how to differentiate between variation and change and have supported similar conclusions based on text analysis.[28]

To recap, then: the comparison of both the raw and the stopwords removed frequency lists of the MB Corpus provided insight into shifts in the microbiome discourse and a method to start investigating language variation and change. What is still unknown is whether these shifts were significant or simply reflective of variation, as many of the same words occurred in the top ten for multiple years, including "bacterial," "human," "gut," and "microbial." The next approach, keyword analysis, emphasizes the differences in word frequencies across corpora and considers the size of each corpus to find statistically significant words on the basis of differences in word frequency. This provides a way to measure saliency while considering word frequency.

Keyword Analysis: Lexical Saliency

Keyword analysis uses frequency lists from corpora and highlights the lexical focus, or lexical saliency, of texts. By comparing the relative frequency of words between different corpora, keyword analysis reveals which words occur more often in one or the other.[29] A "keyword" (or

---

[25] Ming Xu, Hua Cai, and Sai Liang, "Big Data and Industrial Ecology," *Journal of Industrial Ecology*, 2015, *19*:205–210, https://doi.org/10.1111/jiec.12241.

[26] Geert Huys, Tom Vanhoutte, and Peter Vandamme, "Application of Sequence-Dependent Electrophoresis Fingerprinting in Exploring Biodiversity and Population Dynamics of Human Intestinal Microbiota: What Can Be Revealed?" *Interdisciplinary Perspectives on Infectious Diseases*, 2008, https://doi.org/10.1155/2008/597603.

[27] Yanping Zhu *et al.*, "Mechanistic Insights into a Ca2+-Dependent Family of α-Mannosidases in a Human Gut Symbiont," *Nature Chemical Biology*, 2010, *6*:125–132, https://doi.org/10.1038/nchembio.278.

[28] On how to use word frequency for analysis see Allison Burkette and William A. Kretzschmar, Jr., *Exploring Linguistic Science: Language Use, Complexity, and Interaction*, 1st ed. (New York: Cambridge Univ. Press, 2018), p. 147; and Jacqueline Marie Hettel, "Harnessing the Power of Context: A Corpus-Based Analysis of Variation in the Language of the Regulated Nuclear Industry" (Ph.D. diss., Univ. Georgia, 2013). For work differentiating between variation and change see William Labov, *Sociolinguistic Patterns* (Philadelphia: Univ. Pennsylvania Press, 1972); and Paul Baker, *Sociolinguistics and Corpus Linguistics*, 1st ed. (Edinburgh: Edinburgh Univ. Press, 2010), p. 69.

[29] Baker, *Sociolinguistics and Corpus Linguistics*; and Nicole Kronberger and Wolfgang Wagner, "Keywords in Context: Statistical Analysis of Text Features," 2000, https://s3.amazonaws.com/academia.edu.documents/46889788/T4.5.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1540501900&Signature=iUAite6rQEqX9KGbk9MgVs04Xzw%3D&response-content-disposition=inline%3B%20filename%3DKeywords_in_context_Statistical_analysis.pdf.

"keywords") is a statistically significant word (or words) found by means of a comparison between two corpora. Essentially, the frequency or number of times a word occurs in one corpus is compared to the frequency of the same word in a different corpus. In most instances, there is a reference corpus and a corpus of interest. Keyword analyses have been used to show differences in word usage between individuals and to pick out important phrases used by social groups; they are another approach that builds on the usage of frequency lists and concordance and have been used to gain insight into language variation and change.[30] Within biomedicine, keywords have been used to analyze opinions of practice-based research and comments from patients, to characterize how knowledge diffuses in clinical settings, and to assess the evolution of research fields.[31]

Different statistical tests are used to determine keywords, including log-likelihood, Dice similarity coefficient, mutual information score, and Chi-square. A discussion of differences between the statistical tests and relevant nuances is beyond the scope of this essay but can be found in Michael Oakes's *Statistics for Corpus Linguistics*.[32] Generally, the result from a keyword analysis is a list of statistically significant words, or keyword list, that shows both positive and negative keywords. Positive keywords are words that are more likely to appear in one corpus compared to the other, and negative keywords are words that are unusually infrequent in one corpus compared to the other.

