International Journal of Digital Humanities (2025) 7:393–416
https://doi.org/10.1007/s42803-025-00113-9

REVIEW ARTICLE

Survey on information requirements on the Google Books Ngram Corpus

Fabian Richter[1] · **Federico Matteucci**[1] · **Peter Reimann**[1] · **Klemens Böhm**[1]

Received: 29 August 2025 / Accepted: 15 November 2025 / Published online: 8 December 2025

Abstract

The development of word frequencies over time is the subject of research in different branches of the humanities. Large temporal n-gram corpora have been created for this purpose, most notably the *Google Books Ngram Corpus*. While the concrete research questions vary between the different research works, there are similarities in the more abstract underlying information requirements, i.e., the structure of queries against a potential database system. Based on a systematic literature review, we extract these information requirements, leading to a categorization of existing articles into macro-areas of information requirements. Furthermore, we collect existing query systems for temporal n-gram corpora and evaluate their expressiveness regarding the information requirements we found.

Keywords Google Books Ngram Corpus · Information requirements · Literature survey · Query language

✉ Fabian Richter
fabian.richter@kit.edu

Federico Matteucci
federico.matteucci@kit.edu

Peter Reimann
peter.reimann@kit.edu

Klemens Böhm
klemens.boehm@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany

1 Introduction

Analyzing the evolution of word frequencies in collections of written documents over time is a valuable tool often used, for instance, in the fields of linguistics (Schlüter and Vetter, 2020; Breit, 2017; Davies, 2014), psychology (Younes and Reips, 2018; O'Sullivan et al., 2019; Teepe et al., 2023), and philosophy (Willkomm et al., 2018; Schmidt-Petri et al., 2021). To help researchers access this information, a number of diachronic n-gram corpora have been created. These include the *Google Web 1T* corpus (Brants and Franz, 2006) and the n-gram corpus of the Norwegian National Library (Birkenes et al., 2015). Among them, the Google Books Ngram Corpus (GBNC) (Lin et al., 2012; Mann et al., 2014; Michel et al., 2011) is by far the largest, comprising approximately 6% of all books ever published (Lin et al., 2012). For this reason and due to its vast influence, we focus on this specific corpus. The GBNC contains the yearly frequencies of billions of n-grams in several different languages. It was constructed from parts of the *Google Books* collection, with contributions from international libraries and universities. The exact composition of the corpus is, however, unknown.

Information requirements *When analyzing the GBNC, researchers use query systems to search for some specific kind of information, such as significant changes in word frequencies or correlations to real-world events. Understanding the researchers' information requirements (or information needs) is therefore crucial for designing useful query systems (Taylor, 1968; Owei, 2003). However, the current literature on the topic is limited. A non-systematic, small-scale study on 11 publications has investigated which subsets of the GBNC are particularly interesting for researchers (Richter and Böhm, 2024) and proposes a workflow to efficiently extract such subsets. In addition, Richter et al. (2025) examined existing query systems for temporal n-gram corpora with regards to their general expressiveness and usability. A similar effort is the Archive Unleashed project (Ruest et al., 2021), investigating how researchers from the humanities use web archives in general, with a focus on usability by non-technical users. Notable examples of query systems that are tailored to the Digital Libraries community are LexiDB (Coole et al., 2016, 2020) for corpus management and SchenQL (Kreutz et al., 2020, 2022) for metadata analysis.*

A systematic literature review on information requirements regarding temporal n-gram corpora is still missing. This gap in the existing literature sparked the research questions for this paper:

- ***Q1***: What kind of (implicit or explicit) information requirements do researchers have when working with the GBNC?
- ***Q2***: Are existing query systems for the GBNC powerful enough to satisfy these information requirements?
- ***Q3***: If not, what is missing in these query systems?

Contributions *Our main contribution is an extensive, systematic literature review on applications of the GBNC in research. We survey, summarize and categorize 69*

existing research articles and extract common information requirements from them. Furthermore, we discuss the ability of existing query systems to satisfy these information requirements.

Paper outline *Section 2 introduces fundamental definitions and notations used in the following. In Section 3, we describe the process and results of our literature review In Section 4, existing query systems and their expressive power are discussed. Section 5 summarizes our findings.*

2 Fundamentals

This section presents the basic concepts used in the next sections, largely following the definitions of Richter et al. (2025). We first describe temporal n-gram corpora (TNCs) and how to transform a text corpus into a TNC. Then, we define the term information requirements.

2.1 Temporal n-gram corpora

A temporal n-gram corpus (TNC) can be built from a temporal text corpus and contains information about the evolution of frequencies of n-grams (i.e., sequences of n words) in that corpus over time. For instance, the GBNC is based on the *Google Books* corpus (Michel et al., 2011). Before defining TNCs, we first fix an alphabet, i.e., a collection of symbols, which we indicate with Σ. The usual choice for Σ, in the context of query systems for the GBNC, is the set of English words, numbers and punctuation symbols, together with a wildcard symbol $*$. In this section, we assume that all the documents and n-grams are built from the same alphabet.

Temporal text corpora *A text document is a semi-structured data object (Abiteboul, 1997; Buneman, 1997)* $d = (c_d, m_d)$*, where*

- c_d is the *content* — a text written with symbols from Σ;
- m_d is the *metadata* — a key-value mapping of properties of the text.

Given a metadata key k, we indicate with $d[k]$ the corresponding value — as a shorthand notation for $m_d(k)$. If k is not in the domain of m_d, i.e., it is not a valid key for m_d, we define $d[k] := \perp$.

A text corpus D *is a set of text documents,* $D = \{d_1, \ldots, d_l\}$*. A text corpus is temporal if the metadata of its documents contains time-related information. More formally, we say that* D *is temporal if there exists some timestamp key* k_t *such that, for any document* $d_i \in D$*,* $d_i[k_t]$ *is a timestamp. This timestamp key can, for instance, be 'publication date'.*

Temporal n-gram corpora (TNC) *An n-gram* $g = (\sigma_1, \ldots, \sigma_n) \in \Sigma^n$ *is an ordered tuple of symbols belonging to* Σ*. The frequency an n-gram* g *occurs in a document* d *is the number of times the concatenated symbols* $\sigma_1 \ldots \sigma_n$ *(ignoring whitespace*

characters) appear in $\mathbf{c_d}$. *The frequency function* $\boldsymbol{f}_{\mathrm{g}} : \boldsymbol{T} \mapsto \boldsymbol{n} \in \mathbb{N}_{\mathbf{0}}$ *associates to a timestamp* $\boldsymbol{T}$ *the total number of occurrences of* g *within all documents in* $\boldsymbol{D}$ *with timestamp* $\boldsymbol{T}$. *W.l.o.g., we also allow* $\boldsymbol{f}_{\mathrm{g}}$ *to take as input an ordered tuple of timestamps* T. *In this case,* $\boldsymbol{f}_{\mathrm{g}}$ *returns an ordered tuple of frequencies.*

We can finally define a temporal n-gram corpus, which contains temporal information on the frequency of n-grams over a certain time span.

Definition 1 (Temporal n-gram corpus (TNC) (Richter et al., 2025)) A *temporal n-gram corpus* is a tuple $\boldsymbol{C} = (\boldsymbol{G}, \mathrm{T}, \boldsymbol{F})$, where $\boldsymbol{G}$ is a set of n-grams, $\mathrm{T} = (\boldsymbol{T_1}, \ldots, \boldsymbol{T_m})$ is an ordered tuple of unique timestamps, and $\boldsymbol{F} = (\boldsymbol{f}_{\mathrm{g}})_{\mathrm{g} \in \mathbf{G}}$ is a family of frequency functions, containing one frequency function per n-gram.

2.2 Information requirements

Information requirements, also known as information needs, are a key component to the design of query systems (Wilson, 1981; Taylor, 1968). In this paper, by information requirement we mean a *formalized need*, i.e., a rigorous and precise description of the user's conscious information need (Taylor, 1968). A simple information requirement regarding the GBNC could be, for example, *'What is the most frequent n-gram in the year 2017?'*. Users typically try to satisfy their information requirements with a query system. As a consequence, query systems should be designed to address the specific information requirements of their potential users (Wilson, 1981).

3 Survey of information requirements for the GBNC

The goal of our survey is to answer the first of our research questions, **Q1**: What kind of (implicit or explicit) information requirements do researchers have when working with the GBNC? **Q2** and **Q3** are further discussed in Section 4.

First, we present the methodology of the survey, including our criteria for literature selection. Then, we classify papers according to how they make use of the GBNC. Finally, we discuss and categorize the information requirements we identified.

3.1 Survey methodology

We conduct a systematic literature survey, following accepted guidelines for such studies (Kitchenham, 2004; Moher et al., 2010; Page et al., 2021) and aim to make our work reproducible by establishing a survey procedure as well as rigorous, objective literature selection criteria. We publish all the relevant code and data.[1]

[1] https://github.com/DrCohomology/GBNC_Survey

3.1.1 Literature selection

Over the years, the GBNC has been the subject of a substantial amount of studies. For instance and just using the Google Scholar search engine with specific keywords,[2] we could easily retrieve about 1000 distinct publications. The number of retrieved publications increases by more than one order of magnitude when including alternate spellings and/or orderings of the keywords.

A major obstacle for our survey is the fact that most of these publications mention the GBNC only as related work or as a side note. These publications are not relevant for our research question, as they do not pose any information requirements towards the GBNC. To filter out such works and to focus on articles that work directly with the GBNC, we consider only publications that contain `google ngram`[3] in the title.

Our reasoning for choosing this specific set of queries is as follows: Mentioning a dataset in the title of a research article indicates that this specific dataset plays an important role in the article. We acknowledge that the inverse is not always true, especially when a dataset is known under different names. What we call 'Google Books Ngram Corpus' is also referred to as 'Google Ngram Corpus' or 'Google Books Corpus' in literature. The first case is covered by our queries, the second one is not. We have considered, but ultimately rejected, other queries as well: Works that have `google books` but not `ngram` in the title usually refer to the broader *Google Books* project, including full texts and metadata, and their number is too large for manual screening. Further restriction to works including `google books corpus` but not `ngram` in their title yields only 11 results, none of which would pass all other selection criteria.

While manual screening of all papers is, when feasible, preferable, we deemed this automatic solution necessary to obtain an impartial and reproducible survey while keeping screening efforts manageable. Furthermore, with this procedure, we intended to include more publications that focus on working with TNCs—rather than just mentioning them.

Thus, 166 results remain, 8 of which were duplicates. We then excluded 55 more papers written in languages other than English; 19 that are blog posts, presentations, or book chapters; and 15 papers of which we could not retrieve the full text. This leaves a total of 69 articles to be considered, the *PRISMA 2020* (Page et al., 2021) flow chart of this procedure is presented in Fig. 1. In some older publications within this set of 69 articles, *Google Ngram Corpus* refers to the *Google Web 1T corpus* (Brants and Franz, 2006) instead of the GBNC. Since the two corpora are structurally similar, we decided to include these publications nonetheless.

3.1.2 Goals and procedure

The core goal of our survey is to answer **Q1**: What kind of (implicit or explicit) information requirements do researchers have when working with the GBNC? A second-

[2] https://scholar.google.com/scholar?q="google+books+ngram+corpus", last accessed August 26, 2025.

[3] or slight spelling variations, namely: google n-gram, google ngrams, and google n-grams; with all queries being case-insensitive

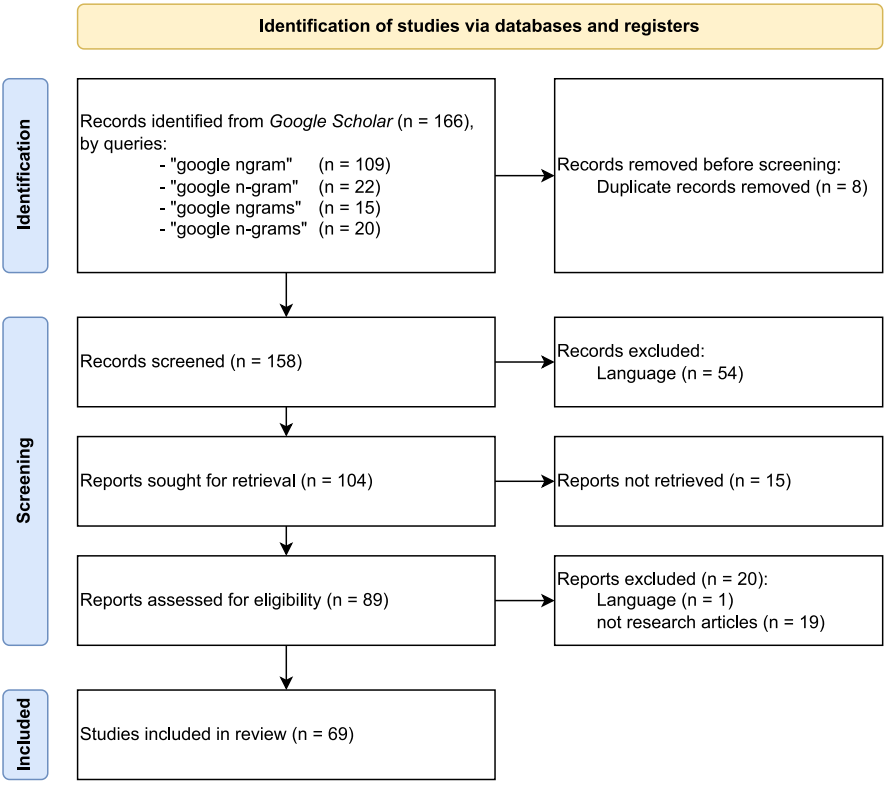

Fig. 1 The *PRISMA 2020* (Page et al., 2021) flow chart of our literature selection procedure

ary goal is to identify patterns within these requirements, for example by examining which types of studies have which information requirements. We therefore perform two orthogonal categorizations: grouping of publications with regard to their usage of the GBNC (cf. Section 3.2.1), and grouping of information requirements into macro-areas (cf. Section 3.2.2).

The categorization procedure was as follows: First, author A1 surveyed an initial set of 30 publications, selected at random. From these publications, A1 extracted the concrete information requirements and assigned them to their respective macro-areas. During this phase, A1 also fixed the four usage groups and assigned the respective articles to them, reserving the option to add more groups if necessary. Second, author A2 then repeated the analysis on another set of 20 publications, assigning publications to groups and information requirements to macro-areas. Differences in nomenclature of information requirements were then resolved by A1 and A2 together. The remaining 19 publications were again assigned to A1, who applied the same procedure as before, the results were checked by A2 afterwards. Details regarding the statistical analyses in Section 3.3.1 are discussed there.

3.2 Survey overview

We classify the publications according to how they use the GBNC and aggregate the information requirements into macro-areas.

3.2.1 Usages of the GBNC

We have observed that the 69 investigated research articles can be grouped into four distinct usage groups. This grouping is based on the viewpoint the respective works take towards the GBNC:

- Some researchers treat the GBNC as *data*, for example by using it to train statistical models. In these works, the GBNC is used as a tool, not as a subject of research in itself.
- In other articles, the GBNC is *analyzed* in its own right. Here, researchers try to understand the inner workings of the GBNC, for example regarding its composition or statistical anomalies within the corpus.
- Several *query systems* have been proposed, aiming to make analyses of the GBNC and similar corpora easier.
- Finally, some articles introduce *new corpora*, similar to the GBNC, but in different languages or tailored to specific domains, for example by only including scientific texts, as in the work of Chumtong and Kaldewey (2017).

3.2.2 Categorization of information requirements

We distinguish six macro-areas of information requirements. Each of these macro-areas contains a finer-grained set of information requirements. The categorization is summarized in Table 1.

Frequency plots *Plotting the frequencies of specific n-grams is the main requirement satisfied by the Google Books Ngram Viewer (Michel et al., 2011; Mann et al., 2014; Lin et al., 2012). These plots are often used to visually identify and explore trends or dependencies between frequencies.*

n-gram filtering *n-gram filtering deals with the search of n-grams matching a specific pattern. We identified specific n-grams search, wildcard search, part-of-speech (PoS) search, and inflection search within the investigated publications.*

Frequency filtering *Frequency filtering is related to filtering for specific n-grams* g *whose frequency function* f_g *satisfies certain criteria. The criteria we found are: similarity between frequency functions, peaks/valleys of the frequency function to occur in certain time periods, and first/last occurrence of an n-gram.*

External knowledge *Sometimes, the information within the GBNC is not enough to satisfy the conscious need of the users. In this case, they might add external knowledge to their query, for example to relate real-world data like stock prices or lit-*

Table 1 Types of information requirements

Macro-area	Information requirements	
Frequency plots	frequency plots	
n-Gram filtering	specific n-grams search	wildcard search
	PoS search	inflection search
Frequency filtering	similarity between frequencies	peaks and valleys of an $\boldsymbol{f}_{\mathrm{g}}$
	first/last occurrence	
External knowledge	real-world data	dictionaries
	other corpora	
Natural Language Processing	sentiment analysis	collocates
	named entity recognition	word embeddings
	topic grouping	
Statistical computations	change point detection	dependency analysis
	trend prediction	similarity between frequencies
	entropy	n-grams clustering

eracy rates to frequencies within the corpus. To handle this, it is important that the query system can easily consider data from different sources. We found the following most common external knowledge bases: real-world data, dictionaries (such as WordNet (Miller, 1995; Fellbaum, 1998)), and other corpora (such as the Corpus of Historical American English (Davies, 2012)).

Natural language processing (NLP) *This macro-area encompasses all information requirements related to analyzing or processing natural language. This includes sentiment analysis, collocates, named entity recognition, word embeddings, topic grouping.*

Statistical computations *Finally, the user's conscious need might require a more sophisticated statistical analysis of the n-gram frequencies. We identified change point detection, dependency analysis, trend prediction, similarity between frequency distributions, entropy, and clustering of n-grams.*

3.3 Information requirements on the GBNC

In the following, we first describe different applications of the GBNC in the 69 articles we investigated. Afterwards, we discuss their information requirements according to our macro-areas proposed earlier.

3.3.1 Overview of results

Table 2 summarizes our findings, distinguishing the usage groups of publications and macro-areas of information requirements, as introduced in Section 3.2. Figure 2 shows a box plot for the number of information requirements per publication, broken

Table 2 Overview of the major results of our survey

Usage of GBNC	Reference	Main Information Requirements					
		freq. plots	n-gram filter	freq. filter	ext. knowl.	NLP	stat. comp.
	Basile et al. (2016)				✓		✓
	Bhowmick et al. (2015)						
Corpus	Bochkarev et al. (2016)						✓
	Nazar and Renau (2012)		✓				
as Data	Perrie et al. (2013)		✓		✓	✓	
	Savinkov et al. (2021)						✓
	Yu et al. (2007)		✓				
	Bai and Huang (2018)	✓	✓				✓
	Bochkarev et al. (2020a)					✓	
	Bochkarev et al. (2020b)		✓				✓
	Bochkarev et al. (2018)						✓
	Brandt (2018)	✓	✓				
	Bukhtoyarov and Bukhtoyarova (2018)	✓	✓		✓		
	Caruana-Galizia (2016)	✓	✓		✓		✓
	Chen and Huang (2022)	✓	✓				
	Chiru and Dinu (2017)			✓	✓		✓
	Clark et al. (2022)	✓	✓				
	Coraci et al. (2020)	✓					
	Dönmez (2018)	✓	✓				✓
	El-Ebshihy et al. (2018)	✓	✓			✓	✓
	Friginal et al. (2014)	✓	✓		✓		✓
	Galeev and Solovyev (2018)	✓	✓				
	Grant and Walsh (2015)	✓	✓				✓
	Guan et al. (2024)		✓		✓	✓	✓
	Gulordava and Baroni (2011)						✓
	Hill and Simha (2016)					✓	
	Inkpen and Islam (2010)					✓	✓
	Islam and Inkpen (2009)		✓	✓			
	Islam et al. (2012)		✓		✓		✓
	Joubarne and Inkpen (2011)		✓		✓	✓	✓
	Juola (2013)						✓
	Klein and Nelson (2009)			✓			✓
Corpus Analysis	Koplenig (2017)	✓	✓				✓
	Lin et al. (2012)		✓				
	Macedo (2013))	✓	✓				
	Madsen and Slåtten (2022)	✓	✓				
	Moskovkin et al. (2019)	✓	✓				
	Nestik et al. (2022)	✓	✓		✓		✓
	Olimid et al. (2023)	✓	✓				
	Olimid et al. (2024)	✓	✓				
	O'Sullivan et al. (2019)	✓	✓				
	Pekina et al. (2018)		✓				✓
	Radimský (2022)		✓				

Table 2 (continued)

Usage of GBNC	Reference	Main Information Requirements					
		freq. plots	n-gram filter	freq. filter	ext. knowl.	NLP	stat. comp.
	Ribeiro and Lima (2017)	✓	✓				
	Richey and Taylor (2020)	✓	✓		✓		✓
	Roth (2013)	✓	✓				
	Roth (2014)	✓	✓				
	Roth et al. (2017)	✓	✓				
	Sampsel (2021)	✓	✓		✓		✓
	Skrebyte et al. (2016)	✓	✓		✓		
	Solovyev et al. (2019)	✓	✓		✓		
	Sparavigna and Marazzato (2015)	✓	✓				
	Teepe et al. (2023)	✓	✓				✓
	Younes and Reips (2018)	✓	✓		✓		✓
	Younes and Reips (2019)		✓				✓
	Zeng and Greenfield (2015)	✓	✓		✓		✓
	Zięba (2018)	✓	✓				
	Zyukina et al. (2020)	✓	✓				
	Aleksandrov and Strapparava (2012)		✓		✓	✓	
Query	Davies (2014)		✓				
	Grabowski and Swacha (2012)		✓				
System	Habeeb et al. (2020)						
	Mann et al. (2014)	✓	✓				
Design	Schlüter and Vetter (2020)						
	Schmidt-Petri et al. (2021)		✓	✓	✓	✓	✓
Corpus	Alsmadi and Zarour (2018)				✓		
	Chumtong and Kaldewey (2017)				✓		
Construction	Ivanov (2014)		✓		✓		
	Ivanov and Solovyev (2020)		✓		✓	✓	✓

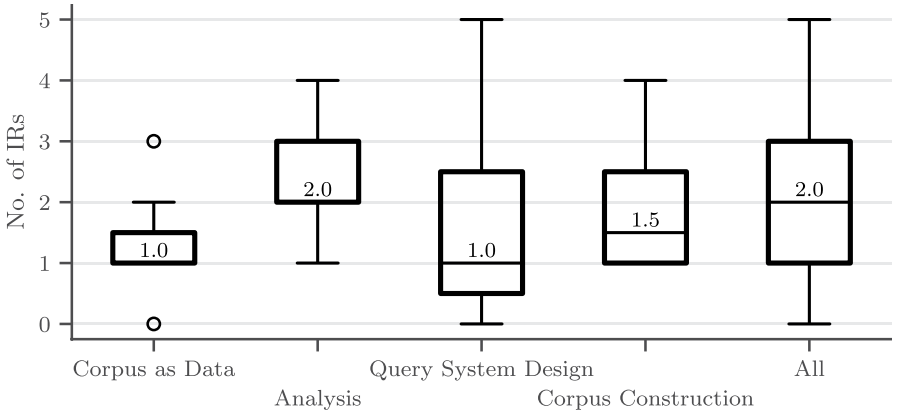

Fig. 2 Box plots for the number of information requirements per group of publications

down by the usage groups described in Section 3.2.1. Figure 3 further illustrates how the frequency of different macro-areas of information requirements introduced in Section 3.2.2 varies across the different groups of publications. The majority of the publications we surveyed mention between one and three information requirements. The most frequent information requirements are n-gram filtering, frequency plots, and statistical computations, while frequency filtering is the least common.

The numbers and types of information requirements per publication vary with the usage group of the publication. Publications using the GBNC as data have, in general, the least amount of information requirements. The most common ones in this group are n-gram filtering, statistical computations, and integration of external knowledge. Frequency plots are notably absent fom this group—which is to be expected, as frequency plots are mostly a tool for visual data exploration to analyze the GBNC itself. Publications analyzing the GBNC typically have at least two information requirements, most often frequency plots and n-gram filtering. Only a small number of these publications need frequency filtering. Publications which design a new query system usually consider between one and two information requirements. Almost half of them include n-gram filtering, while only one includes statistical computations. The other information requirements are fairly evenly distributed. Finally, publications that use the GBNC to build another corpus have between zero and four information requirements. 3 out of 4 publications need to incorporate external information, e.g., external text collections on which TNCs are constructed or specific dictionaries to restrict the GBNC to smaller subsets.

Figure 4 shows that frequency plots and n-gram filtering tend to co-occur together very often. A frequency plot can only display a limited number of time series while still being readable, so filtering is a necessary step for plotting. The focus on n-gram filtering instead of frequency filtering might be because n-gram filtering is much better supported by query systems, as detailed in Section 4. On the other hand, frequency plots and NLP are (weakly) mutually exclusive. One possible explanation for this is that most publications with frequency plots analyze the GBNC itself, which contains only fragments of natural language. Not all NLP methods are applicable to such frag-

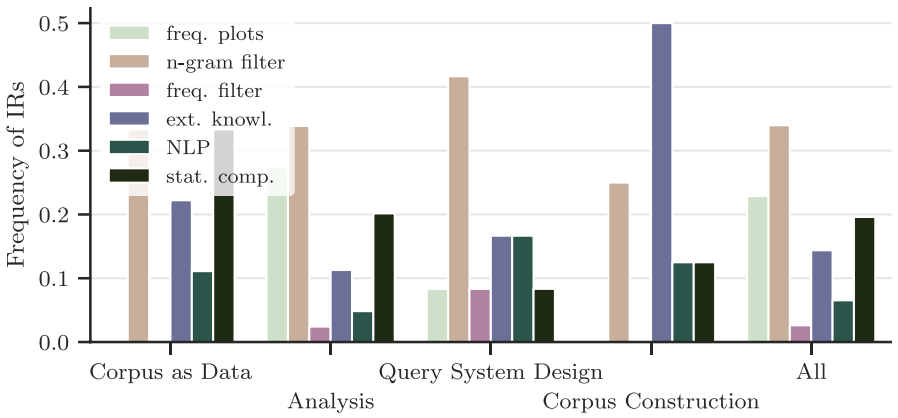

Fig. 3 Distribution of macro-areas of information requirements across the groups of publications

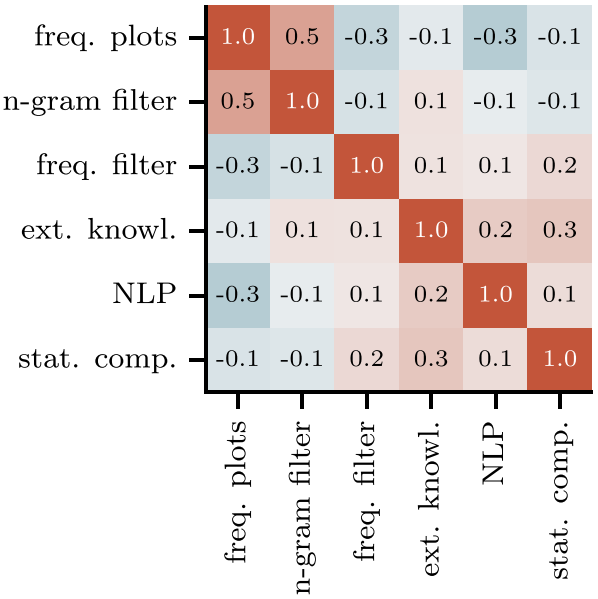

Fig. 4 Correlation between macro-areas of information requirements

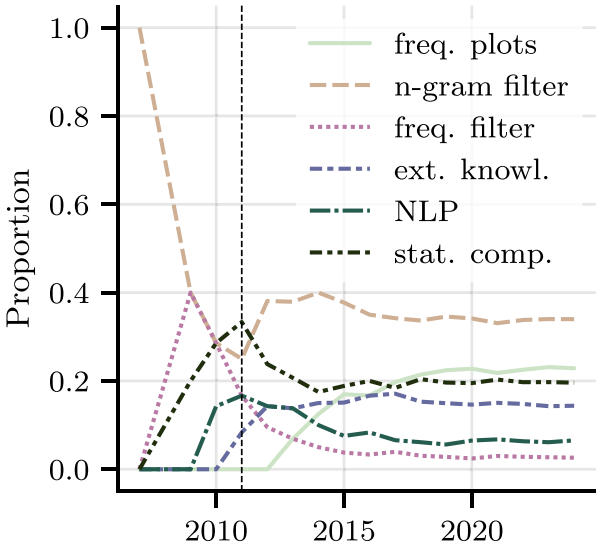

Fig. 5 Proportion of information requirements over the years. The GBNC and GBNV were published in 2011 (Michel et al., 2011)

ments. In other types of publications, a few suitable NLP methods are nonetheless applied to these language fragments, or to external texts, but these works then usually do not rely on frequency plots.

Finally, Fig. 5 shows the proportion of the different information requirements over the years. It is notable that frequency plots start to be used in 2013, after the introduction of the *Google Books Ngram Viewer* (Michel et al., 2011) in 2011. Similarly, n-gram filtering gained popularity over the years and is the information requirement mentioned most frequently since 2012. A contrary effect can be seen for frequency

filtering and statistical computations, two types of information requirements the *Google Books Ngram Viewer* does not support at all.

In the following, we detail on some of the major findings on our survey of information requirements within each of the usage groups of the GBNC. Here, publications are not grouped by their information requirements, but by their perspective on the GBNC and the specific tasks they tackle. While correlations between the two systems of categorization exist (see Fig. 3), they are, in principle, orthogonal to each other.

3.3.2 Corpus as Data

In several of the investigated works, the GBNC is used as training data for statistical models, or as a generic benchmark dataset.

Statistical models *Basile et al. (2016) use the GBNC to train vector-embedding models and to detect temporal shifts in word meanings using these models. Nazar and Renau (2012) train a frequency-based grammar checker on the GBNC data. Savinkov et al. (2021) train a neural network on the GBNC to improve part-of-speech and case tagging in Russian texts. Yu et al. (2007) use the n-gram data from a smaller corpus to derive semantic relations between words.*

Other applications *Bhowmick et al. (2015) use the GBNC as a benchmark dataset for Apache Hadoop and Apache Pig.*[4] *Bochkarev et al. (2016) use the GBNC to verify Heap's Law, an empirical observation relating the number of unique words within a text to its length.*

3.3.3 Analysis

The largest group of publications in our review considers the analysis of the GBNC itself as their subject of research.

Linguistics *Temporal n-gram corpora track the development of n-gram frequencies over time, so the detection of syntactic and semantic shifts is one of their natural applications. Galeev and Solovyev (2018) examine the popularity of different verb forms over time to detect syntactic changes. In the work of Pekina et al. (2018), the GBNC is used for probabilistic estimates of vocabulary sizes and to detect the appearance of new words in a language. Chen and Huang (2022) use co-occurrence information from the GBNC to determine semantic change of a specific Chinese word. Similarly, Gulordava and Baroni (2011) use the distributional similarity of n-grams to identify semantic changes. El-Ebshihy et al. (2018) train word embeddings on the GBNC and apply change point detection to identify 'interesting' n-grams and time series. Koplenig (2017) investigates the adverse effects that the lack of metadata within the GBNC can have on such change detection tasks, e.g., due to potential undetectable imbalances in corpus composition.*

[4] https://hadoop.apache.org, https://pig.apache.org/

Joubarne and Inkpen (2011) compare the semantic similarity of terms annotated by human experts to values automatically constructed from the GBNC. A similar effort is made by Islam et al. (2012), who compare different similarity measures based on the GBNC data.

Inkpen and Islam (2010) and Islam and Inkpen (2009) use a temporal n-gram corpus to derive spelling correction rules. Hill and Simha (2016) generate 'fill-in-the-blanks' questions from the GBNC data to assist with contextual language learning.

Lin et al. (2012) introduce syntactic annotations to the GBNC, mainly dependency parsing results. Bochkarev et al. (2018) use these to estimate the dynamic complexity of both English and Russian texts over time. Pekina et al. (2018) conduct a similar study based only on the estimated vocabulary sizes.

Radimský (2022) uses the GBNC to construct a morphology database of Romance languages, while Bochkarev et al. (2020a) use the GBNC data for Named Entity Recognition, a standard task of natural language processing.

Other scientific disciplines *Not only linguistic research is performed on the GBNC, researchers from other disciplines are interested in the data as well. Most frequently, authors from these other disciplines focus on a small set of n-grams within the GBNC, specific to their respective field and research question.*

Bukhtoyarov and Bukhtoyarova (2018) investigate the relationship between the frequency of terms related to 'revolution' to actual historical revolutions. Their study considers five different languages and a small set of 1-grams in each of these. Macedo (2013)) conducts a similar study on the history of 'eurocentrism'. Olimid et al. (2024) consider a list of n-grams and their frequencies as a proxy indicator for the degree of participation of women in political decision making. Olimid et al. (2023) apply the same methodology on terms concerned with governance styles and democracy. Juola (2013) uses the GBNC to show that culture in the USA has grown more and more complex over time. Zięba (2018) gives a general overview of the GBNC in socio-cultural research. Younes and Reips (2019) present guidelines for improving the quality of GBNC research, using religious terms as a demonstration.

Sparavigna and Marazzato (2015) and Brandt (2018) look at terms from specific disciplines, i.e., material science and geology, to draw conclusions regarding the history of the fields. Similar studies analyze the history of architecture (Bai and Huang, 2018), climatology (Grant and Walsh, 2015), and education (Moskovkin et al., 2019; Sampsel, 2021). Ribeiro and Lima (2017) investigate the popularity of certain food idioms over time. In a similar fashion, Dönmez (2018) and Zyukina et al. (2020) analyze the popularity of various human activities.

Bochkarev et al. (2020b) and Guan et al. (2024) investigate the development of color terms within the Russian and English GBNC, respectively. Bochkarev et al. (2020b) look at the prominence of two Russian terms for 'brown', while Guan et al. (2024) link the frequencies of different colors to psychological effects. Further studies on links from psychiatry and psychology to n-gram frequencies have been conducted by O'Sullivan et al. (2019) (history of psychiatry) and Teepe et al. (2023) (mental health and digitalization). Coraci et al. (2020) investigate the impact of diag-

nostic ultrasound, again by using specific n-gram frequencies, comparing the results between the GBNC and PubMed,[5] a database of publications in medicine.

Madsen and Slåtten (2022) study the evolution of different emerging technologies and how they are mentioned in literature, e.g., 'Big Data' technologies or management concepts such as 'Total Quality Management', by looking at their respective frequencies in the GBNC. Roth (2013, 2014); Roth et al. (2017) apply the same methodology to functional differentiation.

Real-world indicators *Several works attempt to relate n-gram frequencies from the GBNC to real-world data. Zeng and Greenfield (2015) use "urban population, household consumption per capita [...] and tertiary school enrollment" in China as proxy indicators for sociodemographic developments. They compute the correlation between these indicators and the frequencies of terms related to individualism and collectivism. Similar studies for Russia are conducted by Skrebyte et al. (2016) and Nestik et al. (2022), for German-speaking countries by Younes and Reips (2018), and for Germany during the times of the Third Reich by Caruana-Galizia (2016).*

Frequency analysis *Some publications analyze the n-gram frequencies themselves. Chiru and Dinu (2017) show that some n-grams have cyclic frequency patterns. Clark et al. (2022) search for n-grams with particularly strong trends in the Chinese GBNC. Friginal et al. (2014) compare the GBNC frequencies to n-gram frequencies generated from the Corpus of Historical American English (Davies, 2012), while Solovyev et al. (2019) compare the GBNC to the Russian National Corpus[6] in a similar fashion. Klein and Nelson (2009) investigate, in different TNCs, the correlation between term counts, i.e., the total number of occurrences of a term, and document frequencies, i.e., the number of documents said term appeared in.*

3.3.4 Query systems

Existing TNC query systems are discussed in more detail in Section 4. Here, we briefly discuss the works that also discuss particular information requirements and were thus part of our literature survey, which the vast majority of systems was not.

Query system implementations *The original query system for the GBNC is the Google Books Ngram Viewer[7] (GBNV) (Michel et al., 2011; Lin et al., 2012; Mann et al., 2014). It allows users to input specific n-grams and retrieve their relative frequencies as plots. Davies (2014) presents an improved version of the GBNV, allowing for more flexible wildcard searches. Another graphical search interface is presented by Schlüter and Vetter (2020).*

This interface focuses on a specific linguistic phenomenon, namely the usage of 'a' or 'an' before words beginning with the letter 'h'.

[5] https://pubmed.ncbi.nlm.nih.gov/

[6] https://ruscorpora.ru/

[7] https://books.google.com/ngrams/

Query languages *Aleksandrov and Strapparava (2012) introduce NgramQuery, a query language that connects temporal n-gram corpora to WordNet (Fellbaum, 1998; Miller, 1995). This allows users to perform queries about semantic relationships, such as retrieving synonyms of a given word. The Conceptual History Query Language (Schmidt-Petri et al., 2021; Willkomm et al., 2018) introduces several operators to represent information requirements from Conceptual History (Koselleck, 1978). Examples of this are an operator for topic grouping and an operator for searching similar frequencies.*

Foundational work *Grabowski and Swacha (2012) present a custom encoding and compression scheme for the GBNC data. The encoding reduces file sizes compared to the files published by Google, so that the GBNC can be searched more efficiently and that query execution times are reduced. Habeeb et al. (2020) describe how to unzip data from the raw corpus for further processing.*

3.3.5 Corpus construction

Alsmadi and Zarour (2018) propose extending the GBNC by including Arabic resources. Chumtong and Kaldewey (2017) construct a TNC out of scientific publications. Ivanov (2014) and Ivanov and Solovyev (2020) use the GBNC to construct corpora of semantic knowledge and relationships between words. In all of these works, external knowledge is used, usually in the form of additional texts.

3.4 Main survey findings

The main result of our survey is the categorization of information requirements. Additionally, we report the following findings regarding the research questions posed in Section 1:

Regarding **Q1**, we find that *the concrete research questions vary between fields, but the underlying, more abstract information requirements are very similar.* Nearly every publication restricts the GBNC to a specific subset of interest, usually a set of n-grams related to the respective research field, in line with the findings of Richter and Böhm (2024). The integration of external knowledge, for example dictionaries or semantic ontologies, is necessary for many different applications of the GBNC. Similarly, statistical computations on n-gram frequencies play an important role in analyzing the GBNC. A query system that allows users to specify restricted subsets of the GBNC and supports these information requirements could therefore be used by researchers across different domains.

Our survey also gives some insight into questions **Q2** and **Q3**, with a more extensive examination of query systems following in Section 4: *Most of the existing query systems that were part of the survey itself only support one or two of the information requirement categories we found.* Especially, support for the integration of external knowledge and statistical computations is rarely available, and thus requires researchers to implement custom solutions.

There is a clear connection between the query systems' capabilities and the popularity of different information requirements. One example for this is the significant under-representation of frequency filtering, which might be explained by a lack of support in most query systems.

4 Existing query systems

The survey presented in Section 3 introduces a set of information requirements that are common throughout literature. In the following, we examine existing TNC query systems with regard to their abilities of satisfying these information requirements.

Different query system for TNCs exist, a broader summary can be found in Richter et al. (2025). In that study, the capabilities of TNC query systems were evaluated from a different point of view, i.e., it is mostly descriptive in the sense that it evaluates what the systems *can* do. On the contrary, the present work in this paper focuses on the more normative aspect of what the systems *should be able to* do in order to answer research questions from literature, as given by the information requirements in Section 3.

The groups of information requriements we identified in Section 3 overlap with those found by Richter et al. (2025), but are not exactly equal. Richter et al. do not consider frequency plots as distinct information requirement, and thus do not list it in their review results. Our groups *n-gram filtering* and *frequency filtering* are contained in their *n-gram operations* and *filtering operations*, respectively. Furthermore, our more fine-grained distinction between the groups *external knowledge*, *NLP* and *statistical computations* is subsumed in their notion of *knowledge, metadata* and partially in *frequency operations*. Applying these correspondences to the results of Richter et al. (2025) yields Table 3. A checkmark in Table 3 represents the ability of a query system to satisfy *at least one* of the information requirements in the respective group.

While the internal architecture of the existing query systems differ, most systems follow a very similar usage pattern: The user specifies a short list of n-grams, poten-

Table 3 Existing TNC query systems

Query System	Information Requirements					
	freq. plots	n-gram filt.	freq. filt.	ext. knowl.	NLP	stat. comp.
Ngram Search Engine (Sekine and Dalwani, 2010)		✓	✓			
GBNV (Mann et al., 2014; Michel et al., 2011; Lin et al., 2012)	✓	✓				
NgramQuery (Aleksandrov and Strapparava, 2012)		✓		✓	✓	
PoliticalMashup Ngramviewer (de Goede et al., 2013)	✓	✓		✓		
GB-Adv (Davies, 2014)		✓				
Slash/A (Todorova et al., 2014)	✓	✓	✓			
NB N-gram (Birkenes et al., 2015)	✓	✓				
CHQL (Schmidt-Petri et al., 2021; Willkomm et al., 2018)		✓	✓	✓	✓	✓
(an:a)-lyzer (Schlüter and Vetter, 2020)						
Icelandic Gigaword n-Gram Viewer (Steingrímsson et al., 2020)	✓	✓				
Meta Trend Viewer (Indig et al., 2022)	✓	✓			(✓)*	
ParlaMint Ngram Viewer (de Jong et al., 2024)	✓	✓		✓		

*Support is planned, according to the authors

tially with some support for wildcards, the system then returns a line plot of their frequencies over time. For a more detailed view, we adopt the conceptual grouping of query systems established by Richter et al. (2025): The *GBNV-like* systems (GBNV, Ngram Search Engine, GB-Adv, Slash/A, NB N-Gram, Icelandic Gigaword n-Gram Viewer) mainly differ in their wildcard support and the underlying corpora. The *sub-corpus-oriented systems* (PoliticalMashup Ngramviewer, Meta Trend Viewer, Parla-Mint Ngram Viewer) offer an additional filtering step: Based on metadata for their specific corpora, they allow a restriction of documents, for example by author or publication source. After this step, their usage is similar to the GBNV-like systems. The three remaining systems are more diverse: The (an:a)-lyzer examines a very specific linguistic phenomenon over time, with user-controllable parameters for the distinction between *UK* and *US* English, and the selection of specific, pre-defined groups of words to plot. NgramQuery and CHQL have no available implementations, but both introduce complex query languages to express information requirements against both TNCs and external knowledge, returning datasets for further processing.

From Table 3, it becomes apparent that no single query system is able to satisfy all the information requirements from literature. The most popular tool, the *GBNV*, only supports frequency plots and n-gram filtering, as do several others. Thus, they are unable to satisfy any frequency-related information requirements. For the systems that support frequency filtering, the actual capabilities vary: The *Ngram Search Engine* supports only frequency thresholds (Sekine and Dalwani, 2010), while *Slash/A* (Todorova et al., 2014) can highlight time periods with significant fluctuations in frequency.[8] *CHQL* supports different, more complex computations, for example nearest-neighbor search for frequency similarity and aggregation over arbitrary sets of n-gram frequencies (Willkomm et al., 2018). In addition, *CHQL* also supports information requirements from the NLP group, for example sentiment analysis. The only other tool that supports NLP information requirements is *NgramQuery*, by defining operators connected to *WordNet* (Fellbaum, 1998; Miller, 1995). The other systems differ mostly within individual groups of information requirements, for example with regards to their flexibility in wildcard usage or the exact frequency filtering predicates they allow.

Of the existing query systems, *CHQL* supports the most categories of information requirements, offering operators for every category except frequency plots. Another system notable for its flexible query language is *NgramQuery*: While it supports only three categories of information requirements, its NLP operators are more powerful than those of the other systems. However, neither of the two systems has a publicly available implementation (Richter et al., 2025), so researchers cannot actually use the systems for their studies or extend them for specific use cases. The other available systems, on the other hand, lack the required flexibility to express complex research questions from literature, as shown by the lack of support for different groups of information requirements.

[8] http://linguistics.chrisculy.net/lx/vistola/tools/slasha.html

5 Conclusions

5.1 Summary

Our work was motivated by three research questions, we summarize our answers as follows:

- **Q1**: What kind of (implicit or explicit) information requirements do researchers have when working with the GBNC?
  - The information requirements are diverse, but fall into six main groups: *frequency plots* (in 35 publications), *n-gram filtering* (52), *frequency filtering* (4), *external knowledge* (22), *NLP* (10), and *statistical computations* (30).
- **Q2**: Are existing query systems for the GBNC powerful enough to satisfy these information requirements?
  - While most groups of information requirements are supported by at least some of the systems, no system supports them all. The most powerful systems have no available implementations.
- **Q3**: If not, what is missing in these query systems?
  - Most existing tools offer n-gram filtering and frequency plots, but support for more complex analyses or the inclusion of external knowledge is very limited.

We answered **Q1** with an extensive literature review in Section 3. We identified four categories of research works, according to their usage of TNCs: *corpus as data*, *analysis*, *query systems*, and *corpus construction*. Across these paper categories, we extracted six main groups of information requirements, namely *frequency plots*, *n-gram filtering*, *frequency filtering*, *external knowledge*, *NLP*, and *statistical computations*. We then discussed these categories' individual contributions. We have shown that information requirements are mostly domain-independent, so a TNC query system supporting these information requirements can be useful for researchers from many different fields.

To answer **Q2** and **Q3**, we researched additional existing query systems for TNCs and compared their capabilities to the information requirements we identified. Notably, most query systems are limited to plotting the frequencies of specific n-grams, with varying levels of wildcard support for n-gram filtering. None of them satisfy all of the information requirements. The two that come the closest, namely *Ngram-Query* (Aleksandrov and Strapparava, 2012) and *CHQL* (Willkomm et al., 2018; Schmidt-Petri et al., 2021), have no available implementations.

5.2 Future work

From the answers we found to our research questions, one direction of future work is immediately apparent: A flexible, domain-independent TNC query system is theoretically possible, but does not exist yet. Publishing such a system could benefit researchers from many different disciplines.

Another direction of future work is the extension of the survey itself. As discussed in Section 3.1.1, we employed a relatively restrictive, title-based filter to publications. A broader survey may be possible in the future, e.g., by using *citation intent classification* (Cohan et al., 2019) instead of a pure keyword selection. Furthermore, the restriction to publications written in English potentially skews the findings towards Anglophone perspectives, making the inclusion of other languages a valuable extension of our work.

Author Contributions F.R. and F.M. jointly conducted the literature survey. F.R. wrote the main manuscript, F.M. prepared the figures. P.R. and K.B. reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Survey data and code to reproduce figures can be found at https://github.com/DrCohomology/GBNC_Survey.

Declarations

Competing interests The authors declare no competing interests.

References

Abiteboul, S. (1997). Querying semi-structured data. In: Proceedings of The 6th international conference on database theory, (pp 1–18). Springer,

Aleksandrov, M., & Strapparava, C. (2012). NgramQuery—Smart Information Extraction from Google N-gram using External Resources. In: LREC, pp 563–568

Alsmadi, I., & Zarour, M. (2018). Google N-Gram Viewer does not Include Arabic Corpus! Towards N-Gram Viewer for Arabic Corpus. *International Arab Journal of Information Technology, 15*(5), 785–794.

Bai, N., Huang, W. (2018). Quantitative analysis on architects using culturomics. *In: Proceedings of the 23rd International Conference of the Association for Computer-Aided Architectural Design Research in Asia* (pp. 257–266)

Basile, P., Caputo, A., Luisi, R., et al. (2016). Diachronic analysis of the Italian language exploiting Google Ngram. CLiC it (pp. 56–60)

Bhowmick, S., Chakraborty, S., & Agrawal, D. P. (2015). Study of Hadoop-MapReduce on google N-Gram datasets. *In: Proceedings of the 2015 IEEE 12th international conference on mobile ad hoc and sensor systems* (pp. 488–490). IEEE

Birkenes, M.B., Johnsen, L. G., Lindstad, A. M., et al. (2015). From digital library to n-grams: NB N-gram. *In: Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)* (pp. 293–295)

Bochkarev, V. V., Khristoforov, S. V., Shevlyakova, A. V. (2020a). Recognition of named entities in the Russian subcorpus Google Books Ngram. *In: Proceedings of the Mexican international conference on artificial intelligence* (pp. 17–28). Springer

Bochkarev, V. V., Lerner, E. Y., & Shevlyakova, A. V. (2016). Verifying heaps' law using google books ngram data. arXiv preprint arXiv:1612.09213

Bochkarev, V. V., Shevlyakovaa, A. V., Parameib, G. V., et al. (2020b). A quantitative study of Russian colour terms buryj and koričnevyj in the Google Books Ngram corpus. *In: Proceedings of the linguistic forum 2020: Language and artificial intelligence* (pp. 1–10)

Bochkarev, V. V., Solovyev, V. D., & Shevlyakova, A. V. (2018). Analysis of dynamics of the number of syntactic dependencies in Russian and English using Google Books Ngram Академия наук Республики Татарстанр 18

Brandt, D. S. (2018). Charting the geosciences with Google Ngram Viewer. *GSA Today, 5*, 66–67.

Brants, T., & Franz, A. (2006). Web 1t 5-gram version 1. https://catalog.ldc.upenn.edu/LDC2006T13

Breit, F. (2017). The distribution of english isograms in google ngrams and the British national corpus. Opticon1826

Bukhtoyarov, M., & Bukhtoyarova, A. (2018). A rough quarter of the millennium. *Revolutions through the lens of google ngram viewer. In: Information technologies in the humanities* (pp. 48–61)

Buneman, P. (1997). Semistructured data. *In: Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems* (pp. 117–121)

Caruana-Galizia, P. (2016). Politics and the German language: Testing Orwell's hypothesis using the Google N-Gram corpus. *Digital Scholarship in the Humanities, 31*(3), 441–456.

Chen, J., & Huang, C. R. (2022). From frying to speculating: Google Ngram evidence to the meaning development of '炒'in Mandarin Chinese. *In: Proceedings of the 36th pacific Asia conference on language, information and computation* (pp. 425–429)

Chiru, C. G., & Dinu, V. N. (2017). Identifying cyclic words with the help of google books n-grams corpus. *Proceedings of the 12th international conference on internet and web applications and services* (pp. 34–39)

Chumtong, J., & Kaldewey, D. (2017). Beyond the google ngram viewer. *Forum Internationale Wissenschaft*

Clark, C., Zhang, L., & Roth, S. (2022). What's trending in the chinese google books corpus. *Global Debates in the Digital Humanities* (pp. 151–169)

Cohan, A., Ammar, W., Van Zuylen, M., et al. (2019). Structural scaffolds for citation intent classification in scientific publications. arXiv preprint arXiv:1904.01608

Coole, M., Rayson, P., & Mariani, J. (2016). LexiDB: A scalable corpus database management system. *In: Proceedings of The 2016 IEEE International Conference on Big Data (Big Data)* (pp. 3880–3884). IEEE

Coole, M., Rayson, P., & Mariani, J. (2020). LexiDB: Patterns & methods for corpus linguistic database management. *In: Proceedings of the 12th language resources and evaluation conference* (pp. 3128–3135)

Coraci, D., Loreti, C., Glorioso, D., et al. (2020). The impact of diagnostic ultrasound in clinical medicine and in nerve evaluation: pubmed and google ngram viewer compared. *Neurophysiologie clinique = Clinical neurophysiology 50*(4), 305–307

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora, 7*(2), 121–157.

Davies, M. (2014). Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics, 19*(3), 401–416. https://doi.org/10.1075/ijcl.19.3.04dav

de Goede, B., van Wees, J., Marx, M., et al. (2013). PoliticalMashup Ngramviewer: Tracking who said what and when in parliament. *In: Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3* (pp. 446–449). Springer

de Jong, A., Kuzman, T., Larooij, M., et al. (2024). Parlamint ngram viewer: Multilingual comparative diachronic search across 26 parliaments. *In: Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024* (pp 110–115)

Dönmez, I. (2018). Human activity analysis and prediction using google n-Grams. *International Journal of Future Computer and Communication* 7(2)

El-Ebshihy, A., El-Makky, N. M., & Nagi, K. (2018). Using google books ngram in detecting linguistic shifts over time. *In: Proceedings of the 9th international conference on knowledge discovery and information retrieval* (pp. 330–337)

Fellbaum, C. (1998). WordNet: An electronic lexical database. MIT press

Friginal, E., Walker, M., & Randall, J. B. (2014). Exploring mega-corpora: Google Ngram viewer and the corpus of historical American English. *EuroAmerican Journal of Applied Linguistics and Languages, 1*(1), 48–68.

Galeev, T., Solovyev, V. (2018). Google Books Ngram as an instrument of teaching foreign language. *In: Proceedings of the 1st indonesian communication forum of teacher training and education faculty leaders international conference on education 2017 (ICE 2017)* (pp. 616–619)

Grabowski, S., & Swacha, J. (2012). Google books ngrams recompressed and searchable. *Foundations of Computing and Decision Sciences, 37*(4), 271–281. https://doi.org/10.2478/v10209-011-0015-8

Grant, W. J., & Walsh, E. (2015). Social evidence of a changing climate: Google Ngram data points to early climate change impact on human society. *Weather, 70*(7), 195–197.

Guan, L., Shi, W., Li, Q., et al. (2024). Have color representations in books changed over the past 200 years? An empirical analysis based on the Google Books Ngram corpus. *Color Research & Application, 49*(1), 65–78.

Gulordava K, Baroni M (2011) A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *In: Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics* (pp. 67–71)

Habeeb, I. Q., Habeeb, Z. Q., Jurn, Y. N., et al. (2020). Constructing Arabic language resources from Google N-gram dataset. *In: Journal of Physics: Conference Series* (Vol. 1530, pp. 1–7). IOP Publishing

Hill, J., & Simha, R. (2016) Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. *In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 23–30)

Indig, B., Sárközi-Lindner, Z., & Nagy, M. (2022). Use the metadata, Luke!–an experimental joint metadata search and n-gram trend viewer for personal web archives. *In: Proceedings of the 2nd international workshop on natural language processing for digital humanities* (pp 47–52)

Inkpen, D., & Islam, A. (2010). Unsupervised approaches to text correction using google n-grams for english and romanian. *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian* (pp 270–285)

Islam, A., & Inkpen, D. (2009). Real-word spelling correction using Google Web 1T n-gram with backoff. *In: Proceedings of the 2009 international conference on natural language processing and knowledge engineering* (pp 1–8). IEEE

Islam, A., Milios, E., & Keselj, V. (2012). Comparing word relatedness measures based on Google n-grams. *In: Proceedings of COLING 2012: Posters* (pp. 495–506)

Ivanov, V. (2014). Extracting frame-like structures from google books ngram dataset. *In: Human-Inspired Computing and Its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I 13* (pp. 18–27). Springer

Ivanov, V., & Solovyev, V. (2020). Ranking concrete and abstract words using google books ngram data. *Journal of Intelligent & Fuzzy Systems, 39*(2), 2229–2237.

Joubarne, C., & Inkpen, D. (2011). Comparison of Semantic Similarity for Different Languages Using the Google n-gram Corpus and Second-Order Co-occurrence Measures. SpringerIn C. Butz & P. Lingras (Eds.), *Advances in Artificial Intelligence* (Vol. 6657, pp. 216–221). Berlin, Heidelberg: Berlin Heidelberg.

Juola, P. (2013). Using the Google N-Gram corpus to measure cultural complexity. *Literary and Linguistic Computing, 28*(4), 668–675.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University, 33*(2004), 1–26.

Klein, M., Nelson, M. L., et al. (2009). Correlation of Term Count and Document Frequency for Google N-Grams. SpringerIn M. Boughanem, C. Berrut, & J. Mothe (Eds.), *Advances in Information Retrieval* (Vol. 5478, pp. 620–627). Berlin, Heidelberg: Berlin Heidelberg.

Koplenig, A. (2017). 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, 32*(1), 169–188.

Koselleck R (1978) Historische semantik und begriffsgeschichte (Vol. 1). Klett-Cotta
Kreutz, C. K., Wolz, M., Weyers, B., et al. (2020). SchenQL: Evaluation of a query language for bibliographic metadata. *In: Proceedings of The 22nd International Conference on Asian Digital Libraries* (pp 323–339). Springer
Kreutz, C. K., Wolz, M., Knack, J., et al. (2022). SchenQL: in-depth analysis of a query language for bibliographic metadata. *International Journal on Digital Libraries, 23*(2), 113–132.
Lin, Y., Michel, J. B., Lieberman, E. A., et al. (2012). Syntactic annotations for the Google Books Ngram Corpus. In: Proceedings of the ACL 2012 system demonstrations, pp 169–174
Macedo, A. L. N. (2013). The history of eurocentrism throughout the google ngram platform. *In: Convergências: estudos em Humanidades Digitais*
Madsen, D. O., & Slåtten, K. (2022). The possibilities and limitations of using Google Books Ngram Viewer in research on management fashions. *Societies, 12*(6), 171.
Mann J, Zhang D, Yang L, et al (2014) Enhanced search with wildcards and morphological inflections in the Google Books Ngram Viewer. *In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 115–120)
Michel, J. B., Shen, Y. K., Aiden, A. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.
Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.
Moher, D., Liberati, A., Tetzlaff, J., et al. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International Journal of Surgery, 8*(5), 336–341.
Moskovkin, V., Saprykina, T., Pupynina, E., et al. (2019). Examination of trends in education with the Google Books Ngram Viewer. *International Journal of Engineering and Advanced Technology (IJEAT)*
Nazar, R., & Renau, I. (2012). Google books n-gram corpus used as a grammar checker. *In: Proceedings of the Second Workshop on Computational Linguistics and Writing (CL&W 2012)*, (pp. 27–34)
Nestik, T., Bochkarev, V., & Levina, V. (2022). Dynamics of the long-term orientation in russian society over the past 100 years: Results of the analysis of the russian subcorpus of google books ngram. *In: International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies* (pp. 126–136). Springer
Olimid, A. P., Georgescu, C. M., & Gherghe, C. L. (2023) Integrated Analysis of Sixty Democracy Governance and Policy Reform Topics using Ngram Tool for Google Platform (1990-2019). Revista de Stiinte Politice (78)
Olimid, A. P., Georgescu, C. M., & Gherghe, C. L. (2024). Societal security, participation and women's representation in political history. A conceptual and graphical analysis using data collection methods in google ngram viewer. Revista de Ştiinţe Politice Revue des Sciences Politiques• No 81:246–258
O'Sullivan, O., Duffy, R., & Kelly, B. (2019). Culturomics and the history of psychiatry: Testing the Google Ngram method. *Irish Journal of Psychological Medicine, 36*(1), 23–27.
Owei, V. (2003). Development of a conceptual query language: adopting the user-centered methodology. *The Computer Journal, 46*(6), 602–624.
Page, M. J., McKenzie, J. E. , Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. bmj 372
Pekina, A., Maslennikova, Y., & Bochkarev, V. (2018). Probability analysis of the vocabulary size dynamics using Google Books Ngram Corpus. *In: Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST 2018)* (pp. 202–207)
Perrie, J., Islam, A., Milios, E., et al. (2013). Using google n-grams to expand word-emotion association lexicon. In: Hutchison, D., Kanade, T., Kittler, J., et al (Eds.) *Computational Linguistics and Intelligent Text Processing* (Vol. 7817, pp. 137–14). Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37256-8_12
Radimskỳ, J. (2022). Towards a diachronic analysis of Romance morphology through Google n-grams. *Corpus* (23)
Ribeiro, S. V., & Lima, P. L. (2017). Google N-grams Viewer and Food Idioms. *EUROPHRAS*
Richey, S., & Taylor, J. B. (2020). Google Books Ngrams and political science: Two validity tests for a novel data source. *PS: Political Science & Politics 53*(1), 72–77
Richter, F., Böhm, K. (2024). A workflow for efficient and interactive analysis of the Google Books Ngram Corpus. *In: Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*
Richter, F., Schäfer, B., & Böhm, K. (2025). A review of query systems for temporal n-gram corpora. *In: Proceedings of the First International Workshop on Scholarly Information Access*

Roth, S., Clark, C., & Berkel, J. (2017). The fashionable functions reloaded: an updated Google Ngram view of trends in functional differentiation (1800-2000). *In: Research paradigms and contemporary perspectives on human-technology interaction* (pp. 236–265). IGI Global

Roth, S. (2013). The fairly good economy: testing the economization of society hypothesis against a Google Ngram view of trends in functional differentiation (1800–2000). *Journal of Applied Business Research, 29*(5), 1495–1500.

Roth, S. (2014). Fashionable functions: A Google Ngram view of trends in functional differentiation (1800–2000). *International Journal of Technology and Human Interaction (IJTHI), 10*(2), 35–58.

Ruest, N., Fritz, S., Deschamps, R., et al. (2021). From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities, 2*(1), 5–24.

Sampsel, L. J. (2021). Teaching the Google Books Ngram Viewer and JSTOR Text Analyzer in the Graduate Music Bibliography Course: Benefits, Issues, and Challenges. *Notes, 77*(4), 539–560.

Savinkov AV, Bochkarev VV, Shevlyakova AV, et al (2021) Neural Network Recognition of Russian Noun and Adjective Cases in the Google Books Ngram Corpus. *In: Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23* (pp 626–637). Springer

Schlüter, J., & Vetter, F. (2020). An interactive visualization of Google Books Ngrams with R and Shiny: Exploring a (n) historical increase in onset strength in a (n) huge database. *Journal of Data Mining & Digital Humanities*

Schmidt-Petri, C., Schäler, M., Schefczyk, M., et al. (2021). The CHQL query language for conceptual history using google books ngrams. *In: Data for History*

Sekine S, Dalwani K (2010) Ngram search engine with patterns combining token, POS, chunk and NE information. In: LREC

Skrebyte, A., Garnett, P., & Kendal, J. R. (2016). Temporal relationships between individualism-collectivism and the economy in Soviet Russia: A word frequency analysis using the Google Ngram corpus. *Journal of Cross-Cultural Psychology, 47*(9), 1217–1235.

Solovyev, V. D., Bochkarev, V. V., & Akhtyamova, S. S. (2019). Google books ngram: Problems of representativeness and data reliability. *In: International Conference on Data Analytics and Management in Data Intensive Domains* (pp. 147–162). Springer

Sparavigna, A. C., & Marazzato, R. (2015). Using google ngram viewer for scientific referencing and history of science. arXiv preprint arXiv:1512.01364

Steingrímsson, S., Barkarson, S., Örnólfsson, G. T. (2020). Facilitating corpus usage: Making icelandic corpora more accessible for researchers and language users. *In: Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3399–3405)

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & research libraries, 29*(3), 178–194.

Teepe, G. W., Glase, E. M., & Reips, U. D. (2023). Increasing digitalization is associated with anxiety and depression: A Google Ngram analysis. *Plos one,18*(4), Article e0284091.

Todorova, V., Chinkina, M., & de Haan, R. (2014). Slash/a n-gram tendency viewer—visual exploration of n-gram frequencies in correspondence corpora. *In: Proc. of the ESSLLI* (pp. 229–239)

Willkomm, J., Schmidt-Petri, C., Schäler, M., et al. (2018). A query algebra for temporal text corpora. *In: Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries* (pp. 183–192)

Wilson, T. D. (1981). On user studies and information needs. *Journal of documentation, 37*(1), 3–15.

Younes, N., & Reips, U. D. (2019). Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. *PloS one,14*(3), Article e0213554.

Younes, N., & Reips, U. D. (2018). The changing psychology of culture in German-speaking countries: A Google Ngram study. *International Journal of Psychology, 53*, 53–62.

Yu, L. C., Wu, C. H., Philpot, A., et al. (2007). OntoNotes: sense pool verification using Google N-gram and statistical tests. *In: Proceedings of OntoLex Workshop* Citeseer

Zeng, R., & Greenfield, P. M. (2015). Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. *International Journal of Psychology, 50*(1), 47–55.

Zięba, A. (2018). Google Books Ngram Viewer in socio-cultural research. *Research in Language (RiL), 16*(3), 357–376.

Zyukina, Z., Voropaeva, Y., & Zyukina, Z. (2020). Intellectual games concept review in THE XIX–XXI century (Google book Ngram Corpus scientific materials base). *In: E3S Web of Conferences* (Vol. 210, p 16035). EDP Sciences