Kestler, Thomas

Article — Published Version

Big Data and Ideational Institutionalism. Reconsidering the Possibilities and Limitations of Google Ngram in the Study of Ideas

Politische Vierteljahresschrift

Suggested Citation: Kestler, Thomas (2024) : Big Data and Ideational Institutionalism. Reconsidering the Possibilities and Limitations of Google Ngram in the Study of Ideas, Politische Vierteljahresschrift, ISSN 1862-2860, Springer Fachmedien Wiesbaden, Wiesbaden, Vol. 66, Iss. 2, pp. 407-420,
https://doi.org/10.1007/s11615-024-00569-4

This Version is available at:
https://hdl.handle.net/10419/323510

Politische Vierteljahresschrift (2025) 66:407–420
https://doi.org/10.1007/s11615-024-00569-4

PVS Politische Vierteljahresschrift
German Political Science Quarterly

RESEARCH NOTE

Big Data and Ideational Institutionalism. Reconsidering the Possibilities and Limitations of Google Ngram in the Study of Ideas

Thomas Kestler

Received: 3 August 2023 / Revised: 5 August 2024 / Accepted: 7 August 2024 / Published online: 16 September 2024

Abstract Ideas and ideational change are difficult to grasp. When exactly did the idea of sustainability emerge? How has the meaning of democracy changed over time? Did ideational change precede or follow institutional change? These and many similar questions are highly relevant to the emerging field of ideational institutionalism, but they are difficult to answer conclusively without reliable data. *Google Books Ngram Viewer*, a tool introduced in 2011, opens up new possibilities in this regard by allowing the analysis of millions of print sources over time. However, after initial euphoria, its use has waned somewhat due to growing doubts about its reliability and the validity of the results it produces. Some of these problems have been addressed, but the debate about the potential and limitations of Google Books Ngram Viewer is ongoing. This research note will contribute to this debate by proposing a systematization for the use of search terms based on the concept of belief systems and by presenting the results of a reliability test for the German language corpus.

Keywords Ideational institutionalism · Corpus linguistics · Belief systems · Google Ngram · Big data

✉ Thomas Kestler
Institut für Politikwissenschaft und Soziologie, Universität Würzburg, Wittelsbacherplatz 1, 97074 Würzburg, Germany
E-Mail: thomas.kestler@uni-wuerzburg.de

Die Rolle von Big Data im ideenbasierten Institutionalismus. Überlegungen zu den Anwendungsmöglichkeiten von Google Ngram für die Untersuchung von Ideen

Zusammenfassung Ideen und ideeller Wandel sind nur schwer zu erfassen. Wann genau ist die Idee der Nachhaltigkeit entstanden? Wie hat sich die Bedeutung des Demokratiebegriffs im Laufe der Zeit verändert? Ging der ideelle Wandel dem institutionellen Wandel voraus oder folgte er ihm? Diese und viele ähnliche Fragen sind im aufstrebenden Bereich des ideellen Institutionalismus von großer Bedeutung, lassen sich aber ohne verlässliche Daten nur schwer schlüssig beantworten. Das 2011 eingeführte Tool *Google Books Ngram Viewer* eröffnet in dieser Hinsicht neue Möglichkeiten, da es die Analyse von Millionen von gedruckten Quellen im Zeitverlauf ermöglicht. Nach anfänglicher Euphorie ist die Nutzung dieses Tools jedoch etwas abgeflaut, da zunehmend Zweifel an seiner Reliabilität und der Validität der Ergebnisse aufkamen. Einige der Probleme sind inzwischen behoben worden, aber die Debatte über die Möglichkeiten und Grenzen von *Google Books Ngram Viewer* ist noch nicht abgeschlossen. Der vorliegende Forschungsbericht leistet einen Beitrag zu dieser Debatte, indem er eine Systematisierung basierend auf dem Konzept der *belief systems* für die Verwendung von Suchbegriffen vorschlägt und die Ergebnisse eines Reliabilitätstests für den deutschsprachigen Korpus vorstellt.

Schlüsselwörter Ideenbasierter Institutionalismus · Korpuslinguistik · Belief systems · Google Ngram · Big data

1 Introduction

In political science, the field of ideational (or discursive or constructivist) institutionalism, sometimes referred to as "fourth new institutionalism," is expanding (Berman 2013; Blyth 1997; Schmidt 2008). However, ideas are difficult to observe, and methodology has not kept pace with the rapid development of ideational research. As Swinkels (2020) notes, the field of ideational studies remains methodologically underdeveloped. This research note aims to strengthen the methodological foundations of ideational research through an evaluation of the German text corpus of *Google Books Ngram Viewer* (short: Google Ngram) and a systematization of search terms to be used for big data analysis.

In recent years, the possibility of systematically observing ideational patterns and developments at the macro level of public discourse has become more feasible thanks to the proliferation of big data. Through digital tools such as Google Trends, searchable archives of parliamentary speeches, and media archives, new possibilities in accessing and managing the flood of ideas contained in millions of sources over long periods of time have emerged.[1] The focus of this research note is on Google

[1] Examples of such archives are Google News Archive (news.google.com/archives), LexisNexis (www.lexisnexis.com), and NewspaperArchive (www.newspaperarchive.com).

Ngram, which represents a potentially valuable tool for studying ideational developments in historical perspective.[2] Google Ngram is based on the content of Google Books. It shows the frequency of words and phrases of up to five words in millions of books in eight different languages since the sixteenth century and can therefore provide unprecedented insights into the history of culture and ideas (e.g., Acerbi et al. 2013; Juola 2013; Younes and Reips 2018).

However, after initial euphoria, the use of Google Ngram has declined considerably due to growing doubts about its reliability and the validity of the results it produces (Juola 2022; Pechenick et al. 2015). For example, can a relative decrease in the use of first-person plural pronouns and an increase in first-person singular pronouns really serve as an indicator of individualization, as some authors have proposed (e.g., Twenge et al. 2013; Uz 2014)? Does an increase in the frequency of the lemma "*heilig*" (holy) in the German text corpus during the Nazi regime provide evidence of greater religiosity in times of crisis, as Younes and Reips (2019) conclude? Obviously, the use of Google Ngram presents some challenges and potential pitfalls. In particular, two issues need to be addressed in order to obtain valid results. The first one is the problem of content validity, i.e., whether and to what extent individual words or word sequences can be used as indicators of cultural patterns or ideational developments. Given the fact that individual terms (1-grams) are often ambiguous and sensitive to exogenous factors such as corpus development, how can we identify search terms that faithfully represent the construct of interest?

A solution proposed by Younes and Reips (2019) relies on the use of word inflections, synonyms, and word clusters to capture a particular cultural pattern such as religiosity. However, by including multiple words from the same semantic field, ambiguities may even increase as the contexts of use become more diverse. For example, the terms "angel" and "creed" from these authors' list of religious terms may appear in different, not necessarily religious, contexts. The challenge, then, is to identify search terms that are exclusive to the concept of interest, yet inclusive enough to capture that concept in its entirety. To this end, this research note proposes a novel strategy. Based on the concept of belief systems, a category of search terms is specified that is both semantically broad and unambiguous and thus suitable for observing ideational developments over time.

The second issue is reliability. Can ideational developments be inferred from word frequencies as shown by Google Ngram? Some caveats are in order. Since Google does not disclose the content of Google Books, word frequencies are derived from largely unknown corpora. Inferences based on word frequencies are therefore susceptible to potential biases and fluctuations in the data. Normalization procedures can help to compensate for corpus discontinuities (see, e.g., Acerbi et al., 2013), but they cannot cure biased contents. Moreover, Younes and Reips (2019) show that different normalization procedures lead to partially inconsistent results. Alternative

[2] There are similar search tools, most notably the HathiTrust Bookworm, which uses data from the HathiTrust Digital Library. It contains over 17 million digitalized texts and offers a wide range of text analysis capabilities (www.hathitrust.org). Nonetheless, the focus of this paper is on Google Ngram because it has been used much more widely than other analysis tools for the study of language and cultural history (Juola 2022).

reliability tests rely on lexical indicators. Koplenig (2015), for example, points to Helvetisms in the German corpus as evidence of discontinuities in corpus composition during the Second World War. Pechenick et al. (2015) observe that several subcorpora in Google Ngram are biased toward professional texts. According to these authors, the proportion of scientific journals has increased over time, possibly leading to an overrepresentation of scientific texts at the expense of popular culture.[3]

Doubts about the reliability of Google Ngram are therefore not unfounded, but we still know little about the composition of the corpus. In particular, it is unclear whether there are significant discontinuities over time and in which subcorpora there is a bias toward certain genres or subject areas. To address these questions for the German language corpus, a sample of publications from the German National Library (*Deutsche Nationalbibliothek*, DNB) will be compared with titles found in Google Books. On this basis, it will be shown that the content of Google Ngram can be considered representative of text production in Germany, at least for the period between 1972 and 2016, which is the time frame covered by the sample.

The main part of this research note is divided into two sections. The following section presents a strategy for tracking ideational developments through appropriate search terms. The next section describes the characteristics of Google Ngram and presents a validity test for its German subcorpora. The conclusion returns to the initial question of the applicability of Google Ngram for the study of ideas.