Creating a keyword list of interest requires both a reference corpus and a comparison corpus or corpus of interest. We can compare a keyword analysis to a standard scientific experiment with a control and an experimental group: the reference corpus is the control group and the corpus of interest is the experimental group. To triangulate statistically significant changes in focal points of the microbiome discourse, we compared the years in the MB Corpus where "gene" was the most frequently occurring word (2006–2007) with the years where "human" was the most frequently occurring word (2008–2010). In this instance, then, the reference corpus was created using the MB Corpus texts from 2006 and 2007, while the corpus of interest was created from the MB Corpus texts from 2008 to 2010. On the basis of findings from earlier studies, we used log-likelihood to evaluate statistically significant words, with a p-value threshold of 0.00001 and a minimum occurrence of 10 for each word.[33]

The results exhibited salient words and larger trends of lexical differences between the two time slices. The lexical differences were reviewed and aggregated into categories based on context following practices used in previous studies.[34] The categories emphasized patient/population, approaches, symptoms, unit of analysis, and location/environment (see Table 3). Close reading helped confirm the actual usage of a keyword in the text and assisted in determining

---

[30] Costas Gabrielatos and Paul Baker, "Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the U.K. Press, 1996–2005," *Journal of English Linguistics*, 2008, *36*:5–38; and Baker, "Acceptable Bias? Using Corpus Linguistics Methods with Critical Discourse Analysis," *Critical Discourse Studies*, 2012, *9*:247–256.

[31] Selene J. Huntley *et al.*, "Analysing the Opinions of U.K. Veterinarians on Practice-Based Research Using Corpus Linguistic and Mathematical Methods," *Preventive Veterinary Medicine*, 2018, *150*:60–69, https://doi.org/10.1016/j.prevetmed.2017.11.020; Inocencio Daniel Maramba *et al.*, "Web-Based Textual Analysis of Free-Text Patient Experience Comments from a Survey in Primary Care," *JMIR Medical Informatics*, 2015, *3*(2):e20, https://doi.org/10.2196/medinform.3783; and Svenja Adolphs *et al.*, "Applying Corpus Linguistics in a Health Care Context," *Journal of Applied Linguistics*, 2004, *1*:9–28, https://doi.org/10.1558/japl.1.1.9.55871.

[32] Michael Oakes, *Statistics for Corpus Linguistics*, 1st ed. (Edinburgh: Edinburgh Univ. Press, 1998).

[33] Paul Baker and Jesse Egbert, eds., *Triangulating Methodological Approaches in Corpus Linguistic Research*, 1st ed. (New York: Routledge, 2016), p. 169.

[34] Clive Seale, Sue Ziebland, and Jonathan Charteris-Black, "Gender, Cancer Experience, and Internet Use: A Comparative Keyword Analysis of Interviews and Online Cancer Support Groups," *Social Science and Medicine*, 2006, *62*:2577–2590, https://doi.org/10.1016/j.socscimed.2005.11.016.

Table 3. Keyword Differences between MB 2006–2007 and MB 2008–2010

Keyword Category	MB 2006–2007	MB 2008–2010
Patient/Population	genes, objects, paralogs	women, subjects, infants, pathogens
Approaches	sequenced, genechip	randomized, identification, exposure
Symptoms	tuberculosis, atrophy, spoilage, streptomyces	infection, diarrhea, irritable, stress, plaque
Unit of Analysis	glycoside, bacteroidetes, pylori, polysaccharide	biofilms, virus, coral, dental, taxa bifidobacterium
Location/Environment	genome, gastric, cecal, colony, gut	vaginal, rumen, bowel, saliva, cystic

the categories for each keyword. To illustrate, the word "women" was categorized as a patient/population keyword in the 2008–2010 corpus when compared to the 2006–2007 corpus. The word "women" occurred 1,441 times in the 2008–2010 texts, compared to only 13 occurrences for 2006–2007, and was found in 138 texts (18 percent of the entire 2008–2010 corpus). These results confirm that the word "women" is used multiple times in different texts; we are not seeing a random pattern or an instance of one word being used at a high frequency in only a few texts. The concordance results further confirmed that the actual usage of the word "women" supported the patient/population conclusion:

> Differentiated adipocytes from overweight non-diabetic *women* were cultured with or without HCA-SX (0.5mg/mL) for 96 hours.[35]

> Healthy female volunteers aged 18 to 50 undergoing gynaecological exam were recruited from the Health Sciences Centre Department of Obstetrics & Gynaecology Colposcopy Clinic in Winnipeg, Manitoba. *Women* undergoing cervical biopsies, menstruating or pregnant were excluded from the study.[36]

Additional supporting evidence for these categorical interpretations came from frequency analyses and concordance.

For more information on characterizing differences in keywords refer to Dell Hymes and John Joseph Gumperz's *Directions in Sociolinguistics: The Ethnography of Communication* and Hymes's *Foundations in Sociolinguistics: An Ethnographic Approach*.[37] Combining the keyword results with the results from the frequency lists and concordance analyses shows changes in word usage that highlight significant differences in the focal points of the microbiome discourse.

Criticisms of keyword analysis emphasize there is no standardized cutoff point and that larger corpora generally produce more keywords than smaller corpora. Keywords results differ depending on the statistical test, the size of the corpora, and the reference corpus. Keywords, moreover, are not mutually exclusive, and a keyword in one corpus can be discovered to be a keyword in the

[35] Francis C. Lau *et al.*, "Nutrigenomic Analysis of Diet-Gene Interactions on Functional Supplements for Weight Management," *Current Genomics*, 2008, 9:239–251, https://doi.org/10.2174/138920208784533638.

[36] Rachel E. Horton *et al.*, "A Comparative Analysis of Gene Expression Patterns and Cell Phenotypes between Cervical and Peripheral Blood Mononuclear Cells," *PLOS One*, 2009, 4:e8293, https://doi.org/10.1371/journal.pone.0008293.

[37] Hymes and Gumperz, *Directions in Sociolinguistics* (cit. n. 19); and Dell Hymes, *Foundations in Sociolinguistics: An Ethnographic Approach* (New York: Routledge, 2013).

comparison corpus. To mitigate these criticisms, many studies combine the results of keyword analyses with close reading and concordance lists in order to provide validation and supporting evidence to help with the interpretations and results.

The results and experiments described in this essay were influenced by larger historical, social, and cultural narratives, but the results do support previous findings on the various and changing interpretations of the microbiome.[38] Additionally, the results described here are part of a larger research project that created the MB Corpus and employed similar approaches to reach novel conclusions and interpretations on the variation, change, and characteristics of the microbiome and the microbiome research community from biomedical texts.[39] It should be noted that the methods and analyses described in this essay are commonplace in the social sciences, digital humanities, and corpus linguistics; they can be used to help support other hypotheses and address new research questions or as the starting point of a text analysis project.

CONCLUSION

Frequency, concordance, and keyword analyses were used to triangulate textual frequency, history, and meaning of words used in the MB Corpus.[40] Insight into changes in the microbiome discourse were found using the MB Corpus and comparing words within year slices of it. A combination of approaches was used to identify variation and change in the microbiome concept, suggest likely interpretations of the usage and meaning of the word "human," characterize how keywords were used, and assist in identify linguistic patterns. Specifically, our results provide additional evidence supporting earlier claims that the microbiome was influenced by biology and microbiology; further, our results point to a possible spectrum of interpretations for the term "microbiome" and highlight the overall complexity of the microbiome vocabulary.

Triangulating complex phenomena is the common thread in text analysis studies. The lack of a formal methodology allows for the combination of qualitative and quantitative data during analysis and is well suited for historical inquiry. Many of the results achieved using the methods described here are quantitative; these methods were combined with in-depth reading, consultation with domain experts, and conversations with biomedical historians. As seen here, text analysis is especially suitable for mixed-methods work when pursuing questions in the history of science and medicine, which unfold in part across a vast collection of printed research materials. Still, without interpretations supported by experts, there is no frame of reference for these results. Historians helped to provide the contextual references, critical frameworks, and historical evidence to support the results from text data analyses. Previous studies combining historical expertise with the results from frequency lists and concordance analyses have characterized the discourse of scientific fields, predicted the location of scientific innovation, and charted the growth of scientific knowledge.[41]

---

[38] Eric Juengst and John Huss, "From Metagenomics to the Metagenome: Conceptual Change and the Rhetoric of Translational Genomic Research," *Genomics, Society, and Policy*, 2009, 5(3):1–19; Rosamond Rhodes, Nada Gligorov, and Abraham Paul Schwab, *The Human Microbiome: Ethical, Legal, and Social Concerns* (Oxford: Oxford Univ. Press, 2013); and Jonathan A. Eisen, "What Does the Term Microbiome Mean? And Where Did It Come from? A Bit of a Surprise . . . ," *Winnower*, 1 Jan. 2015, https://doi.org/10.15200/winn.142971.16196.

[39] Aiello, "Systematic Analysis of the Factors Contributing to the Variation and Change of the Microbiome" (cit. n. 2).

[40] Michael Stubbs, "Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies," *Functions of Language*, 1995, 2:23–55; and Baker and Egbert, eds., *Triangulating Methodological Approaches in Corpus Linguistic Research* (cit. n. 33).

[41] Ali Balaid *et al.*, "Knowledge Maps: A Systematic Literature Review and Directions for Future Research," *International Journal of Information Management*, 2016, 36:451–475, https://doi.org/10.1016/j.ijinfomgt.2016.02.005; Adolphs *et al.*, "Applying Corpus Linguistics in a Health Care Context" (cit. n. 31); and Paul Baker *et al.*, "A Useful Methodological Synergy? Combining

Conventional historical methods are especially important for interpreting the results of text analysis because of the way computers read documents. Simply stated, computers currently do a bad job. Text analyses of many stripes can deliver observations about patterns in documents, but these observations are about the text data alone. Data can never interpret itself, and human interpretation is required to provide the concepts, theories, and knowledge for every algorithm or analysis. Those who argue that data is free from theory ignore the fact that any captured data is shaped or interpreted by the technology collecting the information, the platforms distributing the data, and the data ontology used to organize it. Even if all the capture processes are part of an automated workflow untouched by human interaction, any algorithms used to process the data are the products of science and specific scientific approaches influenced by previous theories, frameworks, and paradigms.[42] Statisticians and computer and data scientists stress the importance of context, domain expertise, and the perils of interpreting patterns without contextual knowledge. But recognizing the importance of context is not enough. Historians can transform general notions of context into crucial parts of the living record, the human stories that make language mean what it means—in one document or a million.

APPENDIX. COMPUTATIONAL REPRODUCIBILITY

In the spirit of engaging historians and providing access to new approaches and resources, we would like to point to valuable resources for those interested in completing a similar analysis. For those familiar with programming languages we recommend using R or Python along with the following texts: *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, and *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras*.[43] For those interested in using software programs to analyze data we suggest Oxford Wordsmith Tools (WT) version 4, as it is free and provides excellent supporting documentation. WT is a suite of software programs for text analysis; it has been used for projects in biomedicine, linguistics, sociology, and computer science since 1996. The most recent version of WT (version 6) and the free version can both be found at https://www.lexically.net/wordsmith/. While WT is for Windows only, there is a large and diverse set of options for text analysis, as well as text-mining software programs, available for specific platforms, operating systems, and user preferences (paid vs. free). The purpose of this essay is not to review all the options, but those interested in computational text analysis should know that there is a range of available options.

Indeed, given that there are so many options it may be difficult to choose between learning to program or picking a text-mining software program. It is not necessary to learn how to program for text analysis, but programming is necessary to customize data workflows and get the most insight from textual data. To help with the decision-making process for software programs

---

Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the U.K. Press," *Discourse and Society*, 2008, *19*:273–306.

[42] Peter Gould, "Letting the Data Speak for Themselves*," *Annals of the Association of American Geographers*, 1981, *71*:166–176, https://doi.org/10.1111/j.1467-8306.1981.tb01346.x; Rob Kitchin, "Big Data, New Epistemologies, and Paradigm Shifts," *Big Data and Society*, 2014, *1*(1):2053951714528481, https://doi.org/10.1177/2053951714528481; and Sabina Leonelli, "Introduction: Making Sense of Data-Driven Research in the Biological and Biomedical Sciences," in "Data-Driven Research in the Biological and Biomedical Sciences," special issue, *Stud. Hist. Phil. Biol. Biomed. Sci.*, 2012, *43*:1–3, https://doi.org/10.1016/j.shpsc.2011.10.001.

[43] Natalia Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis* (Amsterdam: Benjamins, 2015); R. H. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, 1st ed. (Cambridge: Cambridge Univ. Press, 2008); and Bhargav Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras*, 1st ed. (Birmingham: Packt, 2018).

we suggest referring to https://www.kdnuggets.com/software/text.html, which provides a list of commercial, online, and free text analysis tools.

Finally, in an effort to make the research used for this article an avenue for discovery beyond this text, we have provided a selection of the MB Corpus—specifically, a hundred articles converted to .txt files from 2007 to 2010, with twenty-five documents for each year. These documents are a part of the larger MB Corpus, which was itself part of a systematic project to identify contextual factors influencing specific changes to the microbiome concept.[44] This corpus of documents, the MPI 100, is provided for readers as a resource that will allow them to experiment with computational text analysis using the same approaches: https://diging.atlassian.net/wiki/x/B4BaQQ.

[44] Aiello, "Systematic Analysis of the Factors Contributing to the Variation and Change of the Microbiome" (cit. n. 2).