Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge

palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-020-0423-6 OPEN

Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge

Saturnino Luz [1]✉ & Shane Sheehan[1]

ABSTRACT An approach to establishing requirements and developing visualization tools for scholarly work is presented which involves, iteratively: reviewing published methodology, in situ observation of scholars at work, software prototyping, analysis of scholarly output produced with the support of text visualization software, and interviews with users. This approach is embodied by the software co-designed by researchers working on the Genealogies of Knowledge project. This paper describes our co-design methodology and the resulting software, presenting case studies demonstrating its use in test analyses, and discussing methodological implications in the context of the Genealogies of Knowledge corpus-based approach to the study of medical, scientific, and political concepts.

[1] Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK. ✉email: s.luz@ed.ac.uk

Introduction

The analysis of corpora has always been of central importance in the humanities. More recently, the spread of computing technology and the consolidation of the field of digital humanities has transformed the way corpus analysis is done. The use of computational tools has a relatively long tradition in the disciplines of lexicography and corpus linguistics (Svartvik, 2011), and the roots of the widely used Keyword-In-Context (KWIC) technique can be traced back at least to the 1950s, starting with the work of Luhn on concordance indexing (Luhn, 1960). This has been an extremely productive relationship, influencing many other areas of investigation in the humanities (Frank et al., 2018), including corpus-based translation studies (Baker, 1993a; Bernardini and Kenny, 2020).

As explained in the introduction to this special issue, the focus of the Genealogies of Knowledge (GoK) project is on exploring the role of translation and other sites of mediation in shaping the historical evolution of scientific and political concepts. A noteworthy aspect of the project is that it explores these issues through the methodological lens of concordance and collocation analysis, an approach that is strongly influenced by the work of the British linguists J.R. Firth, John Sinclair and Michael Halliday (Léon, 2007; Sinclair, 1991), and shaped by the use of computational tools. Scholarly work in this field of study traditionally proceeds in a "bottom-up" manner. Selected texts are read and analysed by scholars, and synthesis often relies on the investigator's memory and powers of abstraction, as well as their theoretical framework. The use of corpus-based methods can radically change this mode of work. Corpus analysis suggests a "top-down" approach where one usually starts by obtaining an overview of the data and exploring a much larger volume of text than would be practical to do by means of eye and hand alone. This leads to an iterative process in which the investigator switches between overview and detail towards analysis and generalization. Visualization tools can aid this process by providing effective overviews and drawing the researcher's attention to patterns that might otherwise go unnoticed, as well as serving as vehicle for visual explanations (Tufte, 1990).

All modern corpus-based studies of text in the Firthian tradition involve, minimally, computational support for term indexing, search, retrieval and display. However, despite these commonalities, different fields and studies often need to adapt old methods and develop new ones to suit their particular analytical needs. While this work of adapting and developing becomes part of the study's methodology and will affect the research outcomes, this process is rarely documented or discussed at a theoretical level. Taking a broader methodological view of tool development by regarding this activity as a part of the conceptual framework of the corpus-based studies tradition that informs GoK is particularly important, as the project adapts a well established linguistics methodology to the study of spatiotemporal evolution of medical, scientific and political concepts.

Our goal in this paper is to document and discuss the ongoing process of co-design and development of text visualization tools to support the corpus-based investigations conducted as part of the GoK project. In doing so, we hope to establish general methods for the development of such tools in interdisciplinary contexts. We envision this as a first step towards the more ambitious goal of creating the basis for a truly interdisciplinary methodology for scholarly work that breaks the barriers between "developers" and "users" of tools, and that welcomes equally the contributions of interactive systems designers, corpus researchers and humanities scholars.

We start by presenting the development methodology and the design rationale for the software tools developed for the GoK project, covering the steps of methodology review, requirements elicitation, observation of scholarly work, and prototyping activities. We then describe the GoK tools proper, as they exist today, present case studies illustrating the tools in use, and discuss the results and implications of this methodological approach.

Related work

The history of digital humanities dates back to the 1940s when Roberto Busa began work on *Index Thomisticus*, the first tool for text search in a very large corpus. Visualization tools in digital humanities mostly focused on close reading techniques for investigating individual texts until the explosion of interest in distance reading techniques triggered by Moretti's "Graphs, maps, trees" in 2005 (Moretti, 2005).

While there are many visualization techniques and systems developed for digital humanities (Jänicke et al., 2015) concordance-based visualization is quite rare, that is to say you do not often find visualizations which encode lexical and grammatical patterns of co-occurrence around a keyword in digital humanities literature. These co-occurrence patterns are used extensively in related fields such as translation studies and corpus linguistics upon which the foundations of the GoK project stand.

The practitioners of corpus linguistics range across many diverse disciplines in the study of language. For example, McEnery and Wilson (2001) introduce corpus linguistics by covering topics such as: lexical studies, grammar, semantics pragmatics and discourse analysis, sociolinguistics, stylistics and text linguistics, historical linguistics, dialectology and variation studies and psycholinguistics, teaching of languages and linguistics, cultural studies and social psychology. While the exact methodology differs in each case the use of computer generated quantitative information, the investigation of lexical or grammatical patterns in the corpus and the qualitative discussion of the quantitative and textual information are strong components of the techniques.

Corpus linguistics quantitative techniques employ frequency and statistical analysis to collect evidence of language structure or usage. Many of the methods can be thought of as empirical linguist techniques. However, a common misconception is that corpus-based approaches are entirely quantitative and do not require any qualitative input (Baker, 2006). Biber et al. (1998) present a collection of quantitative corpus-based methods, in each case "... a great deal of space is devoted to explanation, exemplification, and interpretation of the patterns found in quantitative analyses. The goal of corpus-based investigations is not simply to report quantitative findings but to explore the importance of these findings for learning about the patterns of language use". These qualitative interpretations are important as "... a crucial part of the corpus-based approach is going beyond the quantitative patterns to propose functional [qualitative] interpretations explaining why the patterns exist". In a linguistic study, before the application of quantitative techniques the formulation of hypothesis and research questions is often informed by qualitative analysis and/or prior knowledge of the texts under investigation. A good quantitative study must be preceded by a qualitative approach if anything beyond a simple description of the statistical properties of the corpus is to be achieved (Schmied, 1993).

Translation studies is a field where corpus-based methods have grown in popularity. Baker's early advocacy for the use of corpus-based methods in the study of translation (Baker, 1993b) has led to its adoption in various sub-fields of translation (Baker, 1995; Olohan, 2002; Rabadán et al., 2009; Zanettin, 2001, 2013). The re-emergence of corpus-based methods had a transformative effect, and the corpus-based methodology has been described as one of

the most important gate-openers to progress in translation studies (Hareide and Hofland, 2012).

Concordance analysis is a core activity of scholars in a number of humanities disciplines, including corpus linguistics, classical studies, and translation studies, to name a few. Through the advent of technology and the ever increasing availability of textual data this type of structured analysis of text has grown in importance (Bonelli, 2010; Sinclair, 1991). Some of the most popular tools which have concordance browsing at their core include *WordSmith Tools* (Scott, 2008), *SketchEngine* (Kilgarriff et al., 2004) and *AntConc* (Anthony, 2004). While there is some variation in advanced features across the range of concordance browsers each provides a windowed concordance which can be explored (via scrolling or multiple pages) and is usually sortable at word positions. This simple feature set is the key to supporting the traditional corpus linguistic methodology of concordance analysis.

As computational tools and methods for concordance and collocation analysis are central to the GoK research programme of extending the methodology of corpus-based translation studies, we focus on the tasks to be supported in this domain. We will return to a review of tools for corpus analysis and visualization in section "Analysis of existing visualizations", once we have defined and conceptualized the tasks involved in greater detail.

Iterative co-design for corpus-based scholarly work

The process of development for the GoK visualization tools involved various steps, including an analysis of the published methods in corpus linguistics on which the GoK project draws, low-fidelity level prototyping and requirements elicitation, high-fidelity prototyping and formative evaluation of these prototypes in use. These steps were not usually performed sequentially, but rather iteratively: progress made by means of one activity often informed our approach at other stages in the process, whilst simultaneously reflecting knowledge and techniques gained during other iteration cycles. In what follows, however, these stages must be shown sequentially, as cohesive blocks, for presentation purposes. Cross connections are indicated as needed, and the reader is warned that some of these are forward references.

Analysis of published methodology. The work of Sinclair in corpus linguistics (Sinclair, 1991, 2003) and Baker and others in translation studies (Baker, 1993a) are the main theoretical influences guiding the GoK methodology. Published methodology in this area is therefore a natural starting point for the identification of aspects of analytical work that can be supported by computational tools. In the case of corpus-based analysis, we were fortunate to be able to rely on the work of John Sinclair, who not only developed the foundations of a method for linguistic analysis which has subsequently influenced a number of research programmes, including GoK, but who also described this method in a detailed tutorial form (Sinclair, 1991, 2003).

In his book *Reading Concordances*, Sinclair (2003) presents 18 tasks. Each task guides the reader through an analysis, describing both the mechanics and analysis required to complete the task. Although a computational element is implied, the tasks described in the book are based on given sets of printed and pre-formatted concordances. For each of these tasks we performed a hierarchical task analysis (Annett, 2003; Newman and Lamming, 1995) by combining or splitting the steps into a series of actions and sub-actions. The goal of such an analysis is to identify the low level actions (for example, sorting a list of words) required to complete higher level tasks (e.g. finding a significant collocate), and to order them in terms of importance or frequency. Using the completed task analysis to detect those actions which are not well supported by current tools or techniques can lead to rational design choices.

Each task was tagged to assist with classification and counting of the actions and sub-actions. We add tags to each task in an iterative process, the first stage of which involved the tagging of each analysis step with potential action tags. On completion of this initial pass, the tags are reviewed for relevance and redundancy, and are collapsed into more generalized actions where relevant. A similar review cycle is performed at the level of the eighteen tasks presented by Sinclair to homogenize the actions across tasks.

As an example let us look at the tags which were applied to the first instruction in the second of Sinclair's tasks. This task focuses on regularity and variation in phrase usage. The first step in the task description asks the reader to use a supplied concordance to "Make a list of the repeated words that occur immediately to the left of gamut. Sort them in frequency order. Then make a similar list of the words immediately to the right of gamut. Ignore single occurrences at present". As this step involves listing all words in frequency order at a position the tags *frequency, word position* are applied. We choose not to apply the *estimate frequency* tag as that would apply if the reader were asked to find the most frequent word at a position where the actual counts would not matter. We are also not looking for *frequent patterns* as we only look at a single position for a single keyword.

This tagging procedure can allow a visualization researcher with limited knowledge of the problem domain to extract meaningful actions. However, this is a subjective process and additional efforts from other perspectives could yield interesting differences or similarities in the action hierarchy and counts.

While most of the tags we used to analyse the 18 tasks represent actions, a few additional tags were chosen to help clarify and add information about the actions and sub-actions. The purely clarifying tags are omitted from the analysis of tag frequency. Examples of these are the tags *expert knowledge*, *combinations* and others are not themselves actions, but are useful in clarifying the objective or operation of the sub-actions. These clarifying tags always appear with other primary tags. The part of speech (*POS*) tag is both a primary action tag and a clarifying tag. The POS primary action is to determine the POS of a word occurrence. The POS clarifying tag represents the use of POS information in another action.

We recorded the distribution of the tags according to the number of tasks in which each appeared and the total number of actions which received the tag as shown in Table 1. At a high level, this table tells us that reading concordance lines (CLs) and

Table 1 Action counts from task analysis.

Tag	No. of tasks in which an action appears	Total action appearances
estimate frequency	16	34
read context	16	31
frequent patterns	15	21
frequency	14	18
word position	13	24
POS: Part of speech	11	23
filter	11	18
sense	10	19
group	7	9
significant collocate	5	7
usage	5	6
phrase	5	6

Number of tasks which feature the action, out of 18 tasks, and total numbers of actions found in these 18 tasks.

actions which require positional statistics are necessary for the style of concordance analysis outlined by Sinclair (2003). Looking in more depth we see that reading the context of a concordance and having an expert classify some linguistic property is a very common action. Estimating word frequency and calculating exact frequencies are also very common tasks which often combine with the position tag.

The actions and sub-action tags generalize the descriptive analysis steps into operations which are common to many of the tasks. In this way, we try to generalize the actions required to analyse a concordance. The results of this process are discussed below.

At the top level of the hierarchy Fig. 1, the highest level actions are defined. These actions are distinct pieces of analysis which form part of the analysis tasks presented by Sinclair (2003). These high level actions are use cases of concordance analysis, such as attempts to investigate meaning, identify phrases or explore word usage properties. We identified these actions during the tagging process by investigating the clusters and order of the tags attached to each task.

Under each of the top level actions we identify sub-actions which contribute to the completion of the top level goals. For example, frequent patterns must be identified before they can be classified as phrasal or non-phrasal usage (Sinclair, 1991). These sub-actions are given in the order they were observed for the top level action. The sub-actions may need to be completed in the order they were observed to successfully complete the top level task; this is often the case with the first sub-action, such as those listed for *Identify Frequent Patterns* and *Identify Phrases*. The second level tasks for the top level *Investigate Usage* action, on the other hand, do not need to be performed in order as they represent the sub-actions required to analyse separate forms of usage. It should also be noted that there is a lot of repetition across the second level sub-actions for each top level task. *Identifying frequent patterns* is an example of a sub-action which appears under every high level action. The type of pattern under

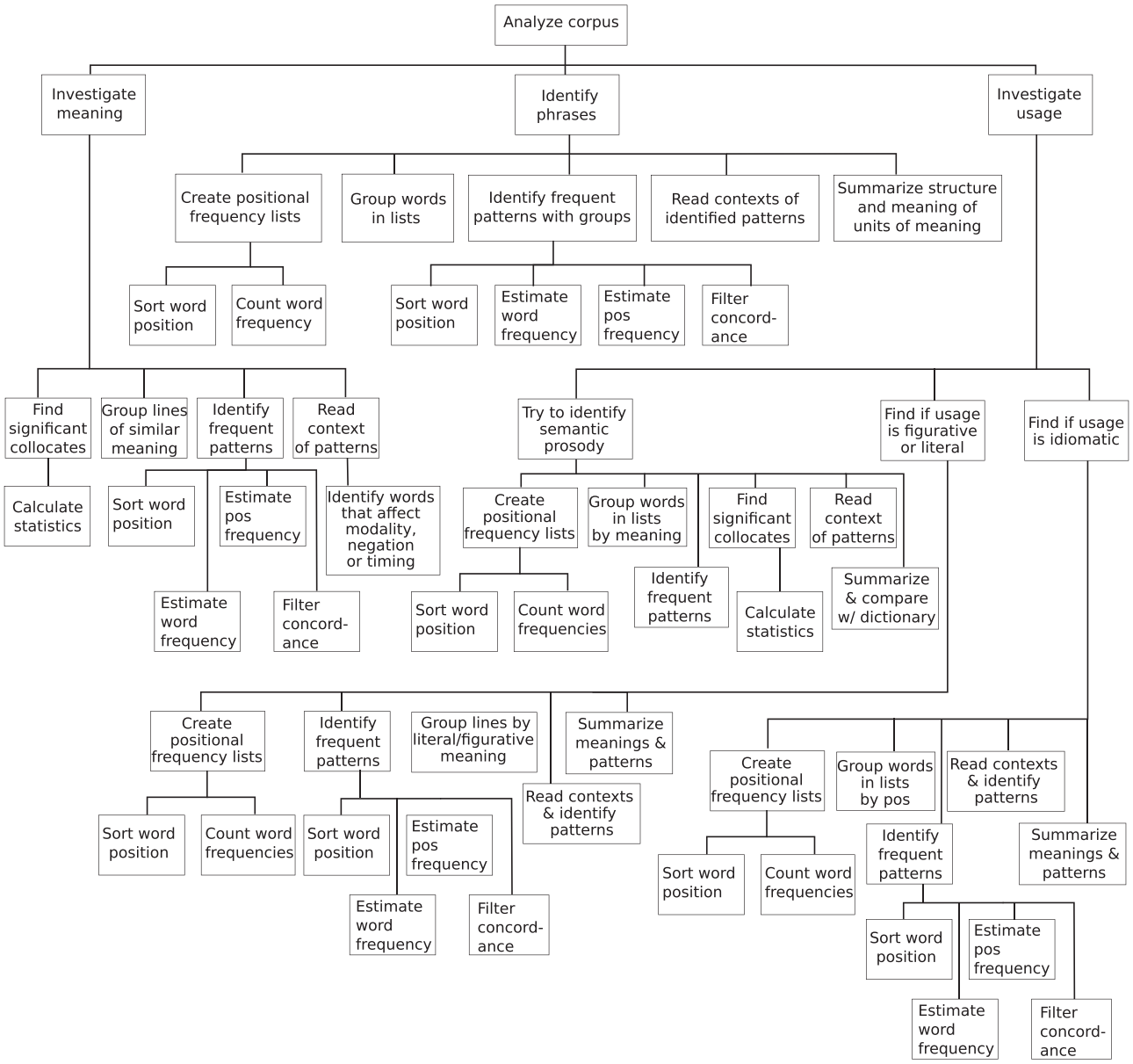

Fig. 1 Hierarchical task diagram. Tasks and action hierarchy for concordance-based corpus analysis.

investigation changes with the task but the mechanics remain relatively unchanged.

As we progress further into the hierarchy we find increasingly lower level sub-actions. These sub-actions describe the mechanics of the analysis. *Sort Word Position*, *Count Word Frequencies* and *Filter Concordance* are very specific low level sub-actions that are easy to identify and perform. The *Read Contexts* mechanics are simple but the purpose of the reading is usually to gain understanding or insight. The mechanics are not enough, an analyst or expert is required for the interpretation of what is read.

All of these actions and sub-actions should be considered when developing tools for research in this domain. If possible, they should all be supported and made more efficient and user friendly or even automated to reduce the workload for the corpus analyst. However, the tag and task analysis presented here would suggest most pressingly that, in order to support the method outlined by Sinclair (2003) tools should allow close examination of individual CLs, while also providing support for analyzing positional frequencies and collocation patterns.

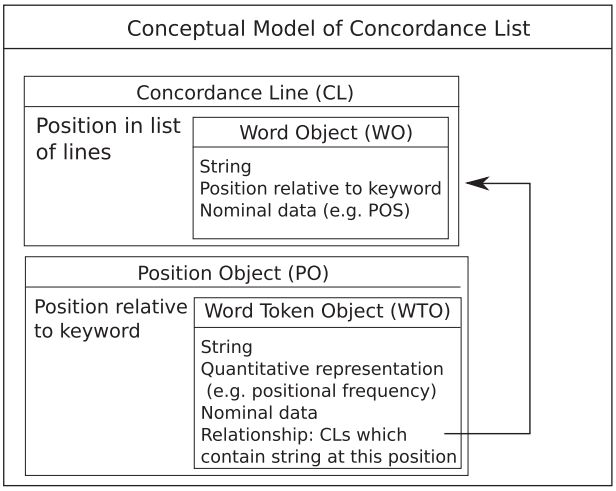

Fig. 2 Diagram representing the conceptual model of concordance lists. Conceptual data model of concordance lists.

Conceptual data model of KWIC. To formalize the data structures, attributes and relationships inherent to the concordance list as revealed by the analysis above, a conceptual model of the KWIC concordance list has been created so that visualizations can be evaluated and designed for in terms of their effectiveness at representing the model. The design of the model seeks to structure the data entities in a manner which best supports the actions described in the task and action hierarchy Fig. 1. Our data model is simple and a natural extension of the task analysis, it is an abstraction of the concordance list which identifies the data attributes required for the identified tasks and actions. Creation of data abstractions by qualitatively analyzing the output of domain characterization effort is typical of good visualization design methods such as the Munzner's Nested Model (Munzner, 2009).

The traditional rendering of a KWIC concordance list evokes a conceptual model consisting of a list of aligned sentence fragments (CLs). In this model each CL has an attribute representing its position in the list (concordance lists are usually presented in alphabetical order) and contains an ordered set of word objects (WO) which make up the string that represents the CL. The WO represent an individual occurrence of a word in a CL and contain its string representation (nominal data), its position relative to the keyword and any other nominal variables (meta-data) available e.g. POS tags.

Many of the actions identified in the task analysis require reading of the CLs ("read context"). In order to read the context the text fragments must be available. Since the linear structure (sentence structure) of the CLs is emphasised in the CL model it is included in the KWIC conceptual model to facilitate this read context action.

Since the WO in the CL model are representative of a single occurrence of a word they do not have as an attribute a quantitative variable such as word frequency. The frequency values are available by counting similar WO within all CLs but the frequencies are not attributes of any entities in the model. We would like our KWIC model to contain these quantitative variables as attributes of some entity since "Estimate Word Frequency", "Count Word Frequencies" and "Estimate POS frequency" are needed in each of the three identified core tasks. We also found that word position was often required in conjunction with these frequency orientated actions. For instance, estimating word frequency at a position relative to the keyword is a common action required for analysis showing up in 13 of Sinclair's 18 tasks.

With this in mind we now conceptualize the concordance lists as an ordered set of position objects (PO). Within each position object there is an attribute for the position relative to the keyword and a set of word token objects (WTO). These WTO differ from the WO in the CL model in several ways. The most important way they differ is that these objects represent all occurrences of a string (or string and nominal variable) at the position in which they reside. That is to say, there will be at most one object with a particular string and nominal variable (meta-data) combination. For example, if POS tags are available there will be one object representing the noun "date" and one representing the verb "date". Each WTO inherits its position as an attribute. Quantitative attributes which represent positional count, frequency or other statistics of the WTO in the KWIC are included after calculation. Finally, an attribute (relationship) which maps each WTO to the CLs in which it occurs is also available. This attribute and the position attribute of the CL WO provide a link between the models unifying the KWIC conceptual model as seen in Fig. 2. This linking of the conceptual models is especially useful for the frequent patterns action, where word combination frequencies between/across positions are required.

Analysis of existing visualizations. Text visualization encompasses many different visual methods. We are mainly interested in comparing visualizations which have been designed for keyword search results represented as a concordance list. However, we also wish to investigate techniques which, while not designed with the concordance in mind, could potentially have applications in concordance analysis.

Both the task analysis and conceptual model described above have identified a split in KWIC tool requirements, with each task requiring some combination of qualitative and quantitative actions. Qualitative actions most often operate on the concept of CLs where readability and the linear structure of the lines are emphasised. While the quantitative actions require positional statistics, they do not often require readability of the individual CLs.

Looking at the conceptual model, task analysis and action counts, we conclude that for qualitative primary actions the most important attribute to present is CL word order, so that the fragments are readable. By choosing an appropriate encoding of CL word order visualizations can aid qualitative concordance analysis.

The quantitative actions often require estimation of frequencies/statistics, or finding exact word frequencies/statistics. These quantitative variables are then explored using positional analysis/estimation of these statistics to identify linguistic patterns. Clearly then, one should prioritize the attributes PO Position (in which position relative to the keyword the word occurs), the qualitative attribute of WTO (how often the word occurs at each position) and to a lesser extent the PO WTO position attribute (which represents the quantitative ordering of the words at a position; i.e the order of a frequency list).

One approach to evaluating the effectiveness of visual encoding is a qualitative discussion of images or video of a visualization system. In what follows, we present a qualitative discussion of visualization systems related to concordance analysis in a semi-structured manner. Paraphrasing Munzner (2009), while visualization experts may draw the same conclusions from the inspection of a system, the validation is strongest when there is an explicit discussion pointing out the desirable properties in the results.

A ranking of visual variables (Bertin, 1983) per data type was proposed by Mackinlay (1986). This ranking of variables is in agreement with a ranking proposed for quantitative data by Cleveland and McGill (1985). Mackinlay's ranking of visual variables for the three data categories (ordinal, nominal and quantitative) are useful to help guide variable choice. If a variable is chosen for a data attribute instead of a higher ranked variable, justification should be given. Often visual encodings are presented without such justification leaving the reader to guess at the authors reasoning (Green, 1998).

We go about evaluating related text visualization methods for concordance analysis by mapping each attribute of our conceptual model (section "Establishing initial requirements") to the visual variables used. We make note of the number of visual variables mapped to the attribute and the ranking of the visual variables for the data type it is representing. The attributes of the KWIC conceptual model are further expanded by categorizing them as nominal, ordinal or quantitative data types. This helps us to apply Mackinlay's rankings which rely on these three data categories.

To begin with, we will evaluate the most widely used concordance visualization, the traditional KWIC visualization (or concordance list). CL position is represented by the vertical position of its enclosed strings. These enclosed strings are rendered horizontally, left to right, in the order they appear in the text fragments. Both CL and WO are mapped to the best visual variables (*position*) for their ordinal position data types. Because of this, we expect it will be easy to identify individual CLs and find where they rank in the chosen ordering scheme (usually alphabetically by a selected or default word position). Similarly, identifying the WO in the order they appear in the text fragments will not be difficult. The concept of a position object can be loosely applied in this visualization; the word positions across CLs can be identified using horizontal position, even though this is made difficult due to the variable length of words. A visual variable that is associative, such as *color hue*, can be encoded to identify words at the same position.

Since the KWIC display is designed with the readability of CLs in mind, the inability to gain an overview of a large concordance is a necessary trade-off. In this rendering the detail is presented at all times, an overview of the entire concordance list is only available for concordances which fit within the screen at a readable font size. Windowing the concordance and scrolling is the usual solution, this works well for viewing individual CLs, but to get an overview of the positional frequencies and patterns a higher level view would be better. Larger screen sizes and higher resolutions can improve the situation but as more data becomes available the scale required becomes impractical.

As this visualization contains no explicit representation of the WTO as such, we expect visual assessment of exact or estimated word frequency to be difficult. Nevertheless, this visualization is the most commonly used tool in concordance analysis where, as we have shown, positional frequencies are regularly required. From our observational study (section "Observational research") and task analysis we found that counting the strings is the usual way to calculate these positional word frequencies. The standard tools used by scholars in the field do not offer positional frequency lists or other tools to make position word frequencies easier to work with. While this visualization is very effective for reading CLs it would seem to be of limited use for quantitative concordance actions.

Several tree-based visualizations (Culy and Lyding, 2010; Luz and Sheehan, 2014; Wattenberg and Viégas, 2008) have been proposed to attempt to bridge the gap between qualitative and quantitative analysis. *Word Tree* (Wattenberg and Viégas, 2008) displays the keyword and either the left or right context, taking the familiar form of a tree structure, in which the keyword is displayed as the root and additional word vertices are connected in text order to each other. The main benefits of this visualization are that the linear structure and readability of the CLs is maintained through the combination of the visual variables connection and horizontal position. Connection defines the word position by the number of edges from the root, and horizontal position per branch provides partial positional groups (ordered positions in a sub-tree). These positional groups allow the frequencies at a position along a branch to be estimated since the words are rendered proportionally in size of their sub-tree. While frequency in a sub-tree is easy to estimate, the frequency at a word position is less clear. Looking at word positions as they move away from the root (keyword) positional frequencies become increasingly difficult to estimate. This is because each branching point can contain words which occur in other branches, leading to the possibility of multiple occurrences of a word object at a position. So at a depth of one from the root each rendered word represents a positional WTO, but deeper into the tree each rendered word is a partial WTO which only represents each occurrence of a token at that position in the sub-tree. Although sorting by frequency is supported at the first position from the keyword, combinatorial explosion causes the estimation of frequency to become more difficult as we view positions deeper into the tree.

An additional problem with estimation of frequency (or other word statistics) using this visualization is that variation in word length causes the variable representing the quantitative information to be inconsistent. Font size, which can be equated with visual variable *area* is used to represent frequency in a branch. The square root scale used by Word Tree and other visualizations should make word area roughly proportional to frequency if not for word length variations. While it seems natural to include quantitative information about a word by scaling the font representing that word, it is worth noting that the visual variable area ranks fifth for the display of quantitative information under Mackinlay's ranking scheme and, additionally, variations of word length complicate the interpretation of the quantitative values. The other tree based visualizations when viewed through the lens of our conceptual model suffer from similar problems of positional branching and frequency representation. They do however solve a readability problem by displaying both left and right contexts simultaneously and connecting them so that a full CL can be read. *Double tree* is designed with different tasks and users in mind, ("linguist's task of exploratory search using linguistic information"), so it is not surprising that it does not map well to our model. In this representation, word position is strongly encoded by using an integral (visual variables which are

perceived together) combination of connection and horizontal position. This contrasts with the loose horizontal positioning in word three where only connection can be used to reliably derive word positions. Double Trees encode frequency using *color saturation*, the eighth ranking variable for quantitative data, combined with the previously discussed branching issues quantitative information is not strongly encoded in this design.

Other visualizations do not place great emphasis on the readability of the CLs. One such visualization is *interHist* (Lyding et al., 2014), a complementary visualization which is used to display quantitative information about its accompanying KWIC view. In this case, the visualization is rendered as stacked bars (rectangles) where height is used to display quantitative information (*length* is the second best variable for quantitative information). While this interface is designed for POS information and does not represent individual tokens or WTOs, it is not difficult to imagine a variant with these rectangles representing WTOs with no change to the visual representation. In *interHist* vertical positioning of the rectangles is not encoded with meaning, making it more difficult to perceive quantitative differences between the rectangles. However, for word frequency all bars will be the same height as each CL will contain the same number of tokens, and since there will be many more words than POS tags rendering the words using a color and a legend could become impractical.

Corpus Clouds (Culy and Lyding, 2011) is a frequency focused corpus exploration tool which consists of composite views of a corpus query. The main display is a word cloud, based on the tag cloud visualization (Viégas and Wattenberg, 2008), where the absolute frequencies of all words returned by a corpus query are displayed. These word clouds map this quantity to *area* using font-size, the limitations of which were previously discussed. This visual encoding does not translate well to our conceptual model since positional concordance frequencies are the quantity of interest, not global frequency lists. Another view in the interface presents a modified KWIC display. The modification is the addition of a small vertical bar, similar to a sparkline, beside each word token in the KWIC view. This makes use of the variable *length* for frequency information, but the effectiveness of the variable is reduced for several reasons. The main limitation is that the size of the bars is restricted, causing only large differences in frequency to be perceived easily. Additionally, comparisons between lines take place in both planes, vertically across CLs and horizontally within lines, again making it difficult to perceive small variations in frequency. The number of KWIC lines which can be displayed per screen is also practically limited if readability is to be maintained.

Structured Parallel Coordinates (Culy et al., 2011) is an application of the parallel coordinates visualization technique to different types of structured language data, one of which is a KWIC plus frequency visualization. This visualization places WTOs, rendered as text labels, on the parallel axis which represent ordered word positions. The CL structure is maintained using connecting lines between WTOs. Statistical information is then placed on additional axes and the connection between the position axes and the quantitative axes are used to express the desired quantitative attributes of the WTOs. An individual quantitative axis is required for each quantity and word position pair. As with all parallel coordinate visualizations the choice of axis orderings is important. In this case the choice was to order the word positions in concordance list order and create the statistical axes to the right of the collection of position axes. This positioning makes it difficult to follow connections from a word position to its related quantitative axis if there are many other axes in between. Additionally, comparing two word positions is perceptually difficult, as the user needs information from four axes for a comparison across two word positions for a single statistic, such as frequency. In Structured Parallel Coordinates the linear order of the sentences is partially maintained through connection. However, since the connected nodes are WTOs the actual sentences are lost and only the preceding and next connections are meaningful.

TagSpheres (Jänicke and Scheuermann, 2017) have a rendering which encodes keyword based co-occurrences as position aware word cloud. Word position is encoded using an integral combination of color and radial position from the central keyword. This creates a strong positional encoding relative to the keyword. The linear structure of the concordance is abandoned in this rendering. The layout attempts to render in close proximity the same token at different word positions, this can help with identifying patterns of frequent co-occurrence for a single word. However, multi-word co-occurrence patterns do not have a clear mechanism for their identification. The layout uses font size, *area*, to represent quantitative information, comparison of quantitative information is made somewhat difficult due to comparison of area between words which may not be positioned on the same horizontal or vertical axis, as well as the issues of encoding font size with non-uniform word lengths. Another visualization which aims at providing an overview of concordances is Fingerprint Matrices (Oelke et al., 2013) where words of interest can be repre sented as rows and columns of an adjacency matrix, with glyphs representing co-occurences across a document. While this allows high density of POs to be encoded, the matrix representation does not map well to the overall concordance list model.

It is also worth considering visualizations which were not explicitly designed with the concordance list in mind. Text visualizations which focus on summarizing or exploration texts can also be applied to a concordance list by viewing the concordance list as a document, instead of as a selection of fragments from source documents. We now discuss some of these visualizations and suggest possible modifications to better fit them to the KWIC model and action requirements. This discussion is not exhaustive, as its goal is to further illustrate visual tools for the text analyst and to suggest possible starting points for new research into concordance visualization, rather than general text exploration or summarization.

TextArc (Paley, 2002) is a radial layout of the sentences in a document, within which WTOs are placed at a location which represents the average position of that word in the document. Font size is often used to represent a quantitative value such as word frequency. In terms of visual variables area in the form of font size is used for quantitative information, Word positions are not identified but perhaps color could be used and reading of the lines is still possible as they are rendered as part of the visualization. While this visualization does not seem to be a good fit for the tasks we have identified, it could give visual insight into which documents a collocated word is mostly contained in. Alternatively by ordering the CLs from multiple documents according to the position in the documents spatial patterns of co-occurrence across documents could be observed.

It is easy to imagine extending *Phrase Nets* (van Ham et al., 2009) to be more positionally aware and using them to look at collocations within a concordance list instead of a document. For example instead of the typical use of PhraseNet to examine a pattern such as "*X* and *Y*" a PhraseNet could be built for the pattern "*happy* one position to left of *keyword*, Y in window 5 words", this visualization would give information about the collocations of the word happy within a five word window of the concordance list. Limitations of this visualization from a concordance analysis perspective include quantitative information representation by font size and, as given, lack of positional information.

Another word cloud based visualization *TagPies* (Jänicke et al., 2018) visualizes keywords and their co-occurrences in a radial layout. While not explicitly mentioned, this interface could be used to visualize the comparative co-occurrences of the same keyword per position by having a separate slice of the pie for each word position (PO). However this visualization would encode the WTO positions as slice shaped word clouds, with word position relative to the centre of a slice and the pie centre having some quantitative interpretation. This visualization, while interesting, does not map well onto our conceptual model.

Sketch Engine, the most popular commercial concordance software, also contains visualization capabilities. These visualizations generate a radial layout of similar words in a corpus. Currently no positional concordance-based visualization is supported (Kocincová et al., 2015).

Similarly, a popular text analysis framework in digital humanities, *Voyant tools* (Miller, 2018; Voyant, 2020) contains a number of visualization components. These components, again, do not address the issue of positional word frequencies relative to a keyword, mostly focusing on keyword frequency and distribution. The built in concordance view is not split by word position only displaying textual columns for the left context, keyword and right context.

Regarding the ranking of visual variables, there are visual variables that are more effective at representing quantity than area (which seems to be used quite often), namely: length, position, angle and slope. It must be noted, however, that domain or task specific requirements ultimately drive one's choice of variables; in this case it just happens that quantitative information segmented by word position is important for many of the fundamental tasks of concordance analysis, and should have high priority when creating an encoding.

Establishing initial requirements. When designing visualization tools it is important to involve expert users in the requirements gathering process. However, simply asking domain experts what they need is rarely productive. Instead, two researchers from the GoK project were asked to list 20 questions which they would like to be able to answer about a corpus. They were asked to list them in order of importance where possible and to make themselves available afterwards to discuss the lists. The request for the lists was made a week before a meeting to discuss the results. It is also important to note that as tool development took place alongside corpus gathering and the preliminary steps of analysis by the GoK researchers, the answers were influenced by the discussions that took place during this phase of the project, as well as by the use of early prototypes and other, general purpose, tools such as concordancers typically used in translation studies. The discussion was used to get more details about the challenges facing users when trying to answer the questions they have identified. This technique is an established method for requirements elicitation in human computer interaction (Marai, 2018) and is becoming more widespread for domain characterization in visualization design.

The list created by the first researcher, Henry Jones (HJ), used a very loose ranking system. The categories and questions follow the order in which they occurred to the researcher and as such may be implicitly correlated with question importance. The three categories identified were *keywords*, *collocational patterns* and *temporal spread*.

The first seven questions identified were all associated with the analysis of a keyword, as shown below:

Keywords

1. How many times is the chosen keyword used across all of the corpus texts as a whole?
2. Is the keyword used with more or less uniform frequency in each of the corpus texts individually or are there significant imbalances in the dispersion of the keyword?
3. Which specific corpus texts use a given keyword with proportionally greater frequency? And lesser? What patterns can we see if we rank the corpus texts by number of hits for this keyword?
4. Which linguistic-grammatical form(s) of the term/concept under investigation (e.g. singular vs. plural, forms suffixed with -ship, -like or -ly) is/are more common across the corpus as a whole?
5. To what extent are the relative proportions of these different word-forms the same or different within each of the corpus texts individually?
6. Are there other related keywords we might study in order to expand our investigation? Can the software suggest keywords that are important to these texts but which we might not otherwise have thought of?
7. If so, are the frequencies of these terms similar or different to the first keyword, both across all of the corpus texts as a whole and in each corpus text individually?

Questions 1 and 4 can easily be answered with a concordance query given that the concordancing tools available can already list all instances of a lexical item as it appears across a corpus of texts. Questions 2, 3 and 5, on the other hand, are not so easy to answer using the tools more commonly used for concordancing and collocation analysis, such as WordSmith tools (Scott, 2001), the Sketch Engine (Kilgarriff et al., 2014) and TEC (Luz, 2011). During discussions with the researcher, the use of spreadsheets to collect word frequencies from repeated concordance queries was offered as a potential solution. Manipulating the subcorpus selection interface to search each text individually or using the filenames displayed in the left hand column of a concordance were both seen as potentially useful techniques. These questions and the technical difficulties faced by the researcher in answering them were evidence of the need for greater metadata integration into concordance tools so that frequency per filename can be quickly estimated. In the remaining *Keyword* questions, 6 and 7, automated keyword suggestion was discussed as a potentially useful technology which could be added to the GoK suite of tools.

In the questions relating to *Collocational Patterns* (below) we found calculating or estimating frequency via the concordance to be part of the solution in six of the seven questions.

Collocational Patterns

1. What are the adjectives that most commonly modify the chosen keyword (LEFT +1) across all of the corpus texts as a whole?
2. What are the adjectives that most commonly modify the chosen keyword (LEFT +1) in each text individually?
3. Are there any adjectives that modify the chosen keyword significantly more frequently in one text when compared with the others?
4. Are these adjectives only used to describe this keyword or are they connected to other keywords in this text?
5. What verbs are most commonly associated with the keyword (normally, RIGHT +1, RIGHT +2) across all of the corpus texts as a whole?
6. What verbs are most commonly associated with the keyword (RIGHT +1, RIGHT +2) in each of the corpus texts individually?
7. Are there patterns of interest in any of the other word-positions relative to the keyword? For example, if the keyword is a label used to describe a particular kind of political agent, we might be interested to look at what

collective nouns are used to group and characterize these political agents (e.g. a mob of citizens, a tribe of politicians: LEFT +2).

The possibility of sorting a concordance according to the items occurring at different word positions, and filtering the concordance via the subcorpus selection interface were suggested as useful techniques for answering the questions. Again, the use of a spreadsheet for collecting the results of various searches was mentioned. Question 4, the remaining question, seemed to have ties to measures of collocation strength with a keyword. By calculating the collocation strength between the keyword and the adjective the user might get some measure related to how strongly they co-occur in comparison with other combinations. The user pointed out that these questions assume the keyword is a noun, but that they believed the tools required for analysis would not change substantially for words of other grammatical categories, due to the analyst's primary interest in general co-occurrence patterns. Consequently, it seems that any techniques which might make the collection of collocation frequency information more efficient would be beneficial.

It is also worth noting that this analysis was conducted on the English-language subcorpus of GoK. Most of the questions and routines discussed so far apply equally to the other languages of GoK (Arabic, Greek, and Latin), implying support for different languages and scripts as a requirement. The questions above, on the other hand, assume English grammar and syntax more specifically.

The final heading for our first set of questions was *Temporal Spread*:

Temporal Spread

1. In what ways do keyword and collocational patterns correspond to the temporal spread of these texts (i.e. given the fact that some of these texts were published in 1850, others in 2012)?
2. Is a particular keyword more frequent in one time-period or another (e.g. within a specific year, decade, or longer historical period e.g. the Victorian era, post-1945, pre-1989, etc.)?
3. Are there time-periods in which the keyword does not feature at all?
4. Can certain adjectives/nouns/verbs be found to collocate more frequently with the keyword (in a particular word-position) in those corpus texts produced within one time-period versus those produced in an earlier or later time-period?
5. Are the changes in the relative frequency of a keyword over time similar or different to the patterns observed with regard to other keywords?
6. To what extent can these patterns be explained by other factors (especially those pertaining to the construction of the corpus itself e.g. the uneven distribution of tokens across the corpus as a whole)?

These questions, particularly 1–4, address similar issues to those catagorized as *Keyword* and *Collocational Patterns* questions. The difference is that these questions need to be answered with regard to the date of a corpus text's publication rather than its filename. Now using just the concordance browser, it seems essential to use subcorpus selections and a spreadsheet or notepad to analyse the patterns across time. Building this metadata frequency information into the tool and using it to interact with and explore the corpus becomes an obvious design goal to solve these issues. Question 5, by contrast, suggests some form of keyword extraction tool would be helpful, where keywords with similar temporal frequency profiles could be identified and grouped together. Question 6 seems to require expert interpretation of the results of other questions in the context of domain knowledge and some understanding of the limitations/design of the corpus.

The second researcher, Jan Buts (JB), decided to split the 20 questions into four categories *Keyword*, *Text*, *Author* and *Corpus*. The sections are ordered by importance, as are the questions within each. The researcher was keen to point out that the questions categorized within different sections will often intertwine and that the importance ranking is only an approximation.

It is interesting that *Keyword* is again given as a topic and is presented as most important by this researcher:

Keyword ("i.e. in case one wants to study a specific keyword in any number of texts")

1. How frequent is the keyword, and where is it ranked in a frequency list?
2. With which words is the keyword most frequently combined, in a span of four positions to the left and right?
3. What is the approximate strength of the collocational patterns observed?
4. Are there intuitive variations of the keyword (both formally and semantically) that occupy similar positions and display similar collocational patterns?
5. Which position does the keyword take in the clause, the sentence, and the text?

Questions 1 and 2 simply deal with keyword frequency and collocation frequency. Collocation strength is again mentioned explicitly in question 3. This is noteworthy as the concordance browser does not facilitate easy investigation of collocation strength. Question 4 calls for analysis of the formal and semantic variations of the keyword, and suggests that an understanding of what these variations might be comes from intuition. It is possible these questions could be also answered by the automatic keyword suggestion recommendations proposed by the first researcher. For question 5, discussion with the researcher revealed that reading each CL individually, using the 'extract' function of the GoK software to expand the context provided in relation to specific lines, and/or searching the original text externally from the corpus were the methods used at each level.

The first question listed under the heading *Text* concerns word frequency in a text and comparisons with other texts or sets of texts:

Text ("i.e. in case one wants to uncover the properties of a certain text")

1. What are the most frequent content words in the text, and how does this compare to other texts of a similar character?
2. What are the most frequent function words and connective elements in the texts, and with which of the content words above do they recurrently combine?
3. What are the most common proper names in the text?
4. Having identified all the above, do they vary in their dispersion across the document?
5. Having established all the above, where are (dis-)continuities situated in the text? (For instance, does the introduction display a different textual character than the body of the text).

The generation of word frequency lists was discussed as a way of answering this question. However, comparing raw frequencies within lists generated from subcorpora of different sizes can be problematic. It would not be a fair comparison, for example, to

contrast the raw frequency of a word in an article of 2500 words with its raw frequency in a book of 100,000 words; some way of automating the calculation of normalized frequencies (e.g. 5 instances per 1000 words) would consequently appear to be an important requirement. Question 2 again makes use of a frequency list to identify frequent words with certain linguistic properties. After the function words have been identified, collocation patterns are analysed. Question 3, on the other hand, relates to proper names and would require manual annotation or some automatic method which recognizes named entities. Finally, questions 4 and 5 are concerned with identifying the position of words within a text with the aim of identifying patterns of dispersion. Visualizing the spatial component of a text and highlighting terms of interest was suggested as a solution. Lexical dispersion plots such as those available in *WordSmith tools*, which display the relative locations of the occurrences of selected words in a text or corpus, are an example of a visualization which helps answer these questions.

Under the *Author* heading four of the five questions are concerned with frequency and collocation strength differences:

Author ("i.e. in case one wants to construct a profile for an author with multiple texts in the corpus")

1. Which words are the most frequent in each individual text written by the author in question, and how do this compare to the overall frequency of words in all the author's texts combined?
2. Which words does the author use significantly often in comparison to other authors similar in temporal, spatial, linguistic, or social context?
3. Who does the author frequently cite?
4. Which multi-word expressions occur significantly often?
5. Given all the above, are there temporal changes to be observed in the author's textual profile?

The questions compare an author to other authors, individual texts of the author, temporal profiles and various other metadata based subcorpus selection options. To answer these questions using the concordance browser would be time consuming, requiring many searches and subcorpus selections combined with note taking or spreadsheets. Question 3 would be especially difficult to answer using existing software: in order to make this process more efficient, the tool would need to be able identify citations and annotate them with metadata tags correctly, a feature that is not well supported in collocation and concordance analysis software. The alternative to automatic methods is manual annotation and linking.

Questions about the *Corpus* focus first on identifying frequent words or patterns, then investigating the collocations of those words and patterns:

Corpus ("i.e. in case one wants to interrogate a corpus varied in textual material")

1. What are the most frequent words, collocations, and other multi-word expressions in the corpus?
2. Can the frequency of the above be attributed to a limited number of texts, or is it characteristic of the corpus as a whole?
3. If the texts in the corpus display varied patterns regarding the above, how are relevant keywords, collocations, and multi-word expressions distributed across the corpus in terms of publication date, source language, author, etc.
4. What can one say about the specificity of the corpus in question in comparison to another specific corpus? (For instance, do certain keywords occur very often in all texts studied, while being very low-frequency in another varied corpus set).
5. Are the texts in the corpus explicitly connected through quotation or other types of reference?

Frequency lists and concordance analysis are employed to answer the questions, before moving to examine the distribution of those patterns of interest across the metadata attributes of the corpus. subcorpus searches, note taking and concordance analysis with a focus on frequency are the main techniques that would be useful.

Question 4 could be answered by comparing frequency lists. As we had previously discussed there are limitations when comparing word frequencies across two lists of different size, and our proposed solution of calculating instances per 1000 words does not take into account distribution of word frequencies. So, if for example a small number of words account for the majority of the occurrences the normalized frequencies for most words will be small and close together. This makes comparisons of word usage difficult and misleading. Solving the disconnect between the raw frequency and the distribution could help with frequency list comparison.

From these two sets of questions provided by the two researchers, we have identified some of the goals, tasks and techniques of our colleagues on the GoK project. We have found many instances in which information related to word frequency in a concordance would be useful. In particular, the examination of frequency information is not limited to the keyword of interest, but typically involves comparison and analysis in relation to other texts and subcorpora, as well as frequencies of collocates. This type of analysis was of high importance in the questions elicited from the researchers. During the discussions we established that these frequencies tend to be estimated visually from a concordance or manually counted. Supporting these actions through visualization was identified as a potential area which could benefit the GoK project and corpus analysis in general. Comparison of frequency lists is another area that was identified as a candidate for tool support. Frequency lists came up quite frequently in the discussion and the comparison of frequency lists was raised as a useful means of gaining insight into the particularities of a subcorpus selection. As noted, however, this comparison can be difficult in practice. Issues around the comparability of different sized lists cause problems, due to frequency and rank not being easily comparable for different sized lists. While this problem was not identified as the most important issue by the researchers, it does seem to be a problem which one cannot address without additional tool support. The third issue for which we considered creating visualization based solutions was metadata based frequency analysis. However, the methods currently used to answer questions related to metadata make use of external tools such as spreadsheets or notes and the analysis of several concordance queries together. The precise tasks which we would be attempting to facilitate were at this point not fully clarified. For this reason we gave priority to the other issues identified.

Software prototyping. At the start of the project, one of the authors (SL), attended meetings where the project team members discussed aspects of the scholarly work to be carried out, and the core research questions to be investigated. This provided the developers with basic intuitions as to how the tasks analysed in section "Analysis of published methodology" could be employed to support the kinds of investigation described in section "Observational research". We then employed low-fidelity prototyping and user interface sketching methods to communicate these intuitions to the research team and create initial designs for

what would eventually become the GoK tools. Note that by "fidelity", here, we mean how close to a finished product the prototype is. Low-fidelity prototypes or 'mock-ups' often take the form of paper prototypes, they are quick to design and easy to alter. An early version of the concordance mosaic visualization was sketched along with different representations of metadata, to be incorporated to the basic software platform (see section "The GoK Software"). Initial ideas for frequency comparison functions were also discussed.

The next step was to implement mock-ups and "bare-bones" working prototypes, illustrating how the tools might work at the level of the user interface. Some of these ideas eventually developed into the tools described next.

The GoK software. Concordancing software (or "concordancers") are common currency in corpus based studies. While arranging and indexing fragments of text for comparison and study dates from antiquity, the advent of computers enabled the systematic creation of *concordances* through the "keyword-in-context" (KWIC) indexing technique, as well as dramatically increasing the volumes of text that can be analysed. Concordancers have since become widely used tools across a broad variety of fields of research, from studies of lexicography and linguistics, to narratology, discourse analysis, and translation studies. The GoK software is based on a set of language processing software libraries (modnlp (Modnlp, 2020)) and a basic concordancer, which have been used in a number of projects (Luz, 2011). This infrastructure supports tasks such as indexing, data storage, metadata management, access and copyright compliance, as well as the basic user interface. As this has been described elsewhere (Luz, 2011), here we will focus on the visualization tools built for GoK specifically. This section gives a brief overview of the software and how it is used, before we delve into its design process in the following sections.

The modnlp client software provides a traditional concordancer interface, comprising frequency lists, a concordance display which allows sorting the CLs according to words to the right or left of the keyword, and functions to display text "metadata" (e.g. title, author, publication date, source language, subcorpus) and restrict the search and display of data based on these metadata. The basic modnlp concordancer is shown in Fig. 3. Through the concordancer the user can search for specific words, word sequences or "wildcards". The results are presented in a KWIC style, with the keyword aligned in the central column and surrounded by its immediate "context" (that is, the words that occur to the right and left of the keyword). Also included in the far left-hand column of the interface is the filename of the text in which each CL can be found. The concordance may be sorted by word position, but also by filename. Lines may be removed from the concordance by the user as necessary.

While the concordancer shows all occurrences of the string of characters searched for (refugee in the case of Fig. 4), the concordances rarely fit a single screen, forcing the user to sort and scroll in order to discover possible collocation patterns. The following tools, implemented as "plug-ins" to the modnlp system, were designed to aid this pattern discovery process, addressing the questions and tasks discussed in sections "Analysis of published methodology" and "Establishing initial requirements".

The concordance mosaic. The *concordance mosaic* tool (referred to as Mosaic for short) provides a concise summary of the KWIC display by presenting it in a tabular, space-filling format which fits a single screen. Mosaic is able to fit hundreds, often thousands of CLs onto a small display by taking advantage of the fact that a small number of types tends to dominate the distribution of tokens at each position in the concordance's context with respect to the keyword (a trend known as Zipf's law (Baek et al., 2011)) and therefore can be represented by a single object (type) on the screen, rather than a repetition of tokens. Thus Mosaic represents positions relative to the keyword as ordered columns of tiles. The design is based on temporal mosaics which were originally developed to display time-based data (Luz and Masoodian, 2004). Each tile represents a word at a position relative to the keyword. The height of each tile is proportional to the word statistic at that position. These tiles can be compared across all positions to evaluate quantitative differences between positional usage. However, the display option labelled *Collocation Strength (Local)* intentionally breaks this cross positional linkage and only allows comparison between tiles at the same position. Colors are used to differentiate between the frequency and collocation strength views of the concordance list. In its simplest form each tile represents the frequency of a word at a position relative to the keyword. In Fig. 4 the Mosaic generated for the keyword *refugee* as it is found in the GoK Internet corpus is presented. The tool is set to display a collocation statistic (cubic mutual information), which emphasizes higher-than-expected word frequencies. Due to the strong visual metaphor of KWIC it should be clear the word *anti* is the most salient (though not the most frequent) word immediately to the left of the keyword (K−1), and that *crisis* is the most salient word immediately to the right of the keyword (K+1); see Baker (this volume). Hovering over any tile will display a tool-tip with the word count and frequency at the relevant position. This relieves the need for manually counting or performing additional searches to retrieve position based word frequencies.

Alternative displays by relative frequency, relative frequency with very frequent words (stop-words such as *the*, *of*, *and*) removed, and scaled according to global statistics (rather than scaled to fill the space of each column, as in Fig. 4) are also available. The user additionally has the option of applying a range of different collocation statistics, such as mutual information, log-log and *z*-score (Pecina, 2010).

Concordance tree. The economy of representation in Mosaic comes at the cost of sentence structure. Mosaic is based on an abstraction of the CLs as a graph, where types are connected with other types in linear succession, with each path in the graph corresponding to a sentence in the collocation (Luz and Sheehan, 2014). However, the tabular layout of Mosaic as juxtaposed tiles does not encode these paths visually, and consequently it is impossible for the user to know at first glance whether two words that occur next to each other on Mosaic also co-occur in the same sentence. For example, the words *de-politicize* and *Syrian* occur next to each other on the mosaic of Fig. 4 (K−2 and K−1, respectively) but there are no sentences in this concordance that feature the expression *de-politicize Syrian*. To discover which sentences (paths) exist for a given word the user must click on that word and observe which collocates appear highlighted (as white tiles) on Mosaic.

An alternative rendering of the underlying graph structure of the concordances is the *concordance tree*. This tool is a variant of the Word Tree design, introduced by Wattenberg and Viégas (2008). It shows the left or right context of a concordance as a prefix tree, where each branch (path along the graph, from root to leaves) corresponds to a sentence in the concordance. The font size of each word at a particular position on the branch is scaled according to the frequency of occurrence of that word at that position.

A fragment of a concordance tree corresponding to the right context of the *refugee* concordance is shown in Fig. 5. While the concordance tree preserves sentence structure, it cannot generally

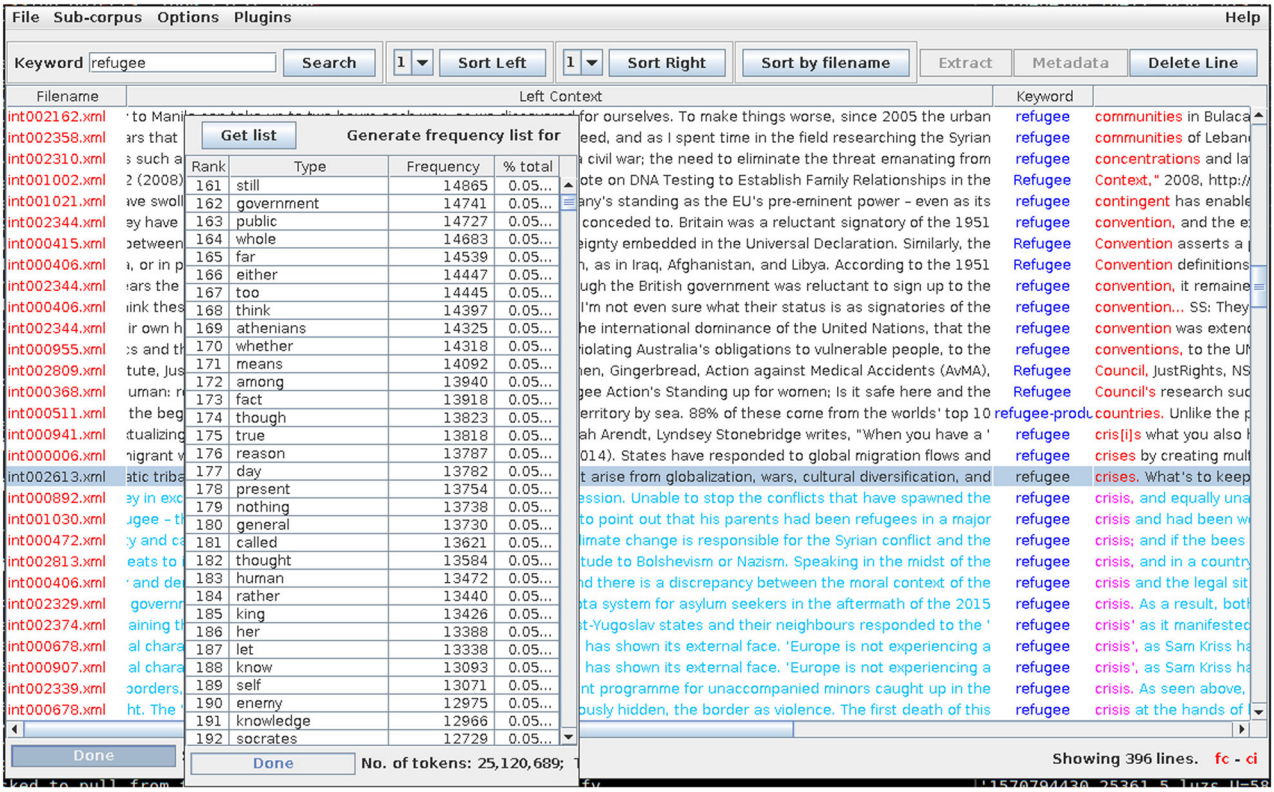

Fig. 3 Modnlp concordancer and frequency list. Highlighted items have been sorted and selected by the user using Mosaic.

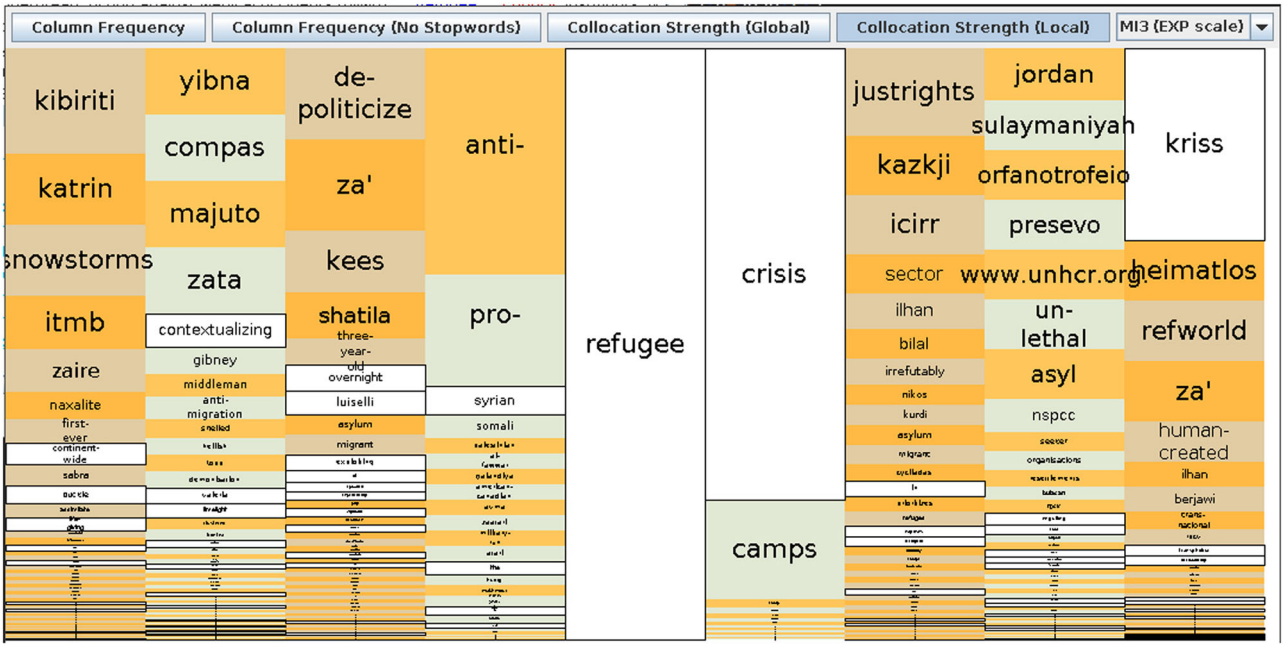

Fig. 4 Mosaic visualization for the concordance of the word "refugee" in the Genealogies of Knowledge corpus. The word "crisis" has been highlighted, causing all words that occur in sentences containing the word crisis at position K+1 to be highlighted as well, and the concordancer to display those occurrences in full context (Fig. 3).

display a full concordance as compactly as Mosaic. In fact, in the worst case it can take up as much space as a full textual concordance. However, as the purpose of visualization is to emphasize frequently occurring patterns, one will typically "prune away" low frequency branches (using the depth-1 width slider; the top right corner of Fig. 5).

Metafacet. The concordancer allows the user to retrieve metadata about each file and section on a line by line basis. This is however a time consuming and challenging process for the corpus analyst if the metadata of a large number of lines need to be investigated. To better support the exploration of metadata in conjunction with concordancing we have implemented the Metafacet tool,

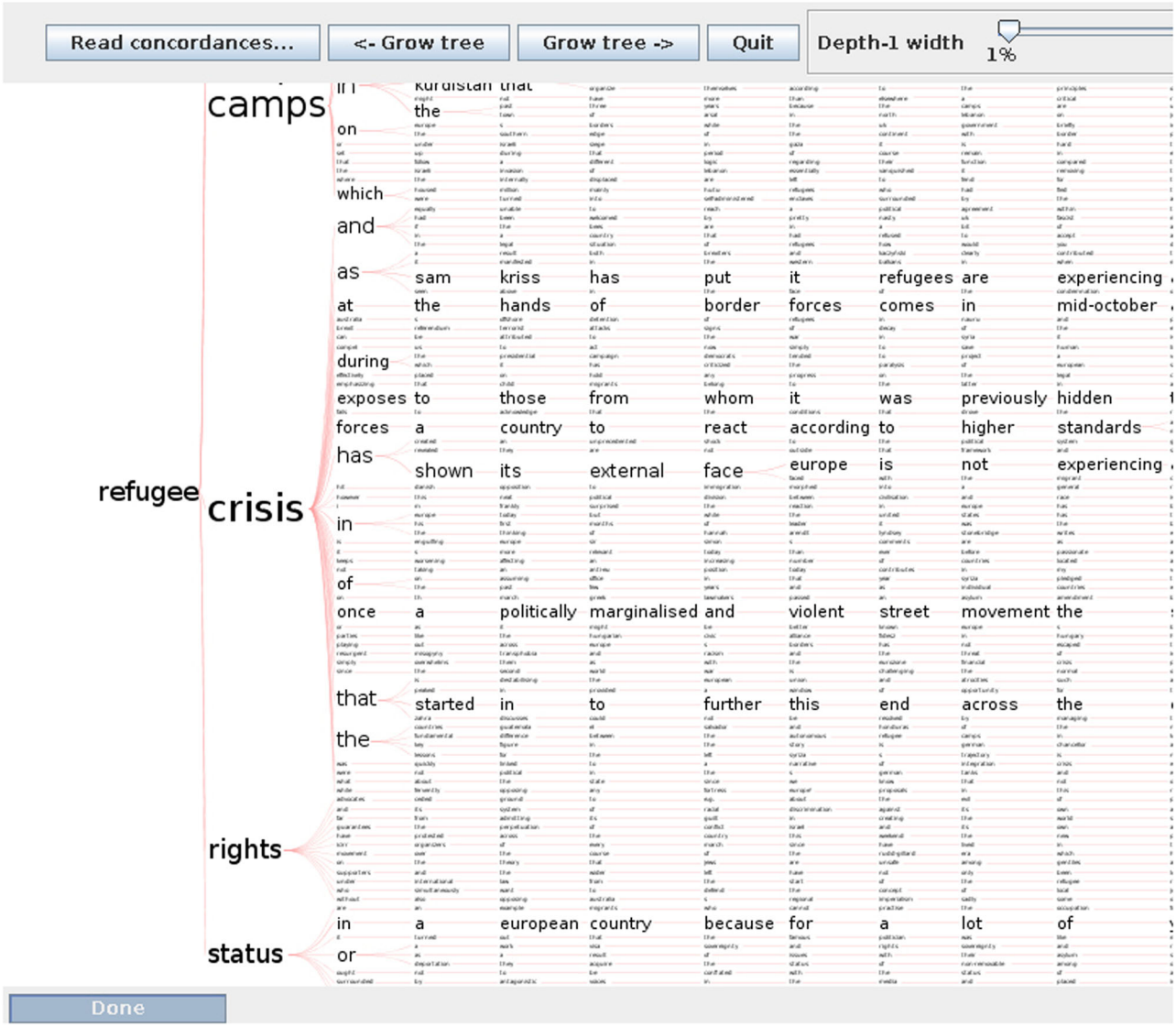

Fig. 5 Fragment of a concordance tree showing the right context of the word refugee in the concordance shown in Fig 4. Concordance tree (fragment) showing the right context of the word *refugee* in the concordance shown in Fig. 4.

which provides a faceted summary of the available metadata. It also allows for the interactive filtering of the concordance list and the Mosaic using all available metadata facets (Sheehan and Luz, 2019).

The interface uses a horizontal bar chart to display CL frequencies per metadata attribute. Colour is used to to help with visual comparison of bar length but the gradient otherwise encodes no special meaning. An attribute is a possible value that a metadata facet can take. For example, "The Nation" being the name of an online magazine whose contents the GoK team has included in the GoK Internet corpus, is an attribute of the facet "Internet outlet".

Figure 6 shows a Metafacet chart for a concordance of *refugee* generated from the Internet corpus, with "OpenDemocracy" selected as the internet outlet for analysis. A drop-down list is used to choose which facet is displayed and the bars can be sorted by frequency or alphabetical order. Moreover, the visualization window can itself be filtered using a sliding scale (positioned on the far right) in order to allow the user to view a smaller portion of the attributes. This conforms to the common visualization design pattern of overview plus detail on demand, whereby users are initially presented with a general summary of the dataset being analysed but retain the option of conducting finer grain analyses according to their interests (Shneiderman, 1996).

Metafacet, when used on its own, provides an interface with which to explore keyword distribution across different metadata attributes. By combining it with the concordance list and Mosaic, the user can navigate the corpus in a new way, viewing the concordance as attributed sets of collocations that can be interactively filtered, sorted and examined.

Frequency comparison tool. The modnlp frequency list shown in Fig. 3 provides detailed statistics on term frequency overall or by subcorpora. However, it does not allow easy comparison of frequencies in different subcorpora. The *frequency comparison* tool allows frequency lists to be compared visually in a statistically valid manner. The functionality of the tool has been described elsewhere (Sheehan et al., 2018), and it has since been modified to operate as a plugin for modnlp and is briefly presented here.

The modnlp concordancer has a subcorpus selection interface which can be used to save named subcorpora for later reuse. These named subcorpora then become available for comparison

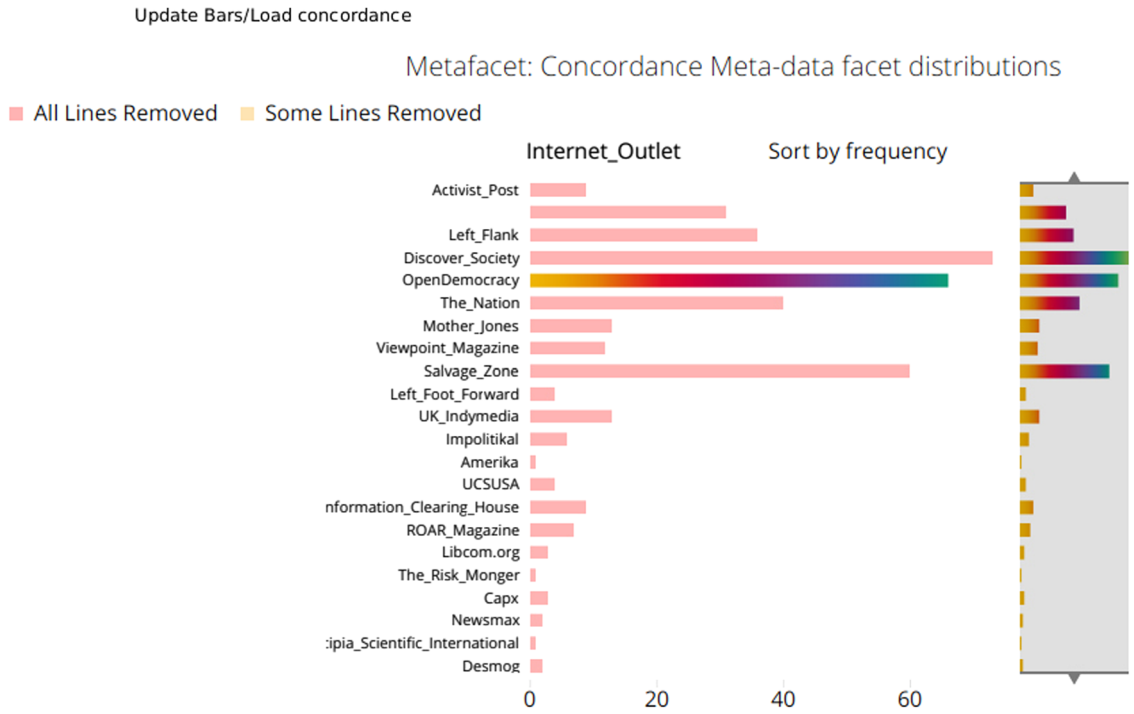

Fig. 6 Metafacet tool showing the attributes (in this case, publication titles) of the facet "Internet outlets".

through the frequency comparison tool. Figure 7 shows a comparison of two pairs of outlets from the GoK Internet Corpus. Frequency information for the outlets *ROAR Magazine* and *Salvage Zone* is displayed on the left, these outlets explicitly adopt a more radical left agenda than others like *The Nation* and *Open Democracy* which are displayed on the right. In this diagram, both axes are log scaled which should yield a linear frequency diagram if the word frequencies follow a Zipfian distribution. Scaling both ranked lists to the same height and comparing a word's position in the distributions enables the user to compare subcorpora of vastly different sizes.

In Fig. 7, the majority of the words have been filtered out to reveal the words with the greatest frequency changes between the two corpora. The words are placed at heights that correspond to the rank order in which they occur in the frequency distributions in their respective corpora. The lines in the middle connect words to the positions in which they appear in the other corpus' ranked distribution. Thus the diagram shows the word "capitalist" is ranked just below 100 (with a frequency of 0.096%) in the radical left outlets, and at nearly 4000th (with a frequency of 0.003%) in the less radical pair. In this case the total number of unique tokens is roughly the same at ~10,000 tokens. The comparison would still be possible if the scales were vastly different, if one of the subcorpora was much larger than the other. The nature of the subcorpora should be evident from the differences in the frequency distributions of these words.

The corpus exploration facilitated by the frequency comparison tool supports hypothesis discovery activities, such as questions 3 and 6 in HJ's keyword related questions list (in "Keywords"), as well as questions 2 and 5, relating to temporal spread (in "Temporal Spread"), as time intervals can be used to define subcorpora for comparison. Similarly, this tool allows JB to investigate question 2 of his author question list (in "Author") and questions 2 and 4 of his corpus list (in "Author"). See also the papers by Baker (2020) and Buts (2020) for examples of uses of the frequency comparison tool in scholarly work.

The images of the above described visualizations were checked using the Coblis color blindness tool (Coblis, 2020). For red, green and blue, the displays are readable without issue. Differentiating between the collocation strength and frequency views for red or green, can be slightly challenging for colour blind individuals as the colour profiles become similar. However, there is enough difference in saturation to tell them apart. In addition, the button shading for the selected interface helps the user identify the option that is selected. This works even for monochromatic images. As colour is used only to help differentiate between items and not as a visual encoding of a data attribute the exact color that displays is of minor importance.

Observational research. To get a better understanding of the techniques and methods used by researchers in the GoK project, we requested time to observe research where the concordance browser was in use. Early versions of Mosaic and Compare Frequencies tools had additionally been integrated into the software by this point.

Interview and case study on the concept of democracy. A researcher from the GoK project (Jan Buts) offered to discuss his methodology and give an outline of the typical analysis process. The initial discussion was not based on a specific case study but describes the general method employed by Jan. After the initial discussion Jan agreed to let us observe a partial re-enactment of an analysis which had already been performed: this related to historical changes in the use and meanings of the concept of *democracy*. Jans's methodology was described as the search for the largest unit of meaning related to a keyword. Meaning in this case should be constructed from the evidence present in the corpus. The corpus is central to the analysis and the technique is in the style of Sinclair (2003), meaning the investigation of collocation, colligation, semantic preference and semantic prosody is performed by an analyst who must be make a conscious effort to

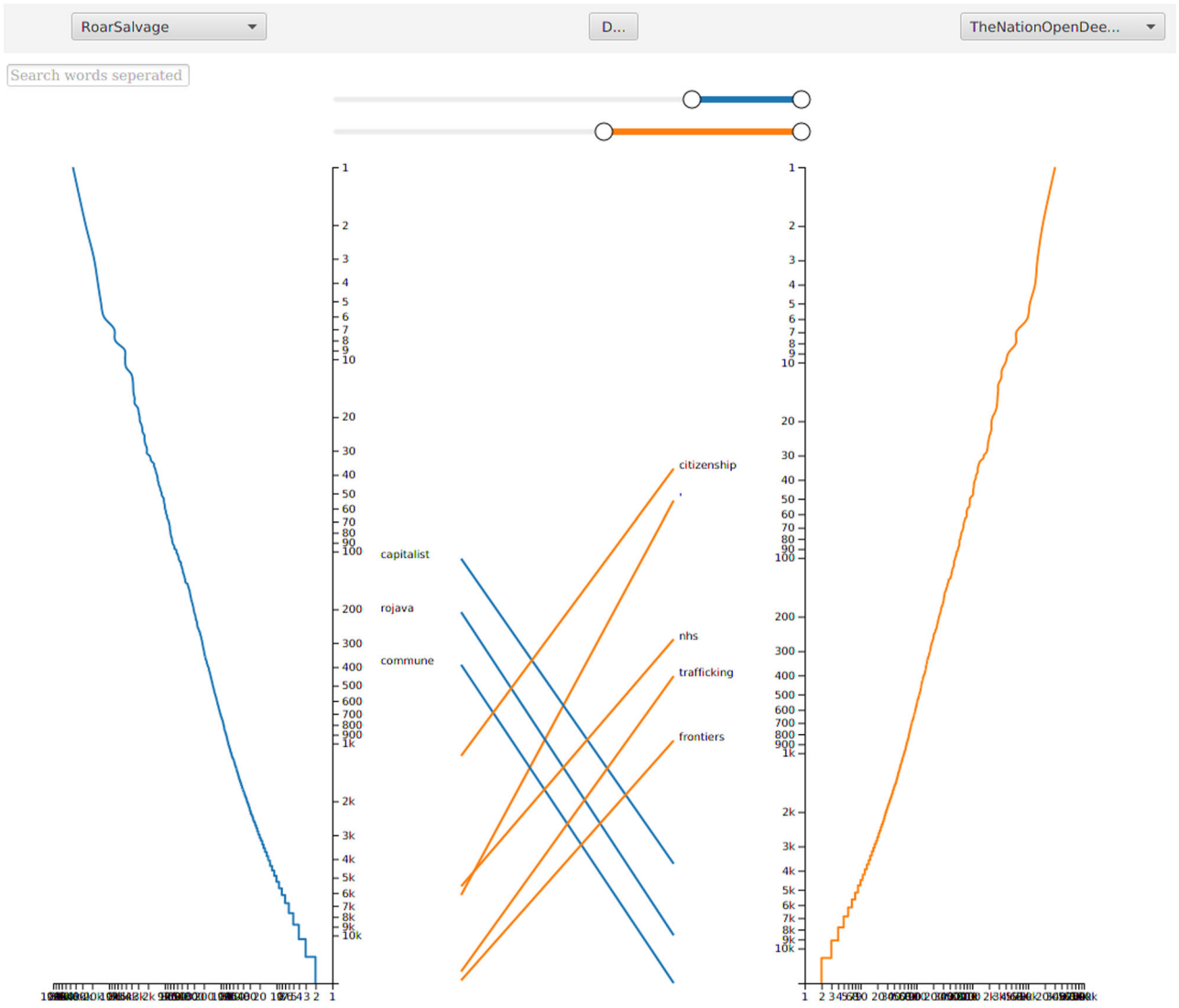

Fig. 7 Compare Frequencies visualization showing words with the largest change in usage between the outlet pair *ROAR Magazine* and *Salvage Zone* (left) and the outlet pair *The Nation* and *Open Democracy* (right).

ignore personal bias. A list of the ordered steps used to perform this type of analysis was provided:

Construction of meaning:

- Sample
- Describe Patterns
- Sample
- Compare Patterns
- Hypothesize

In the steps provided the term "sample" refers to a subcorpus selection and keyword search in a concordance browser: if the concordance is large, the samples may be thought of as subsets of the full concordance list. "Describe Patterns" refers to the process of analyzing the positional frequencies around the keyword. Jan explained that the analysis begins by looking at the patterns of words occurring next to the keyword and expands to additional positions until the discovered patterns describe the meaning of most CLs. The remaining lines would then be analyzed after the core units of meaning had been established. "Compare Patterns" is the process of examining the differences and similarities between the described patterns of the samples. The following clarifying question was posed to Jan:

- Can you give practical details of your typical methodological approach?

> "Investigate the keyword and its neighbouring collocates (Left and right +1). Investigate deviations from frequent patterns then expand the analysis horizon and repeat until the largest unit of meaning is found. Largest unit of meaning should be read as 'overarching', in the sense that the point is not to necessarily go beyond the concordance line, but to construct an abstract unit that can account for as many concordance lines as possible. If interesting patterns which lead to a hypothesis are discovered pursue these. Typical corpus linguistics method applied to unique corpus".

During the discussion some difficulties were reported in relation to working with the GoK corpus. The unique nature of the corpus makes the generalization and the representativeness of

a hypothesis more difficult to explain. Viewing metadata for the CLs is useful but, as it is line specific, it is impractical for analyzing large numbers of CLs. The concordance browser's use for identifying patterns in large numbers of CLs requires sampling multiple times. A broader picture of the concordance which can examine broader and more restricted contexts would be desirable. New tools should complement the concordance, extending its functionality rather than seeking to replace it. The statement that the unique nature of the corpus caused problems required clarification. The following question was posed in relation to this issue:

- How and why does the corpus influence the methodology?

> "If speaking about 'typical/traditional' corpus linguistics (which is always a bit of a stretch), one finds actual/practical lexicography, and analysis of register/text-type, etc., drawing on corpora that are constructed to serve as a sample of the full language or sub-language under investigation. Think of the British National Corpus, for example.
> Our corpus hosts a variety of texts, but it would be difficult to make the case that it is representative of anything outside the corpus itself. Our Internet corpus, for instance, is not a sample that can tell something meaningful about the Internet 'as a whole'. Therefore, rather than making exhaustive analyses of a certain word across the corpus, or tracing a grammatical pattern across its contents, the corpus urges one to study a specific subset of texts, and to complement the findings with sources outside of its confines to make hypotheses about conceptual developments. Consequently, the method will be less repetitive than one would traditionally see, and more meandering, to a certain extent".

When asked at the end of the interview if it would be possible to observe the method in the future, Jan volunteered to give a demonstration of some analysis. The demonstration which we observed was a partial re-enactment of an analysis which had already been performed. The concept of democracy was investigated in a subcorpus of political texts published in English from 1970 onward. Jan commented that "this is in line with the most fundamental goals of the GoK project". The steps taken which were observed and recorded were:

- Begin by searching the keyword *democracy* without any bias for what will be returned;
- Open Mosaic and see if anything stands out (nothing does);
- Look at Column Frequency (No Stopwords) view. Social democracy appears to be a very strong collocation. Click social and look at the CLs now highlighted in the concordance browser;
- Reading the lines reveals that*Social democracy* only occurs in file mod000008 and refers to one book title and its contents. This is only informative about this specific file and the file is then removed from the subcorpus under investigation for the sake of gaining a more balanced overview. (This appeared an unusual step and was recorded as needing further clarification);
- Re-run the search this time ~500 lines were removed from the concordance. *Common* and *Athenian* were recorded as important collocates;
- Mosaic is consulted again, both the Column Frequency and the Column Frequency (No Stopwords) views. These do not seem to show any unexpectedly frequent results;
- Navigate to Collocation Strength (Global) view and investigate the words one position to the left of the keyword;
- Do any of these extreme combinations also have interesting frequency profiles (not single occurrences in the concordance)? Investigate by looking for words which stand out in the Mosaic Column Frequency and Collocation Strength views;
- Did not find any particularly interesting frequent and strong collocations at position left +1;
- Use a regular expression to search for "-acy". Interested in keyword frequency and collocations;
- Note *democracy* is 76% of "-acy" occurrences.
- Looking at other frequent keywords (*aristocracy*, *bureaucracy*): they are mostly negatively framed in the CLs;
- Switch to concordance strength view and observe that the highest ranked keyword is *mediaocracy*;
- Search *mediaocracy* 10 lines returned;
- Use Column Frequency (No Stopwords) view of Mosaic and the concordance browser to establish the semantic prosody of the term, that is, whether it is used negatively or positively;
- Hypothesis: Democracy is the dominant "-acy" and is viewed in a positive way. All other "-acy" are presented as negative. They are presented as threats to democracy.

This description reveals heavy use of Mosaic for analysis. The case study presented seemed to be a partial treatment of the problem and may have skipped some steps which were needed to reach the hypothesis. Jan was asked the following clarifying questions:

- You moved swiftly from removing the file mod8 to investigating collocation strength. After removing the file mod8 you did not re-investigate the collocation frequency of *democracy* and instead moved on to collocation strength. Why?

> "Just for demonstration purposes. In essence, not only were pieces skipped over, the illustration was also fairly preliminary in the following sense: Removing mod8 because it creates some distortion is of course bad practice [if this were the actual research]. The point in doing so is to quickly weed out material unfit for my purposes, until I reach a suitable point of investigation (in this case: democracy turning from one of the competing systems of rule into the only one available, however constantly beleaguered by threats from within). Once this point of investigation is established, the analysis can start out again and I make sure to construct a suitable subcorpus on clearly defined terms that doesn't require me to be rash at the outset of an analysis. The mosaic view can then be approached again as an entry into the data, and all the collocation patterns examined more closely".

- How do you decide what subcorpus to initially investigate? In this analysis books form 1970 to present date.

> Currently the first thing I do (especially when the concordance return is small) is look at overlaps in meta-data property between concordance lines, to get a sense of the whereabouts of the data.

- Would a visualization which shows frequency of a keyword across meta-data facets be useful?

> "Yes. One could, for example, look at differences in dispersion in the use of the word 'terror' pre- and post-9/11, look at whether a certain author evades a word (say, anarchy) that is used by all other authors writing on the

same subject (say, democracy), one could look at whether a magazine has a regional, national or international outlook by comparing the proper names used with those in other magazine, etc".

• In your analysis I struggle to see why you began analyzing the -acy concordance. It does not appear to follow from the previous steps of analysis. Is this an established next step in corpus linguistics? Is it based on experience and domain knowledge or some part of the analysis not presented?

"This has to do with the reduction of bias through the reliance on form. I could, for example, go look at democracy vs. totalitarianism (in my attempt to study contemporary forms of government), but I have no proof that these concepts in fact are alternatives to each other. This would be solely based on intuitions, and as a lot can be argued about language data, I would basically come to prefabricated conclusions if I wanted to (democracy is opposed to totalitarianism in the following senses). Starting out from taking the suffix-cracy and seeing what other terms it attaches to offers a more neutral entry into the data inspired by the actual linguistic form rather than pre-conceived oppositions".

• You did not appear to investigate the collocations of democracy and other (-acys) to determine the usage or context in which they occur, except for meritocracy, I am assuming that this was done and just shown?

Indeed, in the final analysis every term discussed merits close attention to the immediate co-text.

• You use Mosaic extensively in the method? Is that typical of your work

I use the Mosaic every time I access the corpus. Especially at the beginning of an analysis, to get an idea where to start and to make sure I won't, in a later stage, overlook any significant patterns.

• You appear to use the collocation strength view for analysis, what is your opinion on it?

Useful for analysis as it gives extreme combinations. (where the combination rarity is interesting). As it stands the analysis done using the Collocation strength view is difficult to explain. Justifications for the patterns found using this view are usually easier to re-frame as part of the qualitative analysis which involves reading the concordance lines.

• If other statistical measures were available in Mosaic would that be useful?

Yes, we would benefit from a measure of confidence rather than strength, or from a commonly known measure that can simply be mentioned as such in publications.

Case study on the concept of "the people". In this case study we observed an analysis of keywords related to the concept of "the people" featuring in a subcorpus of eight different English translations of Thucydides' *History of the Peloponnesian War*. The researcher (HJ) told us this was early stage exploratory research that, he hoped, would ultimately lead to a publication. The details of the think aloud observation session and interview can be found in Sheehan and Luz (2019). While think aloud user studies are one of the most common user study techniques they need to be carefully controlled. In performing the study, we carefully explained what was required by Henry. That he should not try to think about what he was going to say and should instead try to narrate what he was doing in real-time. We were careful to not bias the study via prompts, we used repeated phrases from a script to help avoid leading questions and interviewer bias.

To summarize the session, the observed method consisted of an analysis of multiple keywords, related to the concept of "the people" using frequency and collocations. The method made use of a concordance browser to select the subcorpora, retrieve the keyword frequencies and to help list the most frequent collocates of the keywords. A spreadsheet was used to keep track of the keyword frequencies per translation and to list the collocates of interest. The process was time consuming and the researcher would benefit from automated methods of extracting the required numerical information, however he was unaware of any tools which could help with this type of analysis when the corpus must be made available through concordance and not as full texts. When investigating the frequent collocates the mosaic was used to save time counting or estimating frequency from the concordance list.

At the end of the observation session Henry explained how the analysis would progress beyond what was observed. For each file and keyword combination the collocation patterns would be identified using the observed technique. Clearly this will be very repetitive and time consuming. Following the collection of these results, the next stage would be to look at the frequency patterns using the table of results. Making bar charts in a spreadsheet application or external tool will be helpful for examining the trends. Temporal patterns are expected but any identified patterns will be investigated using qualitative analysis, which involves reading the CLs which relate to the identified pattern. Understanding the meaning of the concept of *the people* at different times is the goal of these analysis steps. Any differences identified, temporal or otherwise, must take into account individual translator's style, the political context and many other factors. The researcher's knowledge of the domain, the corpus and the individual texts is essential to the analysis. For example, the observation made during discussion with Henry that "There are no translations 1919–1998, during period of huge cultural change in Britain. Possible reasons for this include Suffrage, war or technological revolution" would be difficult to derive from the concordancer as it relates to an absence of texts in the Modern English corpus. Henry explained that information about the authors and texts will influence the analysis. Some examples of the type of information which can be relevant are "the political leanings of the translators which is established relevant knowledge" and knowing that "certain texts are partial translations, abridged versions etc.".

In regard to methodology, Henry described a process of exploration using the corpus tools and pattern search, drawing on the actions described in section "Analysis of published methodology". Common lexical knowledge and knowledge of the literature on specific concepts (e.g. the people) helped identify keywords for exploration. Similar patterns of exploration and use of the GoK tools are shared among other GoK scholars to investigate the role of translation in the evolution of political and scientific discourse. Moreover, Henry believes aspects of this methodology can be used by other projects that use corpora. Lack of familiarity with the software, and lack of documentation were mentioned as barriers to the further development and wider adoption of the methodology. However, tools such as Mosaic were considered intuitive and helpful in "any investigation of

collocations, as it tells you in a very quick and transparent way which are the most common collocates in each word position for a given keyword". A somewhat different investigation pattern was observed in the next case study.

Case study on the concept of statesmanship. This observation session took place after a piece of analytical work had already been completed, but Henry offered to explain and re-enact a portion of the investigation for the purposes of this study. The analysis we observed contributed to a publication (Jones, 2019). The following is a summary of the observed analysis and explanation.

In the GoK Modern English corpus the term *statesman* was found to exist "almost exclusively (90%) in translations from Classical Greek". This pattern was not observed for other semantically connected keywords such as *governor*, *leader*, *ruler* and *citizen* which are more evenly distributed across translations from all source languages rather than only Classical Greek. The analysis which led to at this conclusion involved a simple keyword frequency comparison across the source language metadata recorded for each text in the corpus. This involved selecting the subcorpus of texