informationresearch

PUBLISHED QUARTERLY BY THE UNIVERSITY OF BORÅS, SWEDEN

VOL. 27 NO. 1, MARCH, 2022

Investigating the extent to which words or phrases with specific attributes can be retrieved from digital text collections

Liezl H. Ball, and Theo J.D. Bothma

Introduction. *Digital text collections are increasingly being used. Various tools have been developed to allow researchers to explore such collections. Enhanced retrieval will be possible if texts are encoded with granular metadata.*

Method*. A selection of tools used to explore digital text collections was evaluated to determine to what extent they allow for the retrieval of words or phrases with specific attributes.*

Analysis*. Tools were evaluated according to the metadata that are available in the data, the search options in the tool, how the results are displayed, and the expertise required to use the tool.*

Results*. Many tools with powerful functions have been developed. However, there are limitations. It is not possible to search according to semantics or in-text bibliographic metadata. Analysis of the tools revealed that there are limited options to combine multiple levels of metadata and typically, without some programming expertise or knowledge of the structure and encoding of data, researchers cannot currently retrieve words or phrases with specific attributes from digital text collections.*

Conclusions*. Granular metadata should be identified, and tools that can utilise these metadata to enable the retrieval of words or phrases with specific attributes in an intuitive manner should be developed.*

DOI: https://doi.org/10.47989/irpaper917

Introduction

There are many digital text collections in the world. A well-known example of a large collection of scanned texts is Google Books. Libraries also have, or are creating,

digital collections; consider for example the digital collections held by the Library of Congress. Though some text collections may not be the same size as datasets used in the natural sciences, some collections offer sufficiently large datasets so that researchers from humanities are seen as dealing with big data (Howard, 2017). Because of the rate at which content is generated and materials are digitised, some digital text collections are very large. For example, in 2019, the Google Books collection contained more than 40 million records (Lee, 2019).

Through the use of technology, digital collections can be explored and analysed in ways not possible in a paper-based environment, for example, using computational methods. Computational methods include '*counting [words], looking at their distribution within a text or seeing how they are juxtaposed with other words*' (Dombrowski, 2020). Computational analysis of textual data in collections is becoming increasingly common and are allowing researchers to explore challenging questions in new ways (Nguyen, et al., 2020).

New technologies and techniques to study digital text collections are not limited to programmers; some tools are available that allow researchers with little or no programming experience to study digital text collections. An example of such a tool is the Google Books Ngram Viewer. This is an interactive online tool, which can be used to see how often words were used during a specific period of time. An example of a search in the Ngram Viewer is shown in Figure 1, where the frequency of the words *carriage, coach, chaise, buggy* and *cab* during the period 1850 to 1950 in English fiction is shown in a graph. Interesting studies have been done using this tool. For example, Michel, et al. (2011) notably used it to study linguistic changes, specifically some changes in lexicon and grammar, and subsequently noted cultural trends, specifically the collective memory of people and events, suggesting that the duration of fame is decreasing. Other studies using this tool have been done by, for example Acerbi, et al. (2013), Keuleers, et al. (2011) Li, et al. (2020) and Ophir (2016).

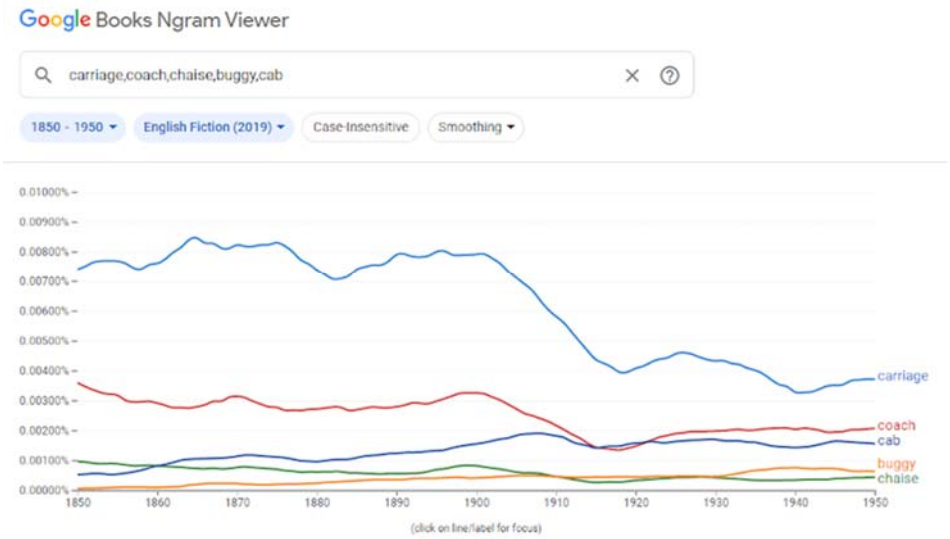


Figure 1: Searching for the words *carriage, coach, chaise, buggy* and *cab* in the Google Books Ngram Viewer

Despite the options that tools such as this one offer, there is a concern that researchers are not sufficiently able to narrow down the collection or dataset they would like to investigate. In other words, to what extent are researchers able to explore (e.g., count words, find examples, view trends) texts, sections of texts or even words that fit the requirements of the researcher?

For example, is it possible to retrieve instances of a word that only appears in direct speech, or to distinguish between homographs, or was written by a certain author, or is in a quotation? Furthermore, how easy is it to perform such an advanced search? Does a user need to understand the structure of the underlying data and use complex commands? This leads to the main problem investigated by this research, namely, to what extent are users able to search for words with specific attributes in a text collection?

In an attempt to address this particular problem, this paper is structured in the following manner. The next section considers concepts and related work that are relevant to this topic, after which the objectives of this paper are discussed. The following section describes the data collected through the evaluation of several tools. In the subsequent section the data are analysed, and the research questions are answered. Lastly, the authors present their conclusion and future work.

Related work

The ability to search in collections and retrieve relevant items is by no means a novel idea. Many tools allow a user to search in large collections of information sources using keywords. Consider for example the widespread use of search engines to search

for information. Haider and Sundin (2019: 1) write that '*search, searching, and with them search engines have become so widely used that we have stopped noticing them*'. Other search platforms are also widely used. Consider platforms to search in databases, such as EbscoHost or ProQuest. Digital libraries also have search options, consider for example HathiTrust, a large digital library, that allows a user to search in its collections.

Various search features are available to allow users to search effectively in digital collections. It is often possible to search according to bibliographic details of information sources. Though the specific options per search service will differ, some examples include searching for an author, searching sources published in a certain time period or searching for a certain type of publication. Advanced search options also allow for searching for a combination of words to appear in a resource or to be excluded from a resource. Various technologies and techniques, such as aspects from Artificial Intelligence, are employed in web and mobile searching with the aim to return relevant results to a user (Cox, 2021).

The focus of this paper will not be on searching in search engines or databases, but on the ability to search for instances of words with specific attributes in a text collection. These attributes are additional information about the texts, sections of texts or words in a text. It could include various types of information, ranging from the author of a quotation in a text to the meaning of a word. Such additional information can be captured in metadata. These metadata could facilitate discovery. This was already seen in Figure 1, where the ability to search according to a date range, is enabled through limited bibliographic metadata, namely a date element. More complex searches in the Google Books Ngram viewer are also possible, as illustrated in the evaluation of this tool. Metadata, and specifically granular metadata, are required to enable such search.

According to Zeng and Qin (2016: 11) '*metadata encapsulate the information that describes any information-bearing entity*'. Metadata can be seen as what can be said about an information object at different levels of aggregation (Gilliland, 2016). One of the important functions of metadata is to facilitate search and discovery (Gilliland, 2016; Haynes, 2018). The purpose of library metadata has been predominantly about providing access to information resources and '*includes indexes, abstracts, and bibliographic records created according to cataloging rules*' (Gilliland, 2016). However, the concept of metadata has been used broadly with slightly different meanings in different disciplines. For example, in the field of natural language processing, annotations (e.g., part-of-speech tags) have been described as '*metadata that provides additional information about the text*' (Pustejovsky and Stubbs, 2012). Zeng and Qin (2016) have also said that '*metadata exist not only in the traditional bibliographic data universe*'. In this

paper the term metadata will be used broadly and seen as additional information about an entity, from the document-level to the word-level. Furthermore, the paper will use the term granular metadata to refer to the detailed attributes of words and texts that are captured (made explicit) to enable retrieval. The purpose of this additional information, as seen by this study, is to facilitate the discovery of words with certain attributes.

One of the main points of criticism of large digital text collections is the absence of (high quality) metadata (e.g., Henry and Smith, 2010; Koplenig, 2017; Nunberg, 2009). In the first place this means that no (or no reliable) filtering can be done. If a user filters a search to search for a word only in works of fiction (intending to exclude other genres), but genre metadata are absent or not reliable, then the results cannot be trusted. Furthermore, unreliable metadata will mean that the composition of the data is not clear (Koplenig, 2017). In large collections, the focus is typically on quantity, which means the composition of the data can be unbalanced (Henry and Smith, 2010).

However, even if high quality bibliographic metadata were available, it seems that there is a need to have more detailed information, for example information about sections in a book. A study was conducted by Fenlon, et al. (2014) to determine the needs of scholars when working with text collections and it was found that scholars require '*more and better metadata that transcend the conventions of the bibliographic record*' (Fenlon, et al., 2014, p. 9). The need for granular metadata was explained by one of the participants in this study who said that often researchers are not necessarily interested in doing research about a book, but about items in the book, such as poems quoted in the book. These items are typically not stated in the bibliographic metadata. It is clear that bibliographic metadata can be too coarse for the needs of some scholars and that granular(bibliographic) metadata could be beneficial. Detailed metadata that could be used to enhance texts were noted by Fenlon, et al. (2014: 7) and include items such as chapter, genre, page, theme, word, document format, entities within text, languages, publication data (Fenlon, et al., 2014: 7). The need for granular metadata has been supported by others, for example Underwood (2015), who has done research on automatically identifying genre, not only on volume level, but also on a page level. The HathiTrust Research Center explains the necessity to '*create a layer of metadata objects that describe finer-grained resources so that scholars can identify them and make use of them in their analysis*' (Jett, et al., 2016, p. 36). This idea is supported by the HathiTrust Research Center by providing an Extracted Features Dataset (Capitanu, et al., 2016). This dataset contains useful metadata about the texts, even at page level.

An example will be given to demonstrate the benefit of granular metadata for retrieval. Assume there is a large text collection, and a user is interested in certain

instances of words in this collection. There are certain requirements for these instances to be retrieved, or in other words, the words must have certain metadata. For example, assume the user is searching for all instances of the word *well*, where it is used as a noun, appears in direct speech, in works published after the year 1800. To answer this question, there must be some information (or metadata) about the words and the texts in which they appear. In this example, it is necessary to have grammatical information (to know the part-of-speech category of the word), structural information (to know if it appears in direct speech) and bibliographic information (to know when the works in which the instances appear were published).

The quality of tools available to engage with collections is important. If extra, granular metadata are available, the data to be queried will be more complex. It has been suggested that powerful, advanced tools are required to work with complex datasets (Hardie, 2012). Lansdall-Welfare and Cristianini (2020) argue that researchers in the field of digital humanities would greatly benefit from tools that enable them to study large digital text collections more easily. Nyhan, et al. (2020) also reflect on their experiences when using university-based high-performance computing infrastructures as opposed to using external, cloud-hosted tools to mine large-scale, digitised text collections. They suggest that though the needs of the scholars in digital humanities have become more complex, the infrastructure to support computationally intensive work has not evolved rapidly. It has also been noted that some datasets are only accessible to those with programming expertise (Bode, 2017).

The idea that users should be able to specify exactly what words or phrases should be retrieved from digital text collections was investigated in a research project conducted by the authors of this paper. Specifically, the use of granular metadata to enhance retrieval from text digital collections was investigated. A comprehensive discussion of the research project is available (Ball, 2020). As part of this project, Ball (2020) investigated the extent to which current tools can be used to retrieve words or phrases with specific attributes from digital text collections. The purpose of this paper is to discuss the findings of this evaluation.

A poster presentation in which the premise of this study was explained, was presented at the 2020 ISIC (Information Seeking in Context) conference (Ball and Bothma, 2020). After the feedback was received, the authors expanded and refined the ideas and present a more comprehensive discussion in this paper.

Objectives

Clearly there is a need for tools that enable scholars to filter and extract very specific instances from digital text collections. This has led to the following questions:

1. What granular metadata will be useful for retrieval from digital text collections on a detailed level?
2. How do current tools support the retrieval of words or phrases with specific attributes from digital text collections?
3. What recommendations can be made for the development of a tool that enables retrieval of words and phrases with specific attributes?

These questions form the basis of the first part of the research project referred to in the previous section.

Method

A grounded theory approach was used in this study, as the data collection and data analysis occurred simultaneously (Pickard, 2017) and based on the analysis of the data certain categories could be developed (Leedy, et al., 2021). By examining the literature and evaluating various tools used to search in text collections, certain granular metadata that could be useful for retrieval were identified and categories for these metadata were developed. A heuristic evaluation of current tools allowed the researchers to see to what extent retrieval of words with specific attributes can be retrieved. A heuristic evaluation is a method in which an interface is evaluated systematically according to a set of principles (Shneiderman and Plaisant, 2010).

There are numerous tools available that allow a user to engage with a digital text collection. Different tools include different search functions that are available to search in text collections. Some have advanced options to filter according to bibliographic metadata, some allow searching according to linguistic metadata. Purposive sampling was used to select the tools that were evaluated in this study. The selected tools can be seen as representing certain search functionality that is found in various tools. The tools are typical of a group of tools or demonstrate unique advanced retrieval options that are pertinent to this study. In the study by Ball (2020), seven tools that allow a user to retrieve words or phrases from a collection were evaluated. Five of these tools will be discussed in this paper, namely, Google Books Ngram Viewer, HathiTrust+Bookworm, Perseus Project, TXM and BNCweb (CQP-Edition). The other two, Voyant Tools and BYU Corpora (now English-Corpora.org) are excluded from this paper as the data from their evaluations do not have significantly different features.

In the next section, the granular metadata that were identified through examining literature and tools will be listed; thereafter, the evaluation of the tools will be discussed. The data collected through this evaluation will enable the researchers to answer the questions posed in this study, which will be done in the subsequent section.

Granular metadata identified in tools and literature

By examining the search functionalities of various tools as well as literature (e.g., Fenlon, et al., 2014; Finlayson, 2015; Lin, et al., 2012; Underwood, 2015), possible metadata useful for retrieval were identified. A variation of the citation pearl-growing strategy was used to find relevant literature and tools to examine. Various tools that are used to explore text collections were identified. These tools were examined and a search for literature discussing these tools was done. Various databases were used to select literature about these tools. Through this a comprehensive set of attributes was identified.

The identified metadata are listed in Table 1. As these metadata are used in the evaluation, they are listed here, but they are discussed in more detail later in the paper.

Table 1: Granular metadata that could be useful in retrieval

Granular metadata categories	Items in category	Explanation
Morphological	Inflected forms (lemma)	Inflection refers to the process where words change in form to denote grammatical distinctions, for example *buy*, *bought*.
	Part-of-speech category	These categories refer to types of words, such as nouns, verbs, adjectives.
Syntactic	Dependency between words	Words in a sentence can be linked to other words, for example the object of the sentence is linked to the verb in the sentence.
Semantic	Meaning	This refers to the different meanings that a word may have. For example, *bank* that refers to a financial institution or the side of a river.
Functional	Logical features	A text may have certain features that become apparent on analysis of the text, for example names, dates or quotations.
	Structural features	Texts are often divided into parts, such as chapters, paragraphs, front matter and back matter.
Bibliographical	Bibliographic detail	These are the metadata that identify the main text or volume, such as title, author, publisher.
	In-text bibliographic detail	Sections in a volume can have different bibliographic metadata than that of the volume, for example a quotation.

These metadata were used in the evaluation of the tools.

Heuristic evaluation of tools

The purpose of this discussion is not to be a comprehensive review of each tool, but to consider the extent to which retrieval of words on a detailed level is possible. As such, there are features of each tool that will not be included in this discussion.

A comprehensive comparison of these tools is available in Ball (2020), but pertinent aspects will be discussed in this paper. This discussion will focus on the metadata that are available to be used in retrieval, relevant search features offered by the tool, the way in which results are displayed, as well as the knowledge or expertise required to use the tool.

Google Books Ngram Viewer

Google Books Ngram Viewer has already been introduced as a tool that allows a user to explore trends of word usage. The data used in the Ngram Viewer are obtained from a selection of over eight million books from the Google Books project, with the oldest texts from the 1500s (Michel, et al., 2011, p. 177). The quality of the optical character recognition data and the metadata played a role in the selection process (Michel, et al., 2011, p. 177). The dataset consists of n-grams extracted from the selection of texts, with the prerequisites that n-grams must occur a minimum of forty times in the corpus and the maximum size of an n-gram is five (Michel, et al., 2011, p. 177). An n-gram is a sequence of items (e.g., words) (Friginal, et al., 2014, p. 51). The frequency of an n-gram in a certain year is the ratio of occurrences of the n-grams in that year to the total number of words in the corpus for that year (Michel, et al., 2011: 177). The third dataset for the Ngram Viewer was made available in February 2020 (Google Books Ngram Viewer Info, 2020). Neither the texts nor the metadata of the texts that are used in the dataset are made available, because of copyright restrictions (Culturomics, 2017, p. 182; Koplenig, 2017).

Researchers are both captivated by and critical of the Google Books Ngram Viewer. It can be used to study cultural and linguistic trends at a macroscopic level (Lin, et al., 2012) and has been used by researchers such as Acerbi, et al. (2013); Juola (2013); Ophir (2016). Several concerns have been raised about the use of the Ngram Viewer to conduct research. One of the main problems is the absence of metadata of the dataset (e.g., Jockers, 2010; Koplenig, 2017: 170). Since the bibliography of the texts used cannot be released, the composition of the dataset is not clear. Results could be affected if the composition of the underlying data changes (Koplenig, 2017, p. 183) and '*the availability of metadata is not just a nice add-on, but a powerful source of information for the digital humanities... size cannot make up for lack of metadata*' (Koplenig, 2017, p. 183, 184). Other concerns are that the main dataset

includes a disproportionate amount of scientific publications (Pechenick, et al., 2015, p. 23) and that it only includes one copy of each item and so does not consider the popularity of an item (Pechenick, et al., 2015, p. 2).

The lack of bibliographic metadata means that a user cannot search for words or phrases and filter according to bibliographic metadata. However, some limited filtering with bibliographic metadata is allowed, namely, publication year and the language of the text. There are texts in eight different languages, and a distinction is made between British and American English. There is also one option to filter the data according to genre and that is by using the English Fiction filter. The search in Figure 1 demonstrates filtering according to genre and time.

The data have been annotated with part-of-speech tags (for example, *fall* can be a noun or verb) and head-modifier dependencies (for example, in the phrase *the morning flight*, *morning* modifies *flight*) (Lin, et al., 2012; Michel, et al., 2011). Automatic taggers and parsers were used to annotate the data using twelve language universal part-of-speech tags and unlabelled head-modifier dependencies (Lin, et al., 2012). With the annotations, a user can therefore search for a word that is part of a specific part-of-speech category by using tags, as well as words that modify other words, as is demonstrated in Figure 2.

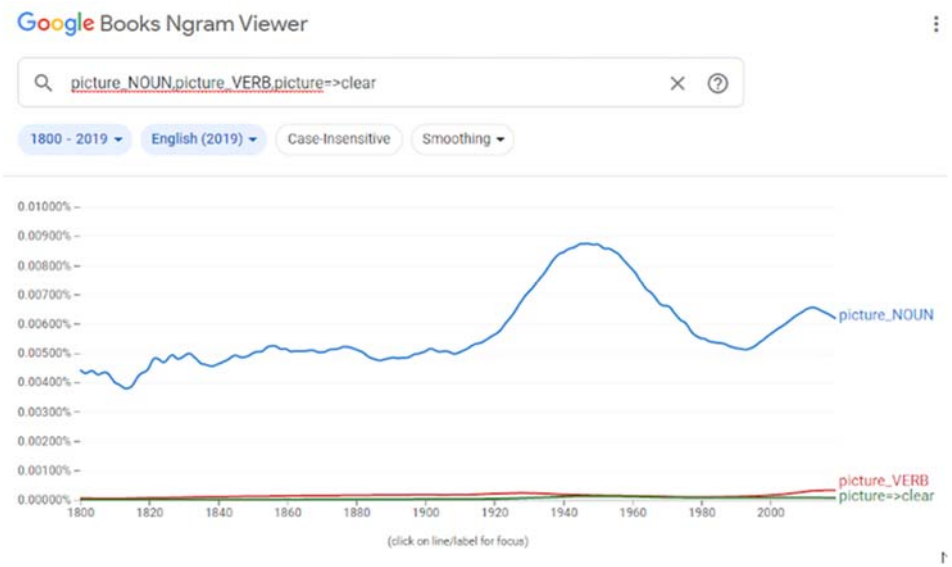


Figure 2: Using morphological and syntactic data in the Google Books Ngram Viewer

A user can search for different words or phrases, separated by commas. A wildcard can be used as a placeholder for a word (for example searching for combinations of *bread and* another word), but not as a placeholder for characters in a word (for example, searching for words ending in *-ly*).

The results of the search are displayed in a graph and do not link to examples in context. Below the graph are links to predetermined searches in Google Books for the terms that were searched for filtered by date. For example, the first link will open a search in Google Books for the term *noble* in books published from 1800–1810, as in Figure 3.

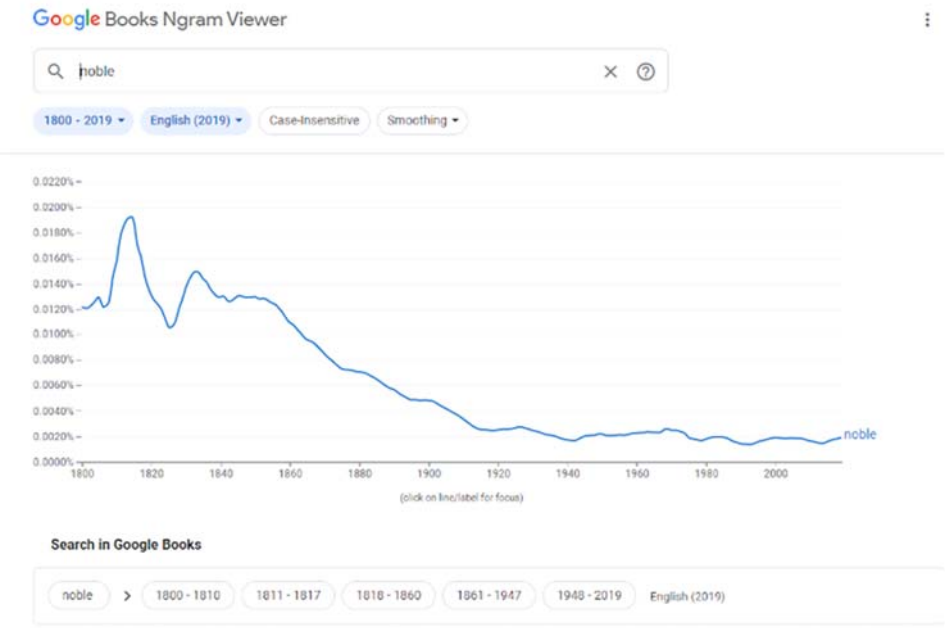


Figure 3: Links to Google Books

The tool is easy to use and could be used by scholars and lay persons.

The Google Books Ngram Viewer allows for searching according to some of the metadata in Table 1. It is possible to search for inflected forms, by part-of-speech categories, and dependency between words. However, it is not possible to search by meaning, logical features or structural features. It is only possible to search by bibliographic detail to a limited extent and it is not possible to search by in-text bibliographic detail.

HathiTrust+Bookworm

The HathiTrust+Bookworm is similar to the Google Books Ngram Viewer, in that it is used to visualise the frequency of the usage of words over time. The HathiTrust Research Center and the Cultural Observatory team, who were involved with the development of the Google Books Ngram Viewer, developed this tool (The iSchool at Illinois, 2014). It uses the material in the HathiTrust Digital Library, a large digital library developed by the HathiTrust consortium, currently holding over seventeen million digitised items (HathiTrust, 2020).

The tool utilises the extensive bibliographic metadata in the HathiTrust Digital Library and allows a user to filter according to 19 bibliographic elements. The date filter is a separate filter. In this example (Figure 4), the frequency of the use of the words *lamp* and *candle* is shown, limited to texts from 1850 to 1950 that were published in the United Kingdom, in the class language and literature, in the narrow class fiction in English, where the resource type is book.

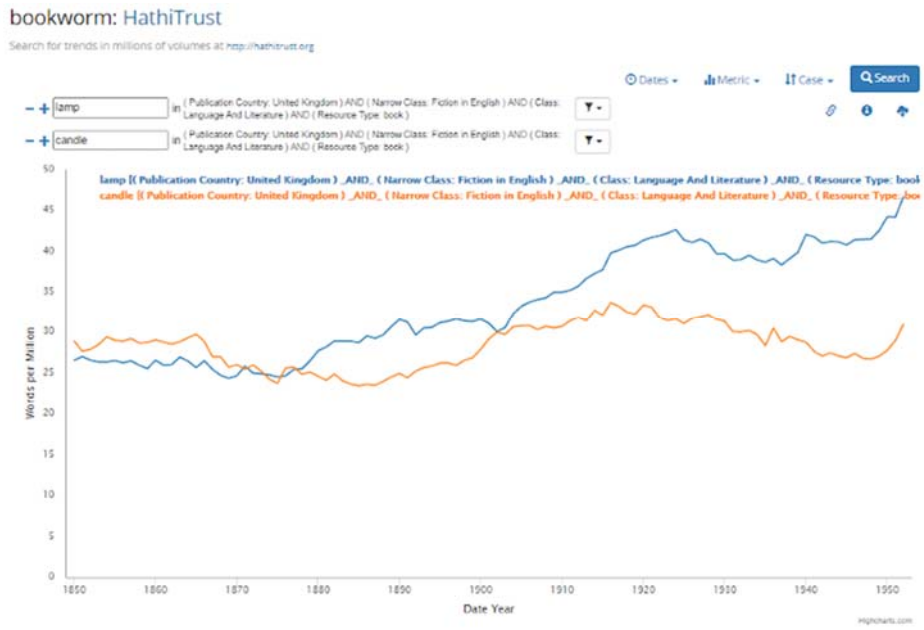


Figure 4: Searching for the terms *lamp* and *candle* in the HathiTrust+Bookworm, filtered according to bibliographic data

The bibliographic filtering is only on volume level. The tool also does not offer many search options. A user can compare several individual words, but not phrases. There is no query language specified.

The tool links the results in the graph with the underlying data in the collection to a limited extent. A user can click on a point in the graph to see top search results for the term in that year, as in Figure 5. The results link to the volumes in the library that contain the term, but the term is not shown in context.

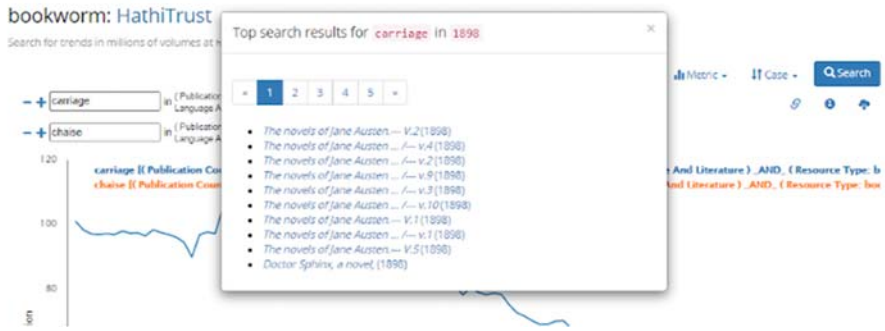


Figure 5: A list of items where *carriage* appears in 1898 in HathiTrust+Bookworm

Similar to the Google Books Ngram Viewer, the tool has a simple interface and is easy to use. There is no explanation for the different bibliographic elements that can be used for filtering, but most fields are self-explanatory.

The HathiTrust+Bookworm does not include many of the metadata in Table 1. It is not possible to search for inflected forms, part-of-speech categories, dependency between words, meaning, logical features or structural features. However, it offers extensive bibliographic metadata on a volume level. Even so, it is not possible to search by in-text bibliographic metadata.

Perseus Project

The Perseus Digital Library was designed to explore new options that online digital collections present to users. The collection originally only contained material from the Greco-Roman world, but currently other collections are also hosted in the library (Perseus Digital Library, n.d.). Where the aim of the Google Books Ngram Viewer and HathiTrust+Bookworm is to give a macroscopic view of a collection, the Perseus Project facilitates detailed study by giving access to individual items. The texts are encoded to make the structure (e.g., paragraphs, sentences) of the text understandable to a machine. The project made a significant contribution to research in the humanities.

Though the texts in this library contain bibliographic data, it is not possible to filter according to bibliographic metadata. The texts are divided into collections and a user browses to a specific text to study or searches for a specific text. The collections are useful, as a user can filter according to collections when searching for instances of words (e.g., Figure 6).

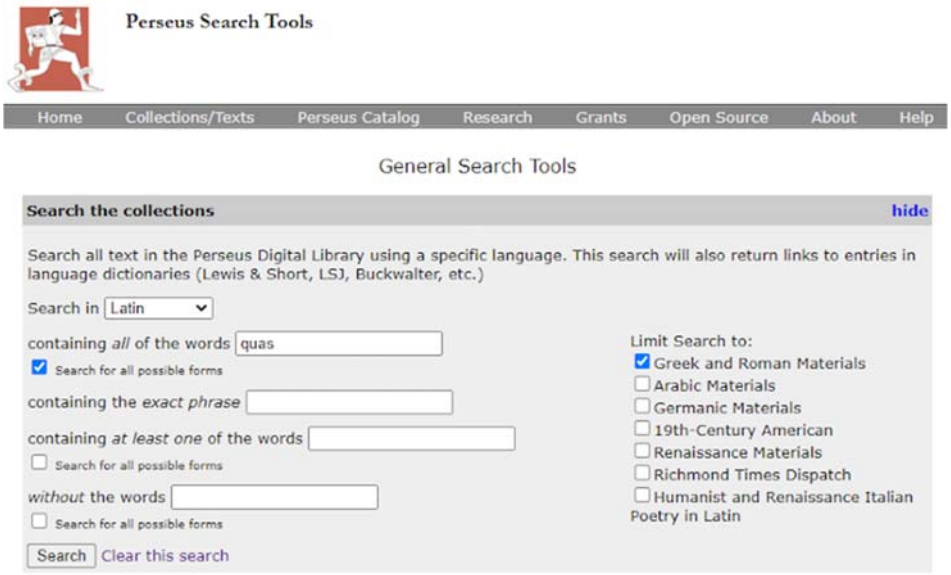


Figure 6: Searching for all forms of a word in the Perseus Project, limited to a specific collection

A user is able to search for all inflected forms of a word. The Perseus Project is able to do this type of searching by saving parsed Latin or Greek words in a database, and forming links from the words in the texts to the database (Rydberg-Cox, et al., 2000). Figure 6 demonstrates a search for the word *quas* in all possible forms, appearing in Latin texts, limited to Greek and Roman materials. The results of the different inflected forms of the search term are shown in Figure 7. Each instance is shown in some context and a user can expand to see more context. Being able to search for an inflected form is very useful, especially for learners of a language.

Figure 7: Results for a word in the Perseus Project

A user can also get more information about a specific word, either by searching for the word or reading a text and then selecting a word to get more information about the word. In Figure 8 a section of Luke (as translated by Saint Jerome) is shown with the word *populo* highlighted. Figure 9 shows more information about the selected word. The tool also uses statistical methods to suggest the most likely meaning and case in the context in which the word appears but, as such, it is not necessarily correct. (See the message from the tool in Figure 9). This could be improved by user votes if the system were widely used.

The two main search options provide the user with the ability to search for information about a specific word (almost like a dictionary lookup) and secondly to search for all occurrences of a word in (a selection of) texts. When searching for entries in the collection, a user can also search for a phrase. The tool does include Boolean operator logic built into the search options (see Figure 6), for example the option *without the words* corresponds to the NOT operator. The tool does not include any use of truncation or other commands from a query language.

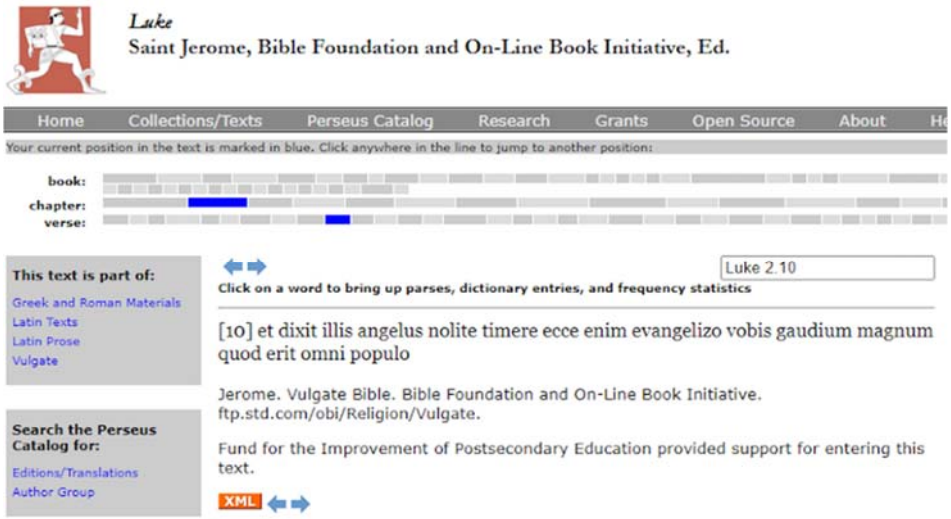


Figure 8: Selecting a word in the Perseus Project

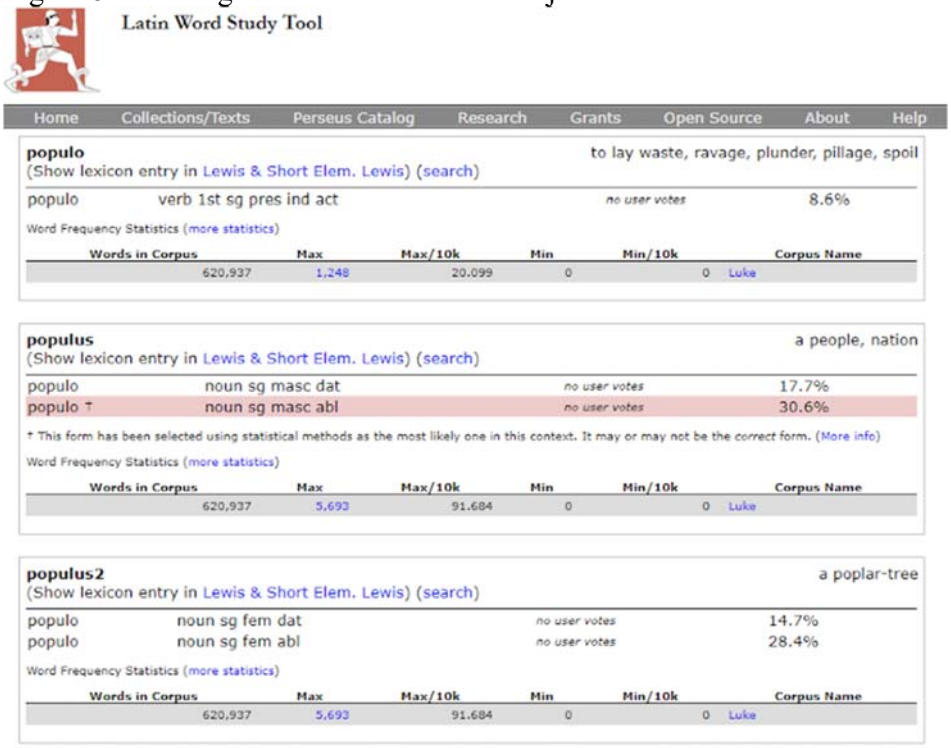


Figure 9: The suggested meaning of a word in the Perseus Project

An interesting aspect of this project is that the encoding reveals sections in one text that are written in different languages. For example, in the *Commentary on the Aeneid of Vergil* by Maurus Servius Honoratus (as reviewed by Georgius Thilo and Hermannus Hagen), there is a Greek word in a Latin text (as highlighted in Figure 10). The encoding shown in Figure 11 makes the different languages explicit. This has an important implication for searching.

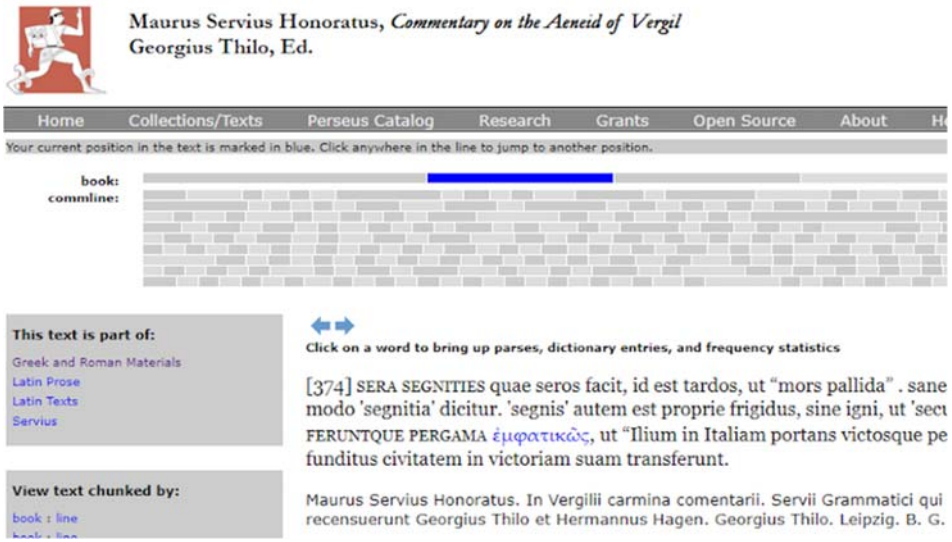


Figure 10: Greek word in Latin text

```
▼<TEI.2>
  ▼<text lang="la">
    ▼<body>
      ▼<div1 type="book" n="2" org="uniform" sample="complete">
        ▼<div2 type="commline" n="374" org="uniform" sample="complete">
          ▼<p>
              <hi rend="caps">sera segnities</hi>
              quae seros facit, id est tardos, ut
            ▼<cit>
                <quote>mors pallida</quote>
                <bibl default="NO"/>
              </cit>
              .
              <delSpan status="unremarkable" to="delend1272"/>
              sane 'segnities' iuxta antiquos dictum est, nam modo 'segnitia'
              <anchor id="delend1272"/>
              <hi rend="caps">feruntque pergama </hi>
              <foreign lang="greek">ἐμφατικῶς</foreign>
              , ut
            ▼<cit>
                <quote>Ilium in Italiam portans victosque penates</quote>
                <bibl default="NO"/>
              </cit>
```

Figure 11: Encoding of languages in the Perseus Project

The Perseus Project is fairly intuitive to use. Simple elements, such as text input fields, checkboxes and dropdown menus are used to enable searching, and descriptive labels are used. The largest collection in this library contains classical works, and one would therefore assume that the users who are drawn to this tool will already have some knowledge about the texts, as well as possibly some knowledge of some of the classical languages.

When considering the metadata in Table 1, the Perseus Project offers searching by some of these metadata. It is possible to search for inflected forms. However, it is not possible to specify which part-of-speech category to search for, nor to search for

dependencies between words or for words with different meanings. It is possible to search for some logical features, specifically people, places and dates. Though the texts in the collections are encoded with structures, it is not also possible to search in these. It is not possible to search using the bibliographic metadata, but some filtering according to the metadata is allowed. It is not really possible to search by in-text bibliographic detail, expect by the language as was demonstrated in this section.

BNCweb (CQP-edition)

Various spoken and written sources were used to construct the British National Corpus (BNC), an English corpus of 100 million words (Burnard, 2009; Grant, 2005: 437–438; Kennedy, 2003, p. 471). An XML edition of the BNC was published in 2007. This encoded version includes part-of-speech information, structures of texts (e.g., quotes, paragraphs, lists) and bibliographic information (University of Oxford IT Services, 2015). There are different ways in which to search the British National Corpus (BNC), of which the BNCweb (CQP-edition) hosted by Lancaster University is one such tool. The original BNCweb is easy to use, but lacks the ability to perform complex queries (Hoffmann and Evert, 2006). The goals of the CQP-edition of the BNCweb was to be as user-friendly as the original version, but to incorporate the powerful Corpus Query Processor (CQP) (Hoffmann and Evert, 2006). According to Hoffmann and Evert (2006, p. 180), one of the advantages of CQP is its ability to integrate metadata and queries.

By using CQP, complex queries can be executed on the corpus. A simplified query language based on CQP was created to allow novices to use the data more easily (Hoffmann and Evert, 2006). Various advanced search features are available in both standard CQP and the simplified language; for example, one can use wildcard characters to search for variations of words, specify part-of-speech tags when searching, search for lemmas or sequences of words and search within XML tags (Evert, 2005). The example in Figure 12 (*AJ0 *** *oo+oo**) is written in the simplified query language and searches for a word that contains a double o preceded by zero or more characters, followed by one or more characters, followed by a double o, followed by zero or more characters, and will find instances such as schoolroom or footloose. These instances must be preceded by any adjective that is up to three tokens away, such as *emerald green leather footstool*.

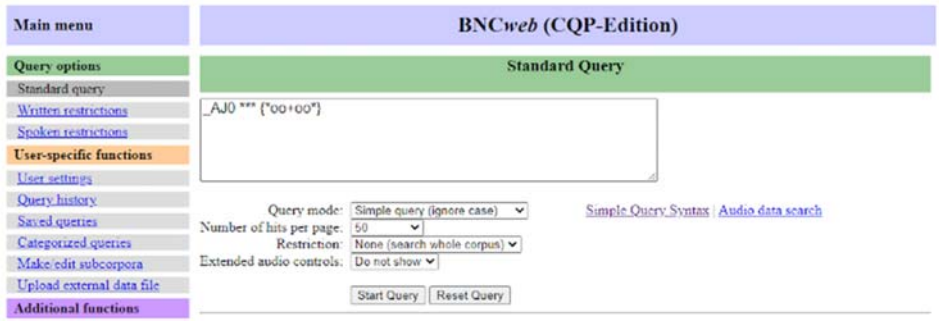
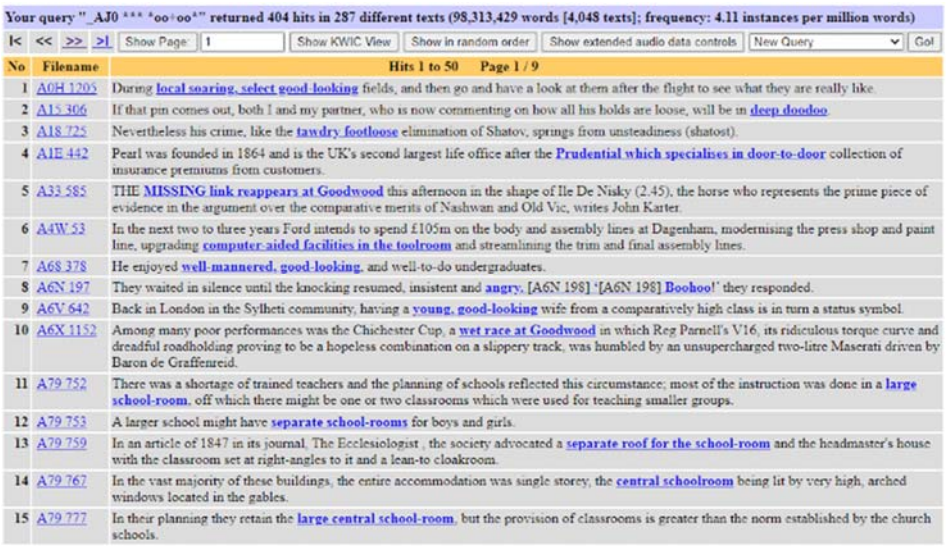


Figure 12: Searching in the BNCweb

It is possible to search in structures of the text, by searching within the XML tags. For example, the query <*quote*>*good*</*quote*> searches for instances of *good* within quotes.

There are bibliographic data (relevant to this corpus) available for each text that is included. For example, publication date, medium of text, genre. This allows a user to search in a specific subset of the corpus. There are no bibliographic data on an in-text level.

By default, the results are displayed in list form, as in Figure 13. There is some context given for each instance, with the search term highlighted. It is possible to link to more context, as well as see information about the text that the instance is from.



Your query "_AJ0 *** *oo+oo*" returned 404 hits in 287 different texts (98,313,429 words [4.04S texts]; frequency: 4.11 instances per million words)

|< << >> >| Show Page: 1 Show KWIC View Show in random order Show extended audio data controls New Query Go!

No	Filename	Hits 1 to 50 Page 1 / 9
1	A0H 1205	During local soaring, select good-looking fields, and then go and have a look at them after the flight to see what they are really like.
2	A15 306	If that pin comes out, both I and my partner, who is now commenting on how all his holds are loose, will be in deep doodoo.
3	A18 725	Nevertheless his crime, like the tawdry footloose elimination of Shatov, springs from unsteadiness (shatost).
4	A1E 442	Pearl was founded in 1864 and is the UK's second largest life office after the Prudential which specialises in door-to-door collection of insurance premiums from customers.
5	A33 585	THE MISSING link reappears at Goodwood this afternoon in the shape of Ile De Nisky (2.45), the horse who represents the prime piece of evidence in the argument over the comparative merits of Nashwan and Old Vic, writes John Karter.
6	A4W 53	In the next two to three years Ford intends to spend £105m on the body and assembly lines at Dagenham, modernising the press shop and paint line, upgrading computer-aided facilities in the toolroom and streamlining the trim and final assembly lines.
7	A68 378	He enjoyed well-mannered, good-looking, and well-to-do undergraduates.
8	A6N 197	They waited in silence until the knocking resumed, insistent and angry. [A6N 198] '[A6N 198] Boohoo!' they responded.
9	A6V 642	Back in London in the Sylheti community, having a young, good-looking wife from a comparatively high class is in turn a status symbol.
10	A6X 1152	Among many poor performances was the Chichester Cup, a wet race at Goodwood in which Reg Parnell's V16, its ridiculous torque curve and dreadful roadholding proving to be a hopeless combination on a slippery track, was humbled by an unsupercharged two-litre Maserati driven by Baron de Graffenreid.
11	A79 752	There was a shortage of trained teachers and the planning of schools reflected this circumstance; most of the instruction was done in a large school-room, off which there might be one or two classrooms which were used for teaching smaller groups.
12	A79 753	A larger school might have separate school-rooms for boys and girls.
13	A79 759	In an article of 1847 in its journal, The Ecclesiologist , the society advocated a separate roof for the school-room and the headmaster's house with the classroom set at right-angles to it and a lean-to cloakroom.
14	A79 767	In the vast majority of these buildings, the entire accommodation was single storey, the central schoolroom being lit by very high, arched windows located in the gables.
15	A79 777	In their planning they retain the large central school-room, but the provision of classrooms is greater than the norm established by the church schools.

Figure 13: Results in BNCweb (CQP-edition)

It is possible to use the tool for basic searches, by just searching for a single word. However, the advanced features would require some knowledge. The tool is probably aimed at people interested in language studies or linguistics.

When considering the metadata in Table 1, the BNCweb (CQP-edition) includes several of these options. It is possible to search for inflected forms and part-of-speech categories. It is, however, not possible to search for dependencies between words or for words with different meanings. It is not possible to search for logical features, but it is possible to search within structures through the use of a query language. It is not possible to search using the bibliographic metadata or in-text bibliographic detail.

TXM

TXM is a software platform developed by textometry teams (Heiden, 2010) and is described as a '*free, open-source Unicode, XML & TEI compatible text/corpus analysis environment and graphical client based on CQP and R*' (Text Encoding Initiative, 2016). The tool is developed to be compatible with TEI encoded data (Heiden, 2010). TXM makes it possible to utilise granular metadata when analysing and searching in a text.

The TXM processing model explains the design of corpora that can be handled by TXM (TXM User Manual, 2018). A very brief explanation of the model will be given here. A corpus may consist of texts, which are described by bibliographic metadata. Each text may have different structures. The word is the smallest unit of a text, and words may have different attributes.

As different corpora may be imported into TXM, the metadata that are available to use in a search will depend on the corpora that is being analysed. Furthermore, the type of encoding may differ between corpora. The VOEUX corpus will be used as an example. This corpus consists of speeches by French presidents. Each speech is encoded with some bibliographic metadata (e.g., the name of the president), functional metadata (e.g., paragraphs, sentences, line beginnings) and morphological metadata (e.g., part-of-speech category, lemma). These metadata can be used in a search.

TXM is a powerful tool offering advanced search and analysis options. A corpus query language (CQL) may be used to construct complex search queries; however, a user is not required to know this query language and may use graphical query assistants to build queries. The tool accommodates the use of truncation and wildcard characters to search for variations of terms.

Three examples will be given here. The first (Figure 14) will demonstrate how the query language can be used, the next example (Figure 15) will illustrate how the graphical query assistant works, and lastly the use of the graphical query assistant (Figure 16) will be used to search in the functional metadata.

Using the query language, a complex query can be executed on TXM, for example [word = ".*t"][word = "mon"] [frpos = "NOM"]. In this query, a specific sequence of words should be retrieved. This sequence should start with a word that ends in the letter t, followed by *mon*, followed by a noun.

text_id	Left context	Pivot	Right context
0009	Françaises, Français ! De	tout mon cœur	, je souhaite une bonne année à la France. Par là
0010	nouvelle année, je vous les offre de	tout mon cœur	. Un cœur que, depuis longtemps, permettez-moi de le
0011	vous, je dis ce soir, avec	tout mon coeur	d'homme et de Français : « Bonne année. » Que
0012	enfants, pour leurs vieux parents, de	tout mon coeur	je dis bonne année.
0029	violences de la nature ont accablés. C'	est mon rô1e	, je crois, que d'exprimer au nom de tous la
0029	, respectée. Elle l'est. C'	était mon devoir	aussi que de la prémunir contre ses divisions, que de témoigner
0038	pour faire reculer le chômage, ce qui	est mon objectif	essentiel. Mais notre croissance repart et elle sera plus forte en
0049	m'autoriserai aucune hypocrisie. J'ai mis	tout mon cœur	, et toute mon énergie à être le Président de tous les
0050	la vérité et j'ai agi. C'	était mon devoir	. Le pire aurait été que, dans cette situation, chaque

Figure 14: A complex query in TXM

Figure 15 shows how the query assistant can be used to create the query.

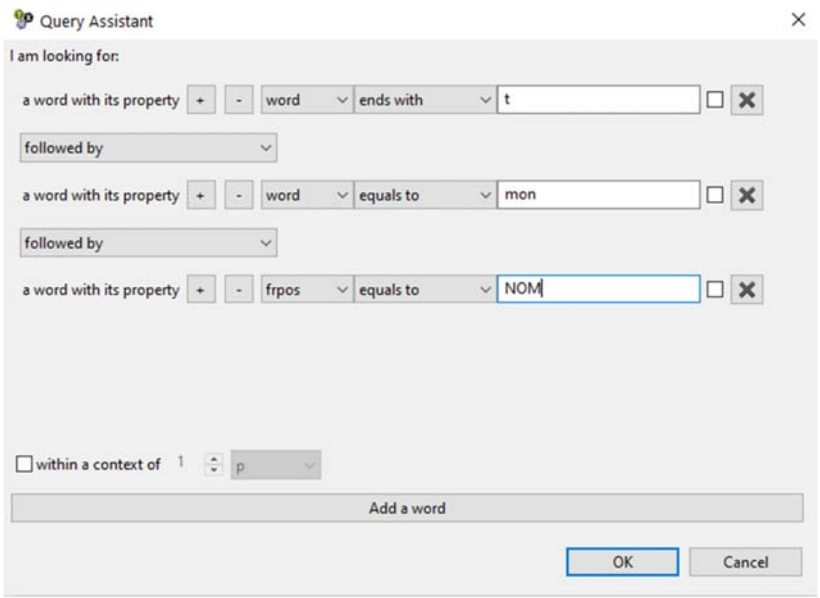


Figure 15: Using the graphical query assistant to construct a query in TXM

TXM allows a user to search in the functional metadata. In the example in Figure 16, the user is searching in paragraphs (p). The numbers of the paragraphs are listed as metadata.

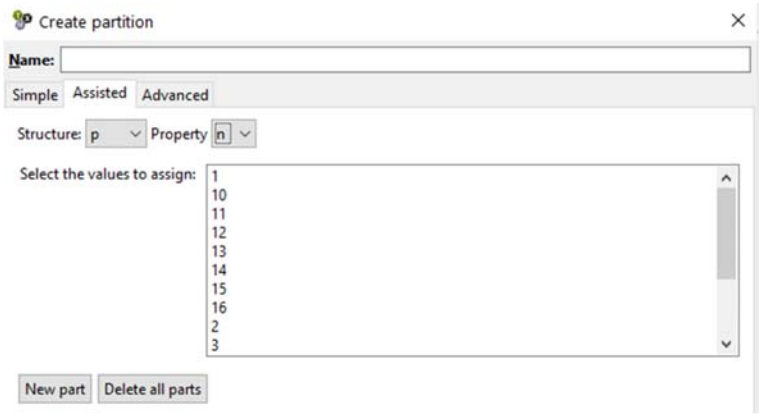


Figure 16: Searching in functional metadata in TXM

One of the distinguishing features of TXM is that it allows a user to return, from the search results, to the main text. Figure 17 shows the search results displayed in keyword in context form. A user can click on a word in the results and the text is displayed in the panel at the top, with the relevant words highlighted.

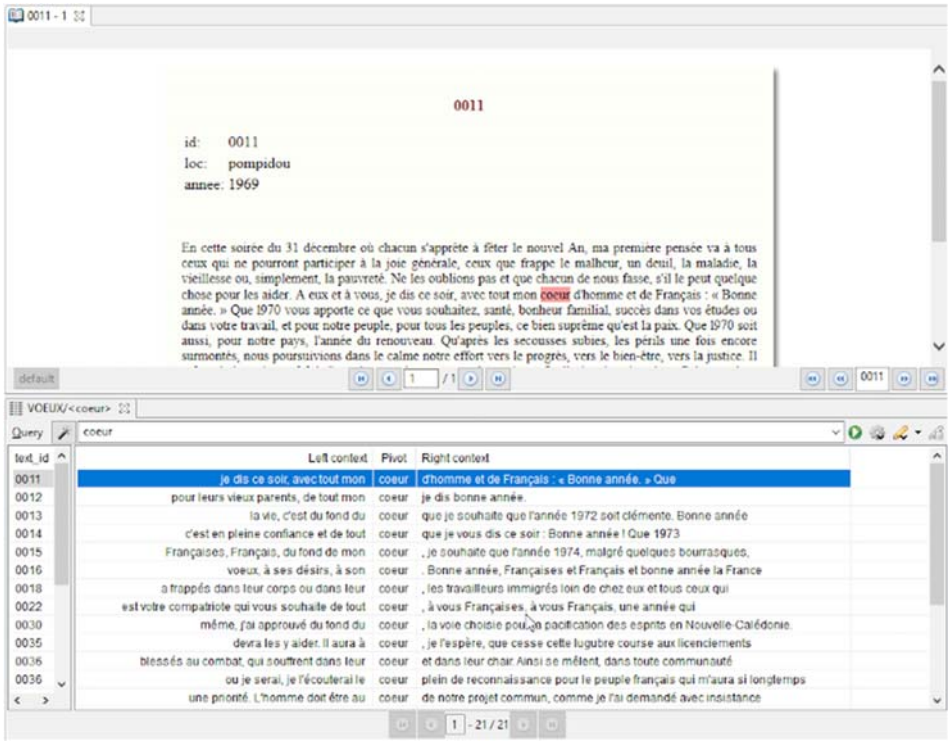


Figure 17: Search results and the main text, as displayed in TXM

TXM is a powerful tool that presents a steep learning curve to the user. There are many features, and the interface is therefore not simple. Furthermore, the user needs to understand the underlying structure and encoding of the corpus to be able to search using the metadata of the corpus. A user will need to learn the query language or learn how to use the graphical query assistant to write effective search queries. It is possibly most suitable for linguists and other researchers who are familiar with (or

willing to learn) the encoding of a specific corpus and the advanced query language in TXM.

TXM allows searching by several of the metadata categories listed in Table 1. The search options will depend on the data in the corpus, but it is possible to search for inflected forms and part-of-speech categories. Depending on the data, the tool allows for logical features and searching within structures. It is not possible to search using the bibliographic metadata or in-text bibliographic metadata. It is not evident that searching for dependencies between words or for words with different meanings are possible.

Discussion

Though the large amount of data available requires advanced searching, and the technology to support advanced searching is available, it is evident from the previous sections that searching on a granular level is not well supported.

At the start of this paper the concept of metadata was discussed. Metadata provide additional information about entities and facilitate discovery. Traditionally, information professionals have focussed on capturing bibliographic metadata of information sources, typically on the text (volume) level. It is argued that such additional granular metadata are crucial when working with digital collections as it is through the information provided in the metadata that researchers can select the sources or specific examples relevant to their need. Unfortunately, not much detail about the sections of the text or the words in the texts are known, and the frustration of not being able to select and collect smaller sections has been noted. If metadata can be extended to also capture and record attributes about texts, sections of texts and words, searching on a granular level could be enabled. Such granular metadata could be used in retrieval to allow researchers to find words, phrases or sections that are relevant to their work.

The review of literature and the evaluation of tools enabled the authors to identify such metadata and develop categories or levels of metadata to enable granular retrieval. This will be discussed in the next section. The subsequent sections will be an analysis of the heuristic evaluation. Though some comments about the ability of users to retrieve on a granular level were given in the heuristic evaluation, this section will answer the question more directly.

Granular metadata to enable detailed retrieval

The value of identifying and recording attributes of texts or sections of texts and noting this in detailed metadata, specifically for retrieval, were highlighted in this study. If a sentence is marked as direct speech, in a document written by a certain author, then a user can search according to these metadata. It was argued that such attributes can be recorded by means of metadata. In the previous section the search options of various tools were evaluated. Through the evaluation it was evident that different tools enable users to search using some metadata. For example, the Google Books Ngram Viewer allow a user to specify the part-of-speech tag when searching for instances of a word. The Hathi-Trust+Bookworm has extensive bibliographic metadata per volume or text. The Perseus Project includes searching for inflected forms. TXM and BNCweb (CQP-edition) allow users to search in sections in the text through tags.

Apart from the metadata observed in the tools, other metadata were identified through reviewing literature. For example, the structures in a text, such as chapters, headings, front matter, can be considered. There could also be stylistic features in a text. One could consider the meaning of words and relationships between words, such as dependencies and anaphoric relations.

There is clearly an abundance of information that could be considered about a text and the words in the texts. It will be important to consider the benefits of different metadata for retrieval. The following categories were identified by Ball (2020): morphological, syntactical, semantic, functional, and bibliographic. These categories were listed with the methodology, but a brief explanation of each category is presented here, as well as the data in each category that could improve retrieval.

The study of words is referred to as morphology and includes studying how words are formed, their internal structure, and types of words (O'Grady, 2010). In this category it could be useful to encode the lemma, which is the distinguished form of a word (Blevins, 2013) and the part-of-speech category of a word. If the lemma of a word is known, it will be possible to search for all inflections of a word, for example, *catch, catching, caught*. The part-of-speech category will enable a user to filter according to types of words, for example, searching for *hand* as a verb or as a noun.

The rules that govern the structure of sentences in a language are known as the syntax of a language. Dependency grammar considers the relationships between words in a sentence and indicate which words are dependent on which (Bird, et

al., 2015). For example, in the sentence *The children eat fudge*, the words *children* and *fudge* depend on (or modify) *eat* (the governor). This could be useful information for retrieval. For example, one could do a search for all instances where a word (e.g., *beautiful*) modifies a specific word (e.g., *day*).

Semantics is concerned with the meaning of words. Knowing the sense of a word will also offer useful filtering options. For example, consider the noun *case*, which can refer to a container or an instance of a particular situation. In this example, filtering by part-of-speech is insufficient, as both instances are nouns. However, the meaning is different, and these differences can be represented by metadata that specify semantic categories.

Texts typically have different features which a human can identify by looking at the layout or by interpreting the content. For example, a paragraph, an address, a heading, a verse are all sections of writing that generally can be identified by their layout. Other features, such as dates or names, can be identified through interpreting the context. This information can allow a user to ask detailed questions. For example, a user could wish to compare the use of a word as used in direct speech to the use of the word in general writing; or a user could wish to search for a specific number as used to indicate a year, not as a general number. In this study these metadata are referred to as functional metadata.

Bibliographic metadata are used to represent a specific information resource to enable the identification and retrieval of resources. Bibliographic metadata are critical to determine the composition of a collection and allow filtering on a document level.

One of the arguments of Ball (2020) is that bibliographic metadata on a document level are not sufficient. It has been mentioned earlier that researchers have expressed the need for metadata on a more detailed level than an individual volume or document. A single volume may contain different texts; for example, a book may contain poems by various authors; a novel may contain quotes by other authors; a text may contain multiple languages; or a book containing a drama may have a long introduction with explanatory notes. Bibliographic metadata that are on the text level and give information about sections in the text (and not only on the volume or document level) may enhance retrieval. This will be referred to as in-text bibliographic metadata.

These granular metadata may be useful to allow scholars to retrieve words or phrases with specific attributes from a digital text collection.

The extent to which current tools support the retrieval of words or phrases with specific attributes from digital text collections

In this paper some tools that are currently being used to study digital collections were investigated. The different tools have different searching options. The ability of the tools to search according to granular metadata is summarised in Table 2 and discussed in this section.

Table 2: The extent to which current tools support the retrieval of words according to granular metadata

Granular metadata categories	Items in category	Google Books Ngram Viewer	HathiTrust+ Bookworm	Perseus Project	BNCweb (CQP-edition)	TXM
Morphological	Inflected forms (lemma)	Yes	No	Yes	Yes	Yes
	Part-of-speech category	Yes	No	No	Yes	Yes
Syntactic	Dependency between words	Yes	No	No	No	No
Semantic	Meaning	No	No	No	No	No
Functional	Logical features	No	No	Yes	No	Yes
	Structural features	No	No	No	Yes	Yes
Bibliographical	Bibliographic detail	Limited	Yes	No	Yes	No
	In-text bibliographic detail	No	No	No	No	No

Some tools allow for detailed filtering on bibliographic data (e.g., HathiTrust+Bookworm), others allow for searching according to part-of-speech categories and inflections (e.g., Google Books Ngram Viewer and BNCweb). The Google Books Ngram Viewer allows searching according to syntactic data. The BNCweb and TXM allow a user to search in functional areas if a user understands the structure and encoding of the data. The Perseus Project shows how a section in a text can be identified as having a different language from the main text and this fact can be incorporated in the search. Some tools show the results in context (e.g., BNCweb). Some tools show trends over time (e.g., Google Books Ngram Viewer and HathiTrust+Bookworm). Some tools have advanced searching options (e.g., the

BNCweb includes truncation to search for word variants, the Perseus Project implicitly includes Boolean operators).

These tools demonstrate that advanced searching in digital text collections is possible. However, there are still limitations in these tools. Tools that have a simpler, intuitive interface typically do not allow for detailed filtering (e.g., Google Books Ngram Viewer). There are tools that allow for retrieval on a fairly granular level, but then the user is required to understand the underlying structure and encoding of the data (e.g., TXM and BNCweb) and either be familiar with a query language (e.g., BNCweb) or be able to navigate a complex tool (e.g., TXM).

In the Google Books Ngram Viewer, there are no metadata about the meaning of words, or functional areas, such as heading or back matter, or in-text bibliographic metadata. In HathiTrust+Bookworm, it is not possible to search according to morphological or syntactic metadata, the meaning of a word or to search in functional metadata. The Perseus Project does not allow a user to search according to part-of-speech categories, syntactic data, semantic data or in functional metadata, but some of the features of the project are worth highlighting. BNCweb does not allow searching according to syntactic or semantic data.

A summary of the findings is presented here.

- From table 2 it is evident that there are a variety of search functions in tools.
- It is also evident that some search functions do not occur in any tools, namely,
    - none of the tools allows a user to search according to semantic data,
    - nor is there an option to search for in-text bibliographic metadata in any of the tools.
- From the analysis of the tools, further inferences can be made.
    - There are limited options to search in multiple levels of metadata simultaneously.
    - There is also no tool that allows a user to combine the metadata that are suggested in this study and enables a user to filter on all levels of metadata.
    - None of the tools examined in this study allows a researcher, who does not have knowledge of a query language or the structure and encoding of the data, to search according to granular metadata.

Recommendations for the development of a tool that enables retrieval of words and phrases with specific attributes

More work should be done to enable retrieval on a granular level, specifically for users with little programming experience, knowledge of encoding standards or the structure of the data. Future research should focus on ways in which texts can be enhanced with detailed metadata by determining which levels of metadata are useful, which elements on each level can provide more information that would assist researchers in conducting searches in collections and how these metadata can be encoded.

Research should not only focus on creating datasets with granular metadata but consider how software tools can use these metadata in retrieval. Ideally, the tool should be accessible to people with different levels of expertise. People with little or no programming expertise should be able to search effectively and efficiently, but advanced users should be able to perform more complex queries. It should not be necessary to understand the underlying data structure and encoding to use the tool. It would be useful if results could link to more context, in other words, the context in which the instance was used. A visualisation function should be considered, in order that trends may be observed.

In the research project conducted by the authors, the focus is exactly that. In the first instance metadata that can describe text in detail were defined and a prototype was developed to facilitate retrieval on a granular level. This will be discussed in further publications.

Conclusion

This paper argues that the ability to retrieve words or phrases from sections in a text, or words or phrases with specific attributes, could allow researchers to be specific in their queries and retrieve only relevant information. There are many attributes that could be considered when dealing with texts and words in texts. These attributes could be captured by granular metadata.

In order to answer research question 1, granular metadata useful for retrieval from digital text collections on a detailed level were identified (Table 1). These metadata are organised into the following categories: morphological, syntactic, semantic, functional and bibliographic. The category of bibliographic metadata should not only consider information at the document level, but should consider bibliographic

metadata in a text, specifically where the information of a section in a text differ from the information of the document.

After identifying useful granular metadata for retrieval for retrieval, a heuristic evaluation was performed to answer research question 2. The evaluation of current tools shows that retrieval on a granular level is limited (summarised in Table 2). Each tool offers some searching, but no tool cover all categories. The tools also in their level of difficultly. Tools that offer some retrieval on a granular level require understanding of data structure and encoding. More should be done to enable researchers to retrieve according to different attributes, as captured in granular metadata, easily and effectively.

In answer to research question 3, it is recommended that further research and development is done to improve granular retrieval. Such work could consider the inclusion of semantic metadata and in-text bibliographic metadata and allows for searching on these metadata fields. Furthermore, tools that offer filtering or searching on multiple levels of metadata should be developed. Such tools should be user-friendly and not require extensive knowledge of linguistics or one or more query languages.

For future research, it is suggested that the categories of granular metadata discussed in this paper are used when developing tools used for retrieval in digital text collections. The authors of this paper are furthermore particularly interested in a way to capture granular metadata for texts, and how to formalise this in a schema and encoding format. Furthermore, as researchers often work with large collections, the possibility to automate some of the encoding should be explored. A prototype or experimental tool that allows a user to retrieve words with certain attributes should be developed.

By allowing for retrieval of words or phrases according to specific attributes, more complex and specific queries can be conducted, enabling researchers to retrieve only what they need.

About the authors

Liezl H. Ball is a lecturer in the Department of Information Science at the University of Pretoria, South Africa. She completed her PhD in 2020. Her research is in the field of digital humanities and investigates how large text collections may be enhanced with metadata to improve retrieval. She can be contacted at liezl.ball@up.ac.za

Theo J.D. Bothma is Professor Emeritus / contract professor in the Department of Information Science at the University of Pretoria, South Africa. He is the former

Head of Department and Chairperson of the School of Information Technology (until his retirement at the end of June 2016). His research focuses primarily on information organisation and retrieval, information literacy and e-lexicography. He can be contacted at theo.bothma@up.ac.za

References

Note: A link from the title, or from "(Internet Archive)", is to an open access document. A link from the DOI is to the publisher's page for the document.

- Acerbi, A., Lampos, V., Garnett, P., & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLoS One, 8*(3), e59030. http://dx.doi.org/10.1371/journal.pone.0059030
- Ball, L. H. (2020). *Enhancing digital text collections with detailed metadata to improve retrieval* [Unpublished doctoral dissertation] University of Pretoria.
- Ball, L. H., & Bothma, T. J. D. (2020). *The capability of search tools to retrieve words with specific properties from large text collections* Proceedings of ISIC: the information behaviour conference, Pretoria, South Africa, 28 September - 1 October, 2020. *Information Research, 25*(4), paper isic2030. http://InformationR.net/ir/25-4/isic2020/isic2030.html (Internet Archive)
- Bird, S., Klein, E., & Loper, E. (2015). *Natural language processing with Python: analyzing text with the Natural Language Toolkit*. http://www.nltk.org/book/. (Internet Archive)
- Blevins, J. P. (2013). Word-based morphology from Aristotle to modern WP (word and paradigm models). In K. Allan, (Ed.), *Oxford handbook of the history of linguistics* (pp. 375–395). Oxford University Press.
- Bode, K. (2017). The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history. *Modern Language Quarterly, 78*(1), 77–106. https://doi.org/10.1215/00267929-3699787.
- Burnard, L. (2009). *British National Corpus*. University of Oxford. http://www.natcorp.ox.ac.uk/corpus/index.xml. (Internet Archive)
- Capitanu, B., Underwood, T., Organisciak, P., Cole, T., Sarol, M. J., & Downie, J. S. (2016). *The HathiTrust Research Center extracted feature dataset (1.0)*. Hathi Trust. https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId=37322778. http://dx.doi.org/10.13012/J8X63JT3.

- Cox, A. M. (2021). *Research report: the impact of AI, machine learning, automation and robotics on the information professions*. CILIP. https://www.cilip.org.uk/page/researchreport. (Internet Archive).
- Culturomics. (2017). *Cullturomics – FAQ*. http://www.culturomics.org/Resources/faq. (Internet Archive)
- Dombrowski, Q. (2020). Preparing non-English texts for computational analysis. *Modern Languages Open*, *1*, 45. http://doi.org/10.3828/mlo.v0i0.294.
- Evert, S. (2005). *The CQP Query Language tutorial (CWB version 2.2.b90)*. http://www-3.unipv.it/larl/cqp-tutorial.pdf. (Internet Archive)
- Fenlon, K., Senseney, M., Green, H., Bhattacharyya, S., Willis, C., & Downie, J. (2014). Scholar-built collections: a study of user requirements for research in large-scale digital libraries. *Proceedings of the Association for Information Science and Technology*, *51*(1), 1–10. https://doi.org/10.1002/meet.2014.14505101047
- Finlayson, M. A. (2015). ProppLearner: deeply annotating a corpus of Russian folktales to enable the machine learning of a Russian formalist theory. *Digital Scholarship in the Humanities*, *32*(2), 284–300. https://doi.org/10.1093/llc/fqv067.
- 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. http://www.e-journall.org/wp-content/uploads/Friginal_et_al_1.1.pdf. http://dx.doi.org/10.21283/2376905X.1.4.
- Gilliland, A. J. (2016). Setting the stage. In M. Baca (Ed.), *Introduction to metadata*. Getty Research Institute.
- Google Books Ngram Viewer Info. (2020). *Google Books Ngram Viewer Info*. https://books.google.com/ngrams/info (Internet Archive)
- Grant, L. E. (2005). Frequency of 'core idioms' in the British National Corpus (BNC). *International journal of corpus linguistics*, *10*(4), 429–451. https://bit.ly/35Cd8a0. https://doi.org/10.1075/ijcl.10.4.03gra
- Haider, J., & Sundin, O. (2019). *Invisible search and online search engines: the ubiquity of search in everyday life*. Routledge.
- Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, *17*(3), 380–409.

https://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf. https://doi.org/10.1075/ijcl.17.3.04har. (Internet Archive)

- HathiTrust. (2020). *HathiTrust Digital Library about*. https://www.hathitrust.org/about (Internet Archive)
- Haynes, D. (2018). *Metadata for information management and retrieval: understanding metadata and its use*. Facet Publishing.
- Heiden, S. (2010). The TXM platform: building open-source textual analysis software compatible with the TEI encoding scheme. In R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, & Y. Harada, (Eds.). *24th Pacific Asia conference on language, information and computation*, Tohoku University, Sendai, Japan, November 2010. (pp. 389-398) https://aclanthology.org/Y10-1044.pdf (Internet Archive)
- Henry, C., & Smith, K. (2010). Ghostlier demarcations: large-scale text digitization projects and their utility for contemporary humanities scholarship. In A. Bishop, C. Clotfelter, A. Friedlander, D. M. Gift, A. Holly, C. Lynch, M. McPherson, C. Moore, S. Nichols, & J. Williams (Eds.), *The idea of order: transforming research collections for 21st century scholarship* (pp. 106–115). Council on Library and Information Resources.
- Hoffmann, S., & Evert, S. (2006). BNCweb (CQP-edition): the marriage of two corpus tools. In S. Braun, K. Kohn, & J. Mukherjee, (Eds). *Corpus technology and language pedagogy: New resources, new tools, new methods, 3*, (pp. 177–195). Peter Lang. http://corpora.lancs.ac.uk/BNCweb/Hoffmann-Evert.pdf. (Internet Archive)
- Howard, J. (2017). *What happened to Google's effort to scan millions of University Library books?* https://bit.ly/3HrhVcu. (Internet Archive)
- Jett, J., Nurmikko-Fuller, T., Cole, T. W., Page, K. R., & Downie, J. S. (2016). Enhancing scholarly use of digital libraries: A comparative survey and review of bibliographic metadata ontologies. In *JSCL'16: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, Newark, New Jersey, USA, June 19 - 23, 2016 (pp. 35-44). Association for Computing Machinery. https://doi.org/10.1145/2910896.2910903.
- Jockers, M. L. (2010). *Unigrams, and bigrams, and trigrams, oh my*. http://www.matthewjockers.net/2010/12/22/unigrams-and-bigrams-and-trigrams-oh-my/ (Internet Archive)
- Juola, P. (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

- Kennedy, G. (2003). Amplifier collocations in the British National Corpus: implications for English language teaching. *Tesol Quarterly*, *37*(3), 467–487. https://doi.org/10.2307/3588400
- Keuleers, E., Brysbaert, M., & New, B. (2011). An evaluation of the Google Books ngrams for psycholinguistic research. In K. Würzner & E. Pohl (Eds.). *Lexical Resources in psycholinguistic research* (pp. 23–27). Universitätsverlag Postdam. (Potsdam Cognitive Science Series 3)
- 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. https://doi.org/10.1093/llc/fqv037
- Lansdall-Welfare, T., & Cristianini, N. (2020). History playground: a tool for discovering temporal trends in massive textual corpora. *Digital Scholarship in the Humanities*, *35*(2), 328-341. https://doi.org/10.1093/llc/fqy077
- Lee, H. (2019). *15 years of Google Books.* https://www.blog.google/products/search/15-years-google-books/ (Internet Archive)
- Leedy, P. D., Ormrod, J. E., & Johnson, L. R. (2021). *Practical research: planning and design* (12th ed.). Pearson Education Limited.
- Li, L., Huang, C. R., & Wang, V. X. (2020). Lexical competition and change: a corpus-assisted investigation of gambling and gaming in the past centuries. *SAGE Open, 10*(3), 1 – 14. https://doi.org/10.1177/2158244020951272.
- Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 8–14 July, Jeju, Republic of Korea. (pp. 169-174). Association for Computational Linguistics. https://aclanthology.org/P12-3029.pdf.(Internet Archive)
- Michel, J., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Lieberman, A. E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. http://dx.doi.org/10.1126/science.1199644
- Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How we do things with words: analyzing text as social and cultural data [Review]. *Frontiers in*

Artificial Intelligence, *3*(62). https://doi.org/10.3389/frai.2020.00062

- Nunberg, G. (2009). *Google Books: a metadata train wreck*. https://languagelog.ldc.upenn.edu/nll/?p=1701 (Internet Archive)
- Nyhan, J., Hauswedell, T., & Tiedau, U. (2020). Reflections on infrastructures for mining nineteenth-century newspaper data. In *Digital Scholarship, Digital Classrooms: New International Perspectives on Research and Teaching: Proceedings of the Gale Digital Humanities Day at the British Library*, 02 May 2019 (pp. 27-38) Gale. https://discovery.ucl.ac.uk/id/eprint/10090119/
- O'Grady, W. (2010). *Contemporary linguistics: an introduction* (6th ed.). Bedford Books.
- Ophir, S. (2016). Big data for the humanities using Google Ngrams: discovering hidden patterns of conceptual trends. *First Monday*, *21*(7). https://doi.org/10.5210/fm.v21i7.5567
- Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One*, *10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041
- Perseus Digital Library. (n.d.). *Perseus Digital Library – about*. http://www.perseus.tufts.edu/hopper/about (Internet Archive)
- Pickard, A. J. (2017). *Research methods in information*. Facet Publishing.
- Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly Media.
- Rydberg-Cox, J. A., Chavez, R. F., Smith, D. A., Mahoney, A., & Crane, G. R. (2000). Knowledge management in the Perseus Digital Library. *Ariadne*, 25. http://www.ariadne.ac.uk/issue25/rydberg-cox. (Internet Archive)
- Shneiderman, B., & Plaisant, C. (2010). *Designing the user interface. strategies of effective human-computer interaction* (5th ed.). Pearson Education.
- Text Encoding Initiative. (2016). *TXM*. https://wiki.tei-c.org/index.php/TXM (Internet Archive)
- Textometrie. (2018). *TXM user manual*. Textometrie. http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf (Internet Archive)

- Underwood, T. (2015). *Understanding genre in a collection of a million volumes*. University of Illinois, Urbana-Champaign. https://hcommons.org/deposits/item/hc:12277/ (Internet Archive)
- University of Illinois Urbana-Champaign. School of Information Sciences. (2014). *Exploring the billions and billions of words in the HathiTrust corpus with Bookworm: HathiTrust + Bookworm Project*. University of Illinois. https://ischool.illinois.edu/research/projects/hathitrust-bookworm-project. (Internet Archive)
- University of Oxford. Oxford Text Archive. IT Services. (2015). *British National Corpus*. University of Oxford. http://www.natcorp.ox.ac.uk/ (Internet Archive)
- Zeng, M., & Qin, J. (2016). *Metadata* (2nd ed.). Neal-Schuman.