2 The Nature of Ideas and the Problem of Content Validity

Ideas are conceptually ambiguous, and their role and place in relation to actors and institutions are contested (Blyth 1997; Mehta 2011; Swinkels 2020). Conceptualizations of ideas range from strategic tools used instrumentally by actors to taken-for-granted scripts and worldviews that underlie social life. On an ontological level, the understanding of ideas also varies considerably. According to one view, which can be described as "postmodernist," ideas are only loosely connected and highly fluid. Ideational elements can be freely combined and constantly recombined in communicative situations or "story games." The opposite view, which might be termed "Hegelian," sees ideational elements as integral parts of tightly integrated knowledge regimes and classificatory orders characterized by a high degree of stability and internal consistency, whether guided by principles of religion or of reason.

In the study of ideas, both views are problematic because there is too much variation in the first case and almost no variation in the second. On the one hand, we should not take it for granted that ideational structures are stable and coherent, because that is what ideational approaches are intended to ascertain (Carpini and Keeter 1993). Conversely, any ideational approach is necessarily predicated on the premise that such structures exist, that they can be described on an abstract level,

[3] Despite Michel et al. (2011) indicating that periodicals had been excluded from the corpus, a considerable number of journals are evident from the ratio of the term "ISSN" (International Standard Serial Number) to "ISBN" (International Standard Book Number). It is notable that the frequency of "ISSN" is significantly higher in the English corpus than in the German corpus.

and that they are causally relevant. Otherwise, it would hardly be justifiable to speak of ideational *institutionalism*.

The concept of belief systems fulfills these requirements and lends itself to a systematic, comparative study of ideas. Belief systems are configurations of ideas and attitudes "in which the elements are bound together by some form of constraint or functional interdependence" (Converse 1964, p. 207; see also Luskin 1987). They are stable but not rigid. Belief systems are critical factors in institutional development because they give way to structured, consistent behavior at the collective level, whether because they reflect the interests and ideas of dominant actors or because they embody a structural kind of ideational power that shapes and constitutes actors' identities and perceived interests (Carstensen and Schmidt 2016).[4] Belief systems are therefore primary domains of ideational research. However, their systematic observation through corpus analysis relies on a number of conditions.

First, given the ambivalent and evolving nature of language, it is essential to have an intimate understanding of the cultural and linguistic context, including a comprehensive command of the language in question. Comparisons between different languages increase the risk of language errors and incorrect inferences. For example, Younes and Reips (2019) posit that Germans became more religious during the Nazi regime because the frequency of religious terms such as "*heilig*" (holy) increased during this period. Aside from the historical implausibility of this claim, a wildcard search in Google Ngram shows that the observed pattern in the German corpus is mainly due to the high frequency of the terms "*Heiliger Geist*" (holy spirit) and "*Heilige Schrift*" (holy scripture) and their declensions, which is hardly evidence of a general increase in religiosity. More plausibly, the observed pattern is due to discontinuities in the corpus during the Second World War (Koplenig 2015).

In addition, misinterpretations may arise from ambivalent indicators. For example, Younes and Reips (2018) use the German adjective "*eigen*" (own) and its inflections as indicators of individualization. However, although the different inflections of the term are unambiguous, "*eigen*" is also used in the sense of "peculiar." As this usage has fallen out of use over time, the frequency of "*eigen*" has decreased, which, therefore, should not be interpreted as contradicting the trend of individualization. Thus, finding appropriate indicators of cultural and ideational developments is a challenging task.

Another challenge to be considered when examining word frequencies over time is shifts in meaning. Take, for example, the term "overkill," which refers to a central concern of the peace movement: the excessive destructive capacity of the nuclear powers. The term began to spread in Germany with the acceleration of the nuclear arms race and reached its highest frequency in the late 1970s. In the course of détente and disarmament, its use declined, but even after the end of the Cold War it retained a relatively high and stable frequency of occurrence.[5] What happened was an expansion of its semantic range and an increasingly figurative use to describe

[4] On the constitutive and motivational power of ideas, see also Kestler (2023).

[5] Google Ngram Viewer, German 2019 (https://books.google.com/ngrams/graph?content=overkill+%2B+Overkill&year_start=1930&year_end=2008&corpus=de-2019&smoothing=3, accessed 26 July 2023).

various kinds of overwhelming and overpowering. Consequently, employing the term as an indicator of the beliefs of peace activists would lead to erroneous conclusions.

Belief systems can help avoid such pitfalls by providing information about the context in which a term is used. Converse (1964) describes belief systems as ideational structures held together by logical, psychological, and social constraints. These constraints imply that belief systems cannot be arbitrarily modified and their components cannot be changed like pieces in a "language game." Belief systems therefore possess a degree of consistency that allows their constituent elements to be identified and placed in context. However, not all parts of a belief system are of equal weight and importance. According to Sabatier (1988, p. 132), (political) belief systems include "value priorities, perceptions of important causal relationships, perceptions of world states (including the magnitude of the problem), perceptions of the efficacy of policy instruments, etc." These elements are organized into concentric spheres around a core of deep, shared beliefs that remain remarkably stable over time, much like Thomas Kuhn's scientific paradigms. For example, Dennis Meadows's 1972 study "The Limits to Growth" advanced the notion that the progressive depletion of natural resources as a result of industrial expansion and population growth would inevitably lead to ecological disaster. This idea has since become a fundamental tenet of ecological thought. Although the explanatory models and the corresponding policy instruments have changed over time, the core belief in the impossibility of continued growth has remained virtually unaffected by changing circumstances and expanding knowledge.

By focusing on such core beliefs, an analysis of ideational structures can find solid ground. While ideas at the periphery of belief systems are more fluid and less coherent, the ideational core remains stable and clearly defined. Terms or expressions associated with core beliefs, such as "class struggle," "limits to growth," and "imperialism," can serve as indicators for identifying the corresponding belief system. Because these terms are part of axiomatic beliefs such as "The history of all hitherto existing societies is the history of class struggles" (Marx) or "Every day of continued exponential growth brings the world system closer to the ultimate limits to that growth" (Meadows), they are highly persistent. Shifts in meaning, as in the case of "overkill," can occur and should be checked by direct searches in Google Books, but they should be the exception rather than the rule, given the constraints inherent in belief systems.

Based on these considerations, an attempt can be made to systematize search terms according to their semantic properties. As shown in the examples above, search terms can vary in their degree of precision and comprehensiveness, which can be conceived in semantic terms as intension and extension. Intension refers to a word's connotation or the range of meanings associated with it. Terms with vague intensions, such as "leaf," encompass different meanings and apply to a wide range of mental representations: a leaf of paper, a leaf from a tree, or a leaf of gold. On the other hand, the extension of a term describes its denotation or the set of objects to which it refers. Narrow terms like "violin" denote only a limited set of objects, whereas broader terms like "tree" or "house" encompass a larger portion of reality.

Content validity when using tools such as Google Ngram depends on both the precision and broadness of search terms. In terms of intension, a search term should

have unambiguous connotations relevant to the topic or construct of interest. In the study of ideas, it should accurately and exclusively reflect a particular belief system. At the same time, it should have a broad extension, capturing as much of the corresponding ideational structure as possible. For example, the term "vote of no confidence" clearly pertains to parliamentary democracy, but it fails to capture the whole concept of parliamentarism. Similarly, "Keynesianism" refers to a particular school of economic thought but does not encompass demand-side economic theories in their entirety. These terms are precise in intension but too narrow in extension.

Only a certain category of terms have both the precision of intension and the breadth of extension that make them suitable for tracking ideas through Google Ngram. Such words have a well-defined connotation while also denoting a sufficiently broad and/or relevant set of referents. This category of words and phrases can often be found at the core of belief systems, such as "class struggle," "limits to growth," and "imperialism."

As shown in Fig. 1, the suitability of search terms varies widely. To validly infer ideational developments, one must focus on the terms found in the upper left square of the figure. This requires a careful examination of the ideational context (or belief system) under analysis and the identification of relevant terms from its core. Context-free terms are of limited use, but identifying core beliefs allows tracking a belief system over time.

However, using Google Ngram for this purpose also requires addressing the second issue mentioned above: the reliability of Google Ngram data.

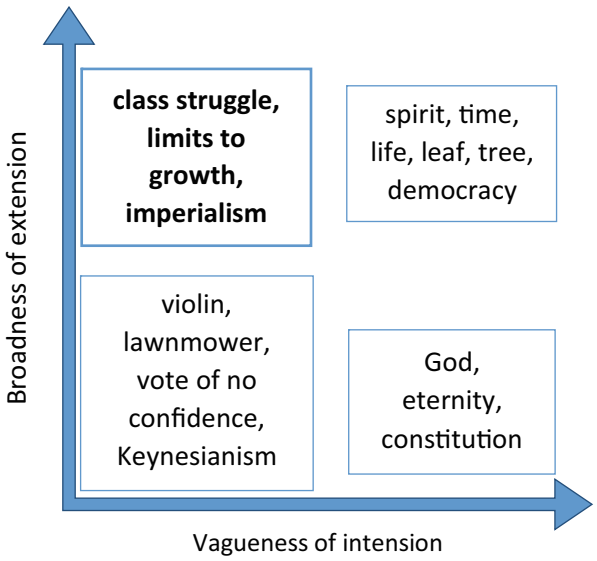

Fig. 1 Two-dimensional systematization of search terms according to semantic characteristics

3 Assessing the Reliability of Google Ngram Data

In utilizing Google Ngram, it is essential to recognize that books represent merely a fraction of the total output of cultural production. Important aspects of culture are reflected not only in books but also in newspapers, radio, television, and, since the advent of the Internet, digital media. Relying exclusively on print sources inevitably yields an uneven picture. To illustrate, between the 1960s and the early 1990s, the frequency of the search term "Rolling Stones" in the general English corpus is less than that of "John Maynard Keynes," which greatly misrepresents the popularity of the rock band the Rolling Stones. Obviously, Google Ngram is not well suited for tracking popular culture. What books do represent is a filtered and condensed picture of the ideas prevailing at a given time. The formative currents of Western thought—the Reformation, the Enlightenment, liberalism, Marxism, socialism—can all be traced back to foundational books and are reflected in text production. This is also true of the economic ideas of J. M. Keynes, whose influence on postwar economic thought certainly exceeds the ideational relevance of the Rolling Stones.

The misrepresentation of popular culture in Google Ngram is due not only to the specificity of the medium book, which represents only a certain segment of cultural production, but also to the fact that it does not account for circulation. Each book is digitalized only once by Google, and all books are weighted equally in the calculation of N-gram frequencies, be it an unnoticed dissertation with only a handful of copies or a bestseller with a print run of millions. This may not be an obstacle to observing general cultural or linguistic patterns, which are reflected in all books. In the study of ideas, however, circulation matters, because it shows the reach and influence of an idea. To distinguish small but highly productive sects from broad ideational currents, additional tools for cross-validation are necessary.[6]

Beyond these caveats, which apply in one way or another to all large corpora, Google Ngram presents an additional problem of reliability. Its data originates from Google Books, a project that started in 2004 with the goal of digitalizing all printed texts. It currently comprises more than 40 million digitalized books from dozens of university libraries around the world, or nearly one-third of all books ever published (Lee 2019), thus providing access to substantial portions of global ideational production for the first time in history. For tracking ideas and ideational developments over time, Google Books is a potential gold mine. In its original form, however, it allows searches only for individual books, which appear in a list ordered by relevance. The listed books can be searched, but only individually and to the extent allowed by copyright regulations.[7]

To circumvent these limitations, two linguists, Erez Aiden and Jean-Baptiste Michel, created a shadow image of Google Books by recording the frequencies

[6] Examples of cross-validation of Google Ngram results are provided by Richey and Taylor (2020), who show, for example, that the frequency of the term "political corruption" correlates strongly with data from Transparency International.

[7] After a series of lawsuits, Google had to significantly restrict access to Google Books, which, in turn, led the company to slow down the scanning process in the years that followed. As a result, the database is not expected to expand any further in the future (Somers 2017).

by year of individual words and sequences of up to five words from Google's stock of digitalized books. Graphical representations of word frequencies from eight languages as well as the raw data can be accessed through the Google Ngram Viewer (https://books.google.com/ngrams/). Aiden and Michel (2013) themselves tracked the evolution of language, showing, for example, when English verbs became regularized or when a word disappeared from the vocabulary. According to Michel et al. (2011), the data also reflect the blacklisting of certain names during the Nazi regime in Germany.

However, there are growing concerns about the reliability of Google Ngram. Critics point to optical character recognition errors and possible bias toward certain types of text. These criticisms cannot be confirmed or refuted because Google's scanning process is largely automated, and the company maintains strict confidentiality about what it has scanned. Furthermore, the Google Ngram corpus differs from the original content of Google Books because the data had to be cleaned of erroneous scans and incorrect attributions. This cleaning reduced the total corpus to about five million books, or 4% of all books ever published (Aiden and Michel 2013; Michel et al. 2011). After an update in 2012, the total Google Ngram corpus grew to eight million books, divided into 22 different subcorpora (Younes and Reips 2019), but it still represents only a fraction of Google Books, which, in turn, includes only a part of global text production. Thus, although Google Ngram is larger than any database assembled to date, it is the result of a double-selection process that may have introduced bias into the data. In particular, the first step of scanning by Google may have introduced a bias due to the particularities of the literature represented in university libraries and varying copyright regulations. We do not know if all subject areas and genres are adequately represented. Nor can we say with certainty whether the corpora remain consistent over time or whether there are significant changes in their composition, given that the corpus has grown disproportionately in recent years. As Koplenig (2015, p. 2) puts it, "We cannot check whether the different diachronic books samples really represent similar things at different moments in time." Thus, examining the consistency and representativeness of the Google Ngram data is essential if it is to be used to measure ideational development.

In addition to normalization procedures and the search for lexical regionalisms (see, e.g., Koplenig 2015; Pechenick et al. 2015), simple consistency tests can be performed based on the frequency of common and neutral word pairs, such as "day"/"night," "heaven"/"earth," or "warm"/"cold," which are presumably not affected by cultural change and are therefore expected to show little variation over time. If the frequencies of such pairs are steady and largely parallel, we may regard this as indicative of a consistent composition of the corresponding corpus. In fact, Google Ngram passes this test with only modest results. Between the years 1900 and 2019, the frequency of "cold" largely parallels that of "warm," but it varies between 46 and 100 occurrences per million words. Similarly, the pairs "heaven"/"earth" and "day"/"night" also run in parallel but with considerable variation.

Part of this variation is explained by changes in the access to literature after the year 2004. When Google began its scanning process, publishers were asked to send new publications in for digitalization, whereas older books had to be made available through university libraries. As a result, newer books are significantly over-

Table 1 Google Ngram corpus size, 1972–2016 and sample sizes and shares available in Google Books

Years	1972–1976	1977–1981	1982–1986	1987–1991	1992–1996	1997–2001	2002–2006	2007–2011	2012–2016	Total	Average
Sample size, *n*	161	214	282	329	457	499	587	535	824	3888	432
Share of sample, %	4.14	5.5	7.25	8.46	11.75	12.83	15.1	13.76	21.19	100	–
Share of sample available in Google Books, *n*	159	211	275	319	444	487	570	520	785	3770	419
Share of sample available in Google Books, %	98.76	98.6	97.5	96.96	97.16	97.6	97.1	97.2	95.27	–	97.35
Number of books in the Google Ngram subcorpus (2012, German), average per year, *n*	5465	6097	5767	5484	5940	7371	11,624	28,684[a]	n. a.	336,706	9554
Share of the Google Ngram subcorpus (2012, German) in total book production in Germany, %[b]	11.7	10.5	9.78	8.33	8.47	9.11	13.49	30.28[a]	n. a.	–	12.71

Sources (accessed 13 June 2024): Deutsche Nationalbibliothek (https://portal.dnb.de/opac.htm), Google Books (https://books.google.de/), Statistisches Bundesamt (https://www.destatis.de/DE/Themen/Querschnitt/Jahrbuch/statistisches-jahrbuch-aktuell.html), Börsenverein des Deutschen Buchhandels (https://www.boersenverein.de/markt-daten/marktforschung/wirtschaftszahlen/buchproduktion/), Google Books Ngram Viewer Exports (https://storage.googleapis.com/books/ngrams/books/datasetsv3.html)
n.a. not available
[a]These values represent the average of 2007, 2008, and 2009, as the data provided by Google cover only the years through 2009
[b]These figures are an approximation, as the calculation is based only on book production in Germany, which includes foreign-language publications but excludes German-language books from Austria and Switzerland

represented in the Google Ngram corpora, especially since 2006. For the preceding decades, corpus development can be regarded as fairly consistent. The number of books per year included in the 2012 German subcorpus increases from around 2000 titles per year in the 1950s to over 10,000 titles in the 2000s, but this increase largely reflects the growing number of book publications in Germany. In relative terms, the development of the corpus is rather continuous (Table 1). After 2006, however, the corpus grows abruptly to almost 30,000 titles per year, representing almost a third of all book publications in Germany. As a result, the composition of the corpus changes significantly, which is probably the cause of the discontinuities in the data. For this reason, Erez Aiden, one of the creators of Google Ngram, recommends using the tool only up to the year 2006 for the study of the history of culture and ideas.[8]

However, observations about the relative size of the corpus tell us little about its composition. We do not know whether the Google Ngram data reliably represent text production in a given language area. In order to clarify this question, it is necessary to evaluate the individual language corpora separately. This analysis focuses on the German language corpus, which is subjected to a reliability test. Since the Google Ngram data originate from Google Books, a random sample of all books published in Germany since the early 1970s is compared with the content of Google Books in order to assess possible biases in the data. This procedure is based on the assumption that biases are most likely due to Google's book-scanning process. If Google Books accurately reflects text production in Germany, we can consider the word frequencies displayed by the Google Ngram Viewer to be reliable, albeit with reduced degrees of freedom due to the smaller corpus size. In particular, older books had to be sorted out more often due to their lower printing quality, so the data become less precise the further back in history we go. In terms of content, however, there should be no systematic bias as a result of technical cleaning.

To test the completeness of Google Books, we relied on a sample of International Standard Book Numbers (ISBNs), which are easier to handle than the complete metadata provided by the DNB. The analysis covered only publications with a German language code (including Austrian and German-language Swiss publications) going back to the early 1970s, when the assignment of ISBNs started.[9] First, all available ISBNs were extracted from MARC 21 files[10], cleared of duplicates, and sorted by country code, resulting in a number of slightly over eight million ISBNs. In the next step, a random sample of 5000 ISBNs was drawn and manually complemented with additional metadata: publication year, Dewey Decimal Classification number, subject area, and keywords. From this sample, maps and calendars (which also carry an ISBN) as well as books in languages other than German were removed, and the publication dates were limited to the years 1972 to 2016, which are completely represented by ISBNs in the DNB data.[11] This step resulted in a total of 3888 titles.

[8] Personal conversation (online), 29 June 2023.

[9] The ISBN system was introduced in the mid-1960s, but it was only over time that it became standard practice. There are also older books carrying ISBNs that were assigned retroactively.

[10] MARC stands for MAchine-Readable Cataloging and is a bibliographic data format.

[11] The data from the DNB were downloaded in 2019, which means that the immediately preceding years are incomplete because not all publications had yet been entered into the cataloging system.

Finally, Google Books was searched for the availability of these book titles to see if the relative weight of publication years and subject areas in the sample matched the distribution of titles found in Google Books. In fact, 97.35% of all book titles were found to be available in Google Books. This means that the stock of books digitalized by Google can be regarded as representative, even without considering the relative weight of subject areas and publication years.

Since Google Books covers almost the entirety of book publications in Germany for the period under study, the frequencies of word usage as cataloged by Google Ngram can be considered reliable, at least until the year 2006. These frequencies can serve as a valuable source for tracing ideational developments, provided that indicators are applied appropriately.

4 Conclusion

The study of ideas has become a burgeoning branch of the new institutionalism, but it faces methodological obstacles, as the quantity and fluidity of ideas raise the problem of inference. How can relevant ideational developments on the macro level be discerned from a necessarily limited set of observations? Google Ngram seems to offer a way out by allowing the observation of ideas on a large scale and over long periods of time. However, concerns about the reliability of this tool have hampered its use in recent years.

The aim of this research note was to assess the possibilities and limitations of Google Ngram in the study of ideas. Tying in with similar efforts, e.g., by Younes and Reips (2019), Koplenig (2015), and Pechenick et al. (2015), problems of validity and reliability were addressed. Building on the concept of belief systems, a systematization of search terms with respect to their semantic properties was proposed to identify search terms of sufficient precision and comprehensiveness to serve as indicators for tracing historically relevant ideas. Because of their persistence, relevance, and unambiguity, terms from the core of belief systems were considered as particularly well suited for the study of ideas through Google Ngram. It should be noted, however, that this type of indicator is not equally suitable for observing more diffuse cultural patterns such as religiosity or individualism. Nevertheless, the two-dimensional matrix of semantic properties may help to improve other approaches such as the use of synonyms or semantic clusters.

To address the issue of reliability, a procedure based on a random sample of book titles drawn from the DNB was applied. It was shown that the content of Google Books represents almost the entirety of book publications in Germany over the last five decades, which leads to the conclusion that the German corpus of Google Ngram, although significantly smaller than the content of Google Books, can also be regarded as representative, at least for the period covered by the sample. The findings of Pechenick et al. (2015) about a strong bias toward scientific texts are not confirmed for the German corpus. However, given the fact that word frequencies in Google Ngram are insensitive to circulation, additional sources and secondary literature should be used for cross-validation. Moreover, given the evolution of

language use and communication, it is advisable to limit the analysis to periods of no more than a few decades to ensure the robustness of the results.

Since not all ideational developments are equally reflected in book publications, Google Ngram may be insufficient as a stand-alone source, but it offers valuable insights, as influential ideas can be expected to leave traces in books and should therefore be observable from word frequencies. Thus, Google Ngram may not solve all the methodological problems of ideational institutionalism, but it can be part of the solution.

Funding This research was supported by the Fritz Thyssen Foundation (grant 20.22.0.006PO)

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest T. Kestler declares that he has no competing interests.

References

Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. The expression of emotions in 20th century books. *PloS one* 8(3):e59030. https://doi.org/10.1371/journal.pone.0059030.

Aiden, Erez, and Jean-Baptiste Michel. 2013. *Uncharted. Big data as a lens on human culture*. New York: Riverhead Hardcover.

Berman, Sheri. 2013. Ideational theorizing in the social sciences since "policy paradigms, social learning, and the state". *Governance* 26(2):217–237. https://doi.org/10.1111/gove.12008.

Blyth, Mark. 1997. "Any more bright ideas?": the ideational turn of comparative political economy. *Comparative Politics* 29(1):229–250. https://doi.org/10.2307/422082.

Carpini Delli, Michael X., and Scott Keeter. 1993. Measuring political knowledge: putting first things first. *American Journal of Political Science* 37(4):1179. https://doi.org/10.2307/2111549.

Carstensen, Martin B., and Vivien A. Schmidt. 2016. Power through, over and in ideas: conceptualizing ideational power in discursive institutionalism. *Journal of European Public Policy* 23(3):318–337. https://doi.org/10.1080/13501763.2015.1115534.

Converse, Philip E. 1964. The nature of belief systems. In *Ideology and discontent* International Yearbook of Political Behavior Research, Vol. 5, ed. David E. Apter, 206–261. New York: Free Press of Glencoe.

Juola, Patrick. 2013. Using the Google N-gram corpus to measure cultural complexity. *Literary and Linguistic Computing* 28(4):668–675. https://doi.org/10.1093/llc/fqt017.

Juola, Patrick. 2022. Google books Ngrams. In *Encyclopedia of big data*, ed. Laurie A. Schintler, Connie L. McNeely, 517–521. Cham: Springer.

Kestler, Thomas. 2023. The motivational power of ideas in institutions and collective action: collectivizing intentional states and bodily awareness. *Human Studies* 46(1):59–78. https://doi.org/10.1007/s10746-023-09661-x.

Koplenig, Alexander. 2015. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* https://doi.org/10.1093/llc/fqv037.

Lee, Haimin. 2019. 15 years of Google books. https://www.blog.google/products/search/15-years-google-books/. Accessed 6 Nov 2021.

Luskin, Robert C. 1987. Measuring political sophistication. *American Journal of Political Science* 31(4):856–899.

Meadows, Dennis L. 1972. *The limits to growth. A report for the club of rome's project on the predicament of mankind*. New York: Universe Books.

Mehta, Jal. 2011. The varied roles of ideas in politics. In *Ideas and politics in social science research*, ed. Daniel Béland, Robert Henry Cox, 23–46. Oxford, New York: Oxford University Press.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182. https://doi.org/10.1126/science.1199644.

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PloS one* 10(10):e137041. https://doi.org/10.1371/journal.pone.0137041.

Richey, Sean, and J. Benjamin Taylor. 2020. Google books Ngrams and political science: two validity tests for a novel data source. *PS: Political Science & Politics* 53(1):72–77. https://doi.org/10.1017/S1049096519001318.

Sabatier, Paul A. 1988. An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy Sciences* 21(2–3):129–168. https://doi.org/10.1007/BF00136406.

Schmidt, Vivien A. 2008. Discursive institutionalism: the explanatory power of ideas and discourse. *Annual Review of Political Science* 11:303–326. https://doi.org/10.1146/annurev.polisci.11.060606.135342.

Somers, James. 2017. Torching the modern-day library of Alexandria. The Atlantic. 20 April 2017. https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/. Accessed 28 Mar 2023.

Swinkels, Marij. 2020. How ideas matter in public policy: a review of concepts, mechanisms, and methods. *International Review of Public Policy* 2(3):281–316. https://doi.org/10.4000/irpp.1343.

Twenge, Jean M., W. Keith Campbell, and Brittany Gentile. 2013. Changes in pronoun use in American books and the rise of individualism, 1960–2008. *Journal of Cross-Cultural Psychology* 44(3):406–415. https://doi.org/10.1177/0022022112455100.

Uz, Irem. 2014. Individualism and first person pronoun use in written texts across languages. *Journal of Cross-Cultural Psychology* 45(10):1671–1678. https://doi.org/10.1177/0022022114550481.

Younes, Nadja, and Ulf-Dietrich Reips. 2018. The changing psychology of culture in German-speaking countries: a Google Ngram study. *International Journal of Psychology* 53(Suppl. 1):53–62. https://doi.org/10.1002/ijop.12428.

Younes, Nadja, and Ulf-Dietrich Reips. 2019. Guideline for Improving the Reliability of Google Ngram Studies: Evidence from Religious Terms. *PloS one* 14(3):e213554. https://doi.org/10.1371/journal.pone.0213554.