informationresearch

VOL. 25 NO. 4, DECEMBER, 2020

Proceedings of ISIC: the information behaviour conference, Pretoria, South Africa, 28th September to 1st October, 2020.

Poster session

The capability of search tools to retrieve words with specific properties from large text collections

Liezl Ball, and Theo Bothma

Introduction. With the increase in the availability of digital text collections for humanities researchers, tools to enable enhanced retrieval are required. If words with very specific properties could be retrieved from a text collection more accurate linguistic and other analyses can be made. There are a range of properties and metadata that could be specified for retrieval, from morphological data up to bibliographic data. Furthermore, the bibliographic data should not only be on item level but extended to the text-level. For example, in an anthology each section could be encoded with the author of that section. Such extended metadata will enable fine-grained retrieval.
Method. In this study, current tools were evaluated to determine to what extent they allow users to retrieve words with specific properties from a text collection.
Analysis. The analysis is limited to the following criteria: interface design, metadata, search options, filtering and search results.
Results. Currently, it is not possible for a user to retrieve words with specific properties from a text collection.
Conclusion. An extended set of metadata should be used to encode text to enable retrieval of words on a fine-grained level.

DOI: **https://doi.org/10.47989/irisic2030**

Introduction

The amount of digital text data available is increasing rapidly. Researchers from different disciplines are interested in using these data in their research. Though one might associate the use of large amounts of data with the natural sciences, researchers in the humanities are increasingly interested in the use of large digital collections for research. Some of the well-known large digital text collections include Google Books, Internet Archive and HathiTrust.

The availability of such data, and technology to process the data, are opening up new possibilities in the field of humanities (e.g., Howard, 2017; Nicholson, 2013; Rydberg-Cox et al., 2000). Some tools have been developed to help users to search for information in large digital text collections, for example the Google Books Ngram Viewer (https://books.google.com/ngrams). Using this tool, a user can see the usage frequency of a term over a period of time. Some interesting research using this tool has been done (e.g., Acerbi et al., 2013; Michel et al., 2011; Ophir, 2016).

Problem statement

Though there are some exciting developments, there are some criticism against current tools and methods. For example, Google Books Ngram Viewer has been severely criticised for the lack of metadata (e.g., Koplenig, 2017). Furthermore, there are some additional features that could enhance the tools, specifically in terms of metadata. If texts are encoded with detailed metadata more powerful searches will be possible. Two examples will suffice. Morphological metadata could be added to enable a user to search for specific inflected forms of irregular verbs; or metadata could be used to indicate structures in a text. For example, if a document contains multiple languages, each section could be encoded with tags to specify the language of that section; this will enable a user to search for words in a specific language on a very detailed level. This means that texts need to be encoded on different levels, from metadata to indicate structure in texts to detailed morphological data.

This study forms part of a larger research project to determine how texts could be encoded on a detailed level to improve retrieval of words or phrases with specific properties from large digital text collections.

The first step in answering the broader question will be to examine the most popular tools used currently for the retrieval of words from a large text collection and answer the following question:

To what extent do current tools allow a user to retrieve words or phrases with specific properties?

Significance of the topic

This research makes an important contribution in the field of information seeking and use. The overwhelming amount of text data available will only be effectively utilised if useful tools are developed that enable users to retrieve the relevant information in terms of their information needs. This study focuses on the ability of users to retrieve words with specific properties from a text collection. By evaluating current tools, recommendations can be made

for further development of advanced search tools used on large digital text collections. Improved tools can help researchers, authors, and other users to observe trends and analyse the results.

Examination of tools

Six tools that can be used to search for words (or phrases) in a digital text collection (corpus) are examined to determine to what extent a user can search for words or sections of texts with specific properties. The six tools that have been identified are the Google Books Ngram Viewer, HathiTrust+Bookworm, Perseus Project, Voyant Tools, TXM and BNCweb. These tools are used to search in large digital text collections. These tools are examined according to the following criteria: interface design, metadata, search options, filtering and search results. More criteria, such as complexity of use, help files and corpus design, could be considered in future studies. Key features in these categories are highlighted. It is beyond the scope of this paper to offer extensive detail of each tool.

Google Books Ngram Viewer

The Google Books Ngram Viewer shows the relative frequency of words (or phrases) used over a specified period of time. The data used by this tool are from a selection of books from Google Books. An example of a search is shown in Figure 1. In this example the tool is showing all instances of *well* where it is a noun, of *well* where it is an adjective and where *well* modifies *done*.

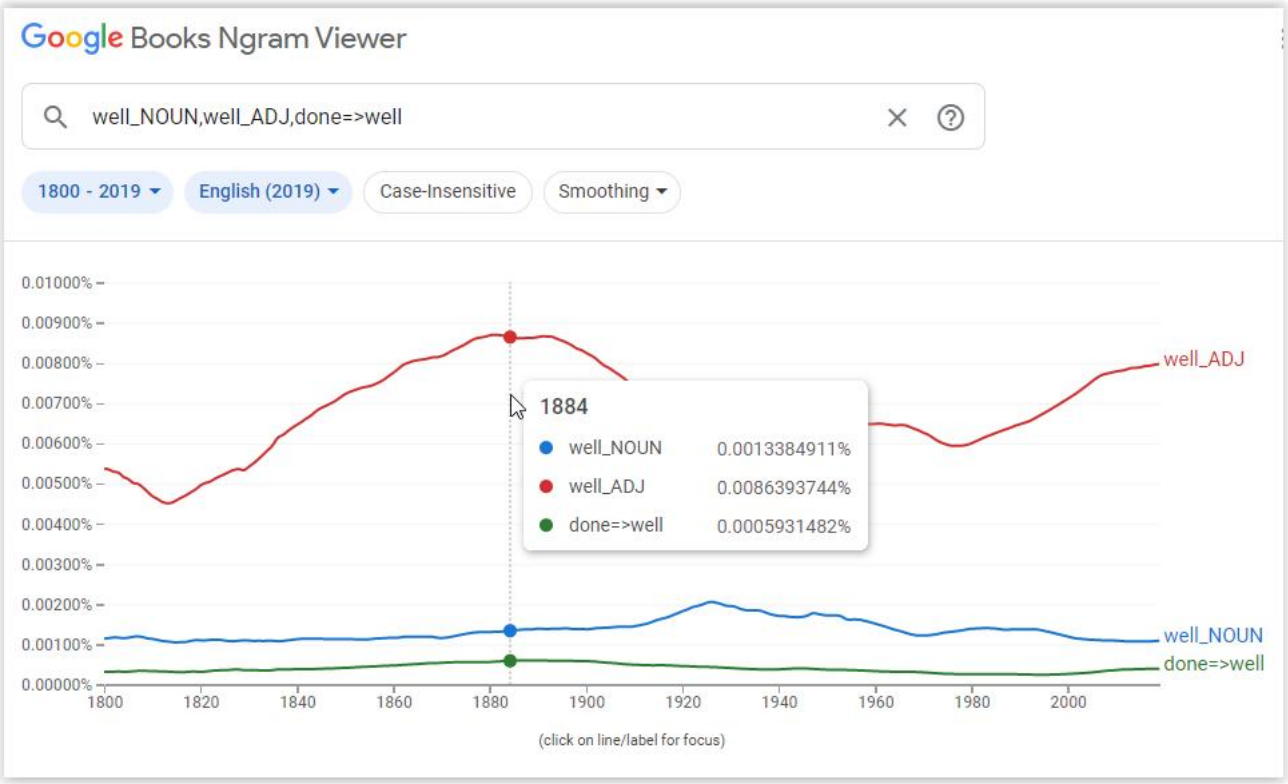


Figure 1: Google Books ngram viewer

- Interface design: The interface is simple and intuitive to use. A user can simply enter a single word or phrase in the search field. An example search is available to demonstrate the use of the tool.
- Metadata: The dataset is annotated with morphological and syntactic data.
- Search options: Apart from entering a single word or phrase, a user can also compare terms by separating the terms by commas. It is possible to search for inflected forms, parts-of-speech categories and words that modify other words (syntactic properties). Truncation can be used to replace whole words in a search. Other search options are also available.
- Filtering: It is possible to filter by date and language. The dataset includes eight languages. There is one subset for genre, which is the English fiction dataset.

Search results: The results are displayed in a graph. The results do not link to the underlying data directly and it is not possible to see examples in context. Below the graph are links to predetermined searches in Google Books for the search terms used in the specific search.

HathiTrust+Bookworm

The HathiTrust+Bookworm) also visualises the frequency of words over a period of time, however a user can filter the results using bibliographic metadata. The data used by this tool come from the HathiTrust Digital Library. In Figure 2 a search for the terms *carriage* and *chaise* is shown. Filters, such as publication country, class and resources type, have been applied.

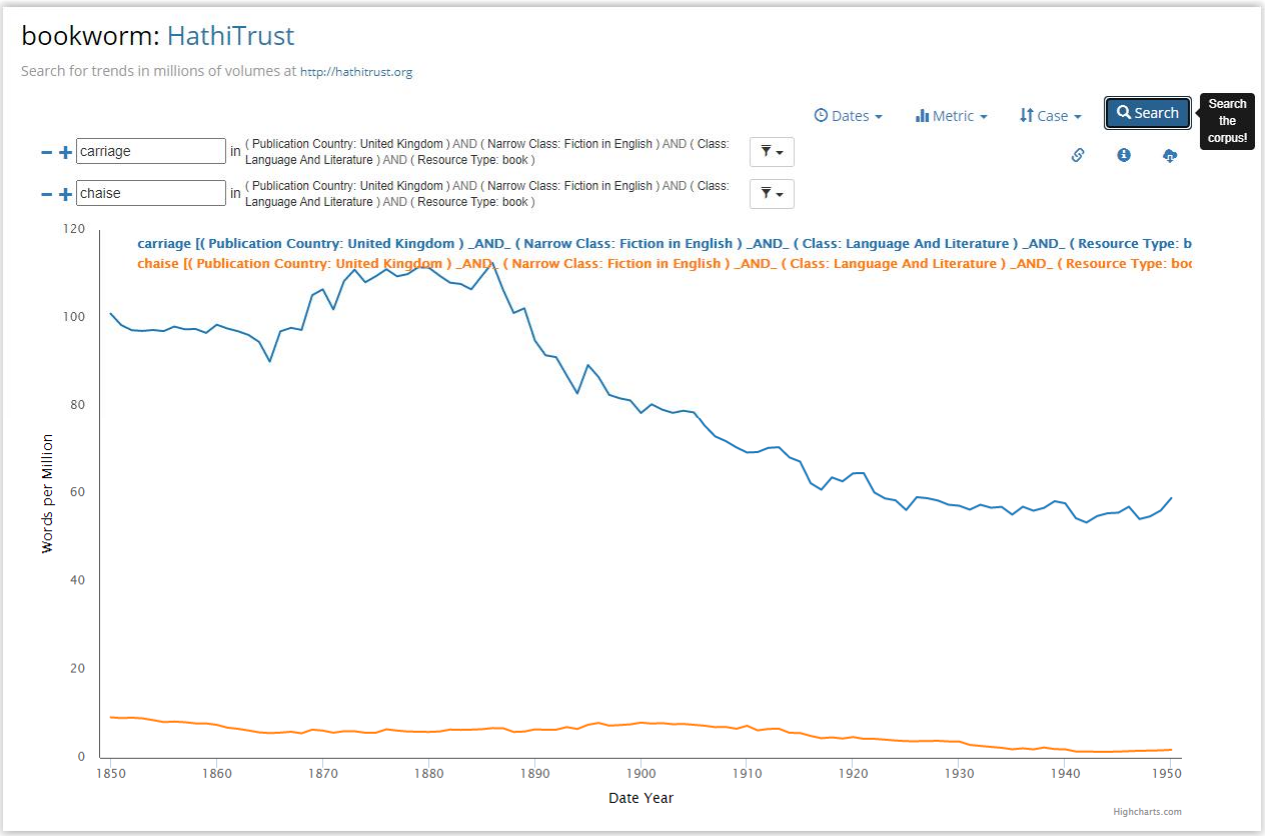


Figure 2: HathiTrust+Bookworm

- Interface design: The interface is clean and simple. A user can enter a search term in the input field and apply filters. More input fields can be added. An example search is available to demonstrate the use of the tool.
- Metadata: Only bibliographic data are available.
- Search options: Single word searches are allowed, but not searches for phrases. No truncation or other commands are available.
- Filtering: There are 19 bibliographic filters that can be applied to each search term. There is an additional filter for date that can be applied to the entire search.

Search results: The frequency counts are shown on a graph. It is possible to click on a point on the graph and see a list of results for that point. The items in the list of results link to the texts in the digital library. However, it only links to the volume, not the term(s) in context.

Perseus Digital Library

The Perseus Digital Library) was specifically developed to explore the possibilities of digital collections. The focus was originally on Greek and Latin material, but the collection has expanded. Figure 3 shows the results of a search for all forms of the Latin word *suis*, limited to the Greek and Latin collection.

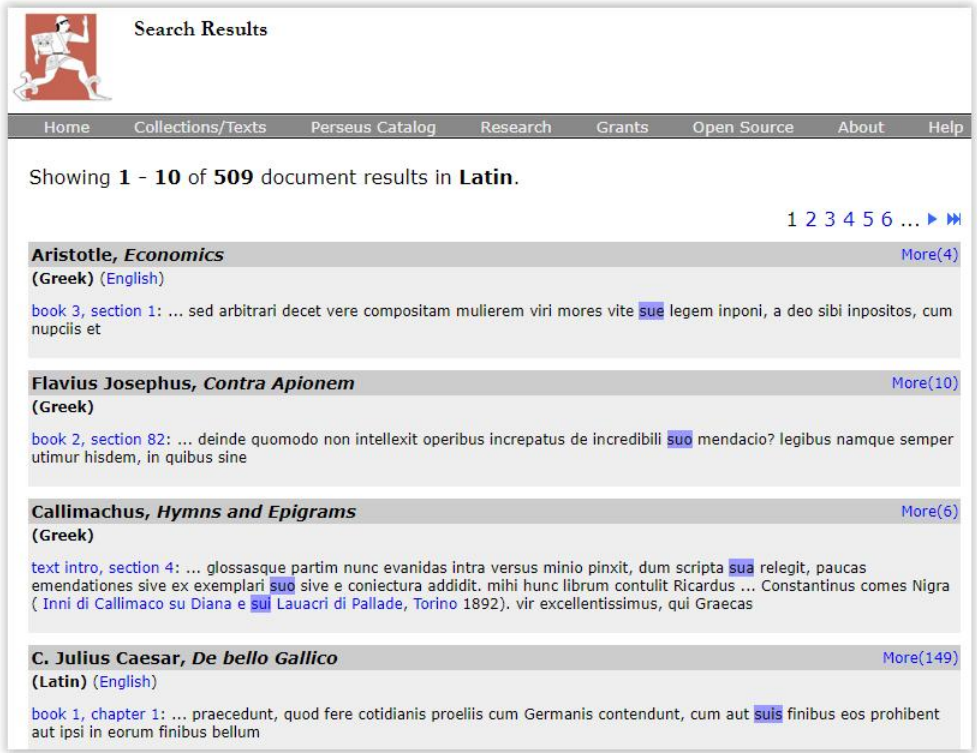


Figure 3: Perseus Project

- Interface design: The interface is fairly simple. Descriptive labels are used for search options. However, it is presumed that the user has some knowledge of the type of materials in the collection.
- Metadata: The texts in the collection are encoded with XML to indicate structures in the text, for example paragraphs. The texts are parsed to enable searching for inflected forms.

- Search options: The general search options include searching for the occurrences of words in texts (as demonstrated in figure 3) or to search for more information about a single word. A user can also select a word as it appears in a text and link to more information about that word. Other search options are also available.
- Filtering: Results can be limited to specific collections.
- Search results: The type of results depends on the type of search. Either a list of instances where a word appears in context is returned, or more information about a word is returned.

Voyant Tools

Voyant Tools) is a free, online tool for text analysis. Texts to be analysed are added by the user. The main page contains several panels, each with a different tool useful for text analysis (for example to view trends). Figure 4 shows a section of the tool.

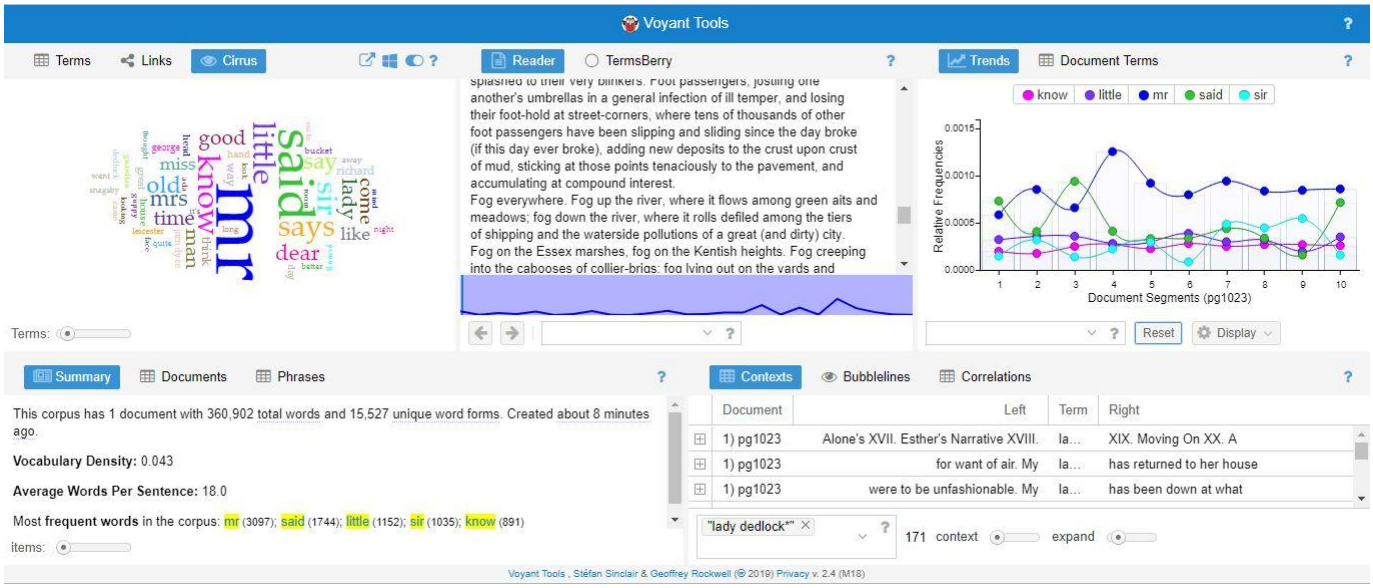


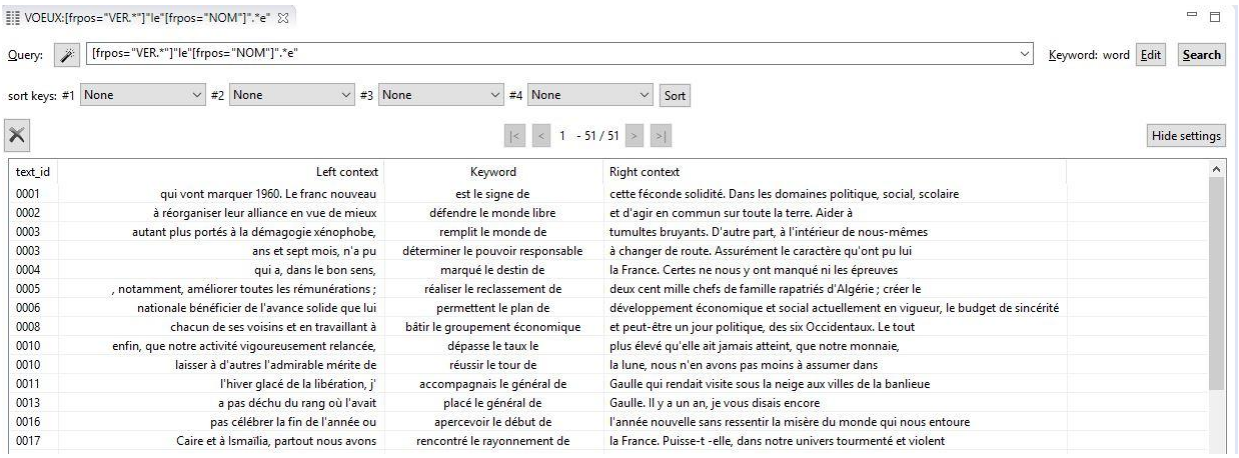
Figure 4: Voyant Tools

- Interface design: Voyant tools has a user-friendly interface, which consists of one platform with different tools.
- Metadata: There are no specific metadata. Texts encoded in XML may be uploaded.
- Search options: Some tools on the platform have a search option, where a user can search for a single word or a phrase. Some features are available to construct a query, for example, wildcard and proximity operators. Integration with XML is possible and XPath expressions can be used to select sections of an encoded texts when importing the texts.
- Filtering: If using XML, then some form of filtering is possible.
- Search results: Each tool will produce different results. Pertinent to this study is the tool that shows occurrences of words in a text in a graph and the tool that shows words in context.

TXM

TXM is a free text analysis environment. One of the main aims was to be compatible with encoded texts (for example, texts encoded with Text Encoding Initiative). It is a powerful tool

and it is beyond the scope of this study to offer an in-depth description of all the features of this tool. Specific features relevant to this study will be highlighted. Different corpora can be imported and studied. A query in the concordance is shown in figure 5. This example query searches for a sequence of four words, where the first word must be a verb, the second word must be *le*, the third word must be a noun and the last word in the sequence must end in an *-e*.



text_id	Left context	Keyword	Right context
0001	qui vont marquer 1960. Le franc nouveau	est le signe de	cette féconde solidité. Dans les domaines politique, social, scolaire
0002	à réorganiser leur alliance en vue de mieux	défendre le monde libre	et d'agir en commun sur toute la terre. Aider à
0003	autant plus portés à la démagogie xénophobe,	remplit le monde de	tumultes bruyants. D'autre part, à l'intérieur de nous-mêmes
0003	ans et sept mois, n'a pu	déterminer le pouvoir responsable	à changer de route. Assurément le caractère qu'ont pu lui
0004	qui a, dans le bon sens,	marqué le destin de	la France. Certes ne nous y ont manqué ni les épreuves
0005	, notamment, améliorer toutes les rémunérations ;	réaliser le reclassement de	deux cent mille chefs de famille rapatriés d'Algérie ; créer le
0006	nationale bénéficier de l'avance solide que lui	permettent le plan de	développement économique et social actuellement en vigueur, le budget de sincérité
0008	chacun de ses voisins et en travaillant à	bâtir le groupement économique	et peut-être un jour politique, des six Occidentaux. Le tout
0010	enfin, que notre activité vigoureusement relancée,	dépasse le taux le	plus élevé qu'elle ait jamais atteint, que notre monnaie,
0010	laisser à d'autres l'admirable mérite de	réussir le tour de	la lune, nous n'en avons pas moins à assumer dans
0011	l'hiver glacé de la libération, j'	accompagnais le général de	Gaulle qui rendait visite sous la neige aux villes de la banlieue
0013	a pas déchu du rang où l'avait	placé le général de	Gaulle. Il y a un an, je vous disais encore
0016	pas célébrer la fin de l'année ou	apercevoir le début de	l'année nouvelle sans ressentir la misère du monde qui nous entoure
0017	Caire et à Ismaïlia, partout nous avons	rencontré le rayonnement de	la France. Puisse-t -elle, dans notre univers tourmenté et violent

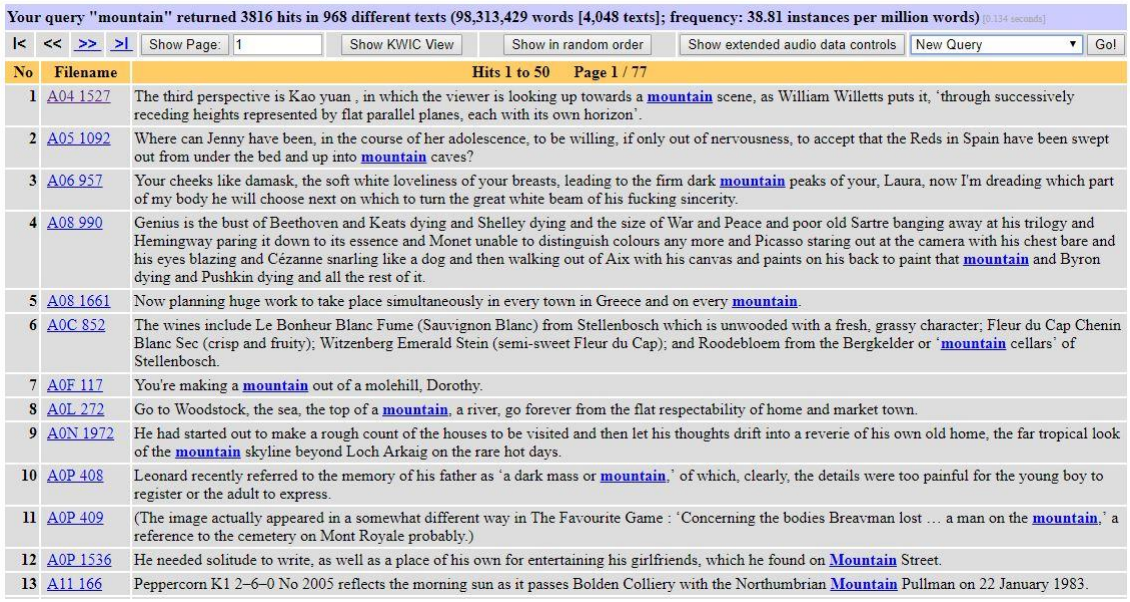
Figure 5: TXM

- Interface design: It is a powerful tool with numerous features, and consequently the tool does not have a simple interface.
- Metadata: Various levels of metadata are available. Bibliographic metadata are included for each text in a corpus. The structure of the texts that are imported may be encoded. Words may be annotated with morphological data.
- Search options: Queries may be written in a query language, allowing for powerful queries. Truncation and wildcard characters can be used. It is possible to search for lemmas, part-of-speech categories, single words or phrases. It is also possible to search in textual structures, such as paragraphs, by using the tags used in the encoding. There are also query assistants which help a user to construct a query through using a graphical user interface.
- Filtering: Filtering is achieved through the search options.

Search results: Various types of results can be retrieved. The concordance lists the retrieved instances in context. It is possible to link to more context.

BNCweb (CQP-Edition)

The BNCweb (CQP-Edition)) is one of the tools that are used to explore the British National Corpus (BNC). The interface is simple, while allowing users to enter complex queries based on a corpus query language. The results of a search are shown in Figure 6.



Your query "mountain" returned 3816 hits in 968 different texts (98,313,429 words [4,048 texts]; frequency: 38.81 instances per million words) [0.134 seconds]

I< << >> >I 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 / 77
1	A04 1527	The third perspective is Kao yuan , in which the viewer is looking up towards a **mountain** scene, as William Willetts puts it, 'through successively receding heights represented by flat parallel planes, each with its own horizon'.
2	A05 1092	Where can Jenny have been, in the course of her adolescence, to be willing, if only out of nervousness, to accept that the Reds in Spain have been swept out from under the bed and up into **mountain** caves?
3	A06 957	Your cheeks like damask, the soft white loveliness of your breasts, leading to the firm dark **mountain** peaks of your, Laura, now I'm dreading which part of my body he will choose next on which to turn the great white beam of his fucking sincerity.
4	A08 990	Genius is the bust of Beethoven and Keats dying and Shelley dying and the size of War and Peace and poor old Sartre banging away at his trilogy and Hemingway paring it down to its essence and Monet unable to distinguish colours any more and Picasso staring out at the camera with his chest bare and his eyes blazing and Cézanne snarling like a dog and then walking out of Aix with his canvas and paints on his back to paint that **mountain** and Byron dying and Pushkin dying and all the rest of it.
5	A08 1661	Now planning huge work to take place simultaneously in every town in Greece and on every **mountain**.
6	A0C 852	The wines include Le Bonheur Blanc Fume (Sauvignon Blanc) from Stellenbosch which is unwooded with a fresh, grassy character; Fleur du Cap Chenin Blanc Sec (crisp and fruity); Witzenberg Emerald Stein (semi-sweet Fleur du Cap); and Roodebloem from the Bergkelder or '**mountain** cellars' of Stellenbosch.
7	A0F 117	You're making a **mountain** out of a molehill, Dorothy.
8	A0L 272	Go to Woodstock, the sea, the top of a **mountain**, a river, go forever from the flat respectability of home and market town.
9	A0N 1972	He had started out to make a rough count of the houses to be visited and then let his thoughts drift into a reverie of his own old home, the far tropical look of the **mountain** skyline beyond Loch Arkaig on the rare hot days.
10	A0P 408	Leonard recently referred to the memory of his father as 'a dark mass or **mountain**,' of which, clearly, the details were too painful for the young boy to register or the adult to express.
11	A0P 409	(The image actually appeared in a somewhat different way in The Favourite Game : 'Concerning the bodies Breavman lost … a man on the **mountain**,' a reference to the cemetery on Mont Royale probably.)
12	A0P 1536	He needed solitude to write, as well as a place of his own for entertaining his girlfriends, which he found on **Mountain** Street.
13	A11 166	Peppercorn K1 2–6–0 No 2005 reflects the morning sun as it passes Bolden Colliery with the Northumbrian **Mountain** Pullman on 22 January 1983.

Figure 6: BNCweb (CQP-edition)

- Interface design: The interface is simple and intuitive.
- Metadata: The data (BNC XML edition) include grammatical, structural as well as bibliographic data.
- Search options: Various advanced search options are available. For example, there are wildcard characters to search for patterns or variations of words, one can search for a word form with a certain part-of-speech tag, lemmas, word sequences and within certain text structures that were encoded.
- Filtering: Filtering is allowed according to the available bibliographic data.

Search results: The default option is to display retrieved instances in context. It is possible to link to more context and see information about the text the instance is from. Other display options are available.

Conclusion

There is increasing interest in using large digital text collections for research purposes. Several tools have been developed to allow a user to explore such data. These tools allow researchers to do more than what is possible using a general search engine. There are options to visualise trends, make selections based on morphological or syntactic properties, retrieve sections of encoded data or filter according to bibliographic data. However, there are several limitations in the tools currently available. Tools with simpler interfaces do not allow users to filter on a detailed level. Some tools allow very specific filtering but are complex. None of the tools in this study allows a researcher with little prior training or knowledge of encoding (such as XML), to search for words (or phrases) with specific properties in a large collection of texts and ideally view the usage frequency of the word(s) over time. More work should be done to determine what metadata can be used to encode texts to enable meaningful retrieval, and how a tool can use such metadata to support retrieval in an easy and intuitive manner.

About the author

Liezl Ball is a lecturer in the Department of Information Science at the University of Pretoria, South Africa. She completed her master's degree in 2016 and is currently working on her PhD. 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 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

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

Howard, J. (2017). What happened to Google's effort to scan millions of university library books? *EdSurge*. https://www.edsurge.com/news/2017-08-10-what-happened-to-google-s-effort-to-scan-millions-of-university-library-books (Archived by the Internet Archive at https://web.archive.org/web/20200707104732/https://www.edsurge.com/news/2017-08-10-what-happened-to-google-s-effort-to-scan-millions-of-university-library-books)

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. http://dx.doi.org/10.1093/llc/fqv037

Michel, J-B., 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 Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182. http://dx.doi.org/10.1126/science.1199644

Nicholson, B. (2013). The digital turn: exploring the methodological possibilities of digital newspaper archives. *Media History, 19*(1), 59-73. http://dx.doi.org/10.1080/13688804.2012.752963

Ophir, S. (2016). Big data for the humanities using Google Ngrams: discovering hidden patterns of conceptual trends. *First Monday, 21*(7). http://dx.doi.org/10.5210/fm.v21i7.5567

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/ (Archived by the Internet Archive at https://web.archive.org/web/20200712070233/http://www.ariadne.ac.uk/issue/25/rydberg-cox/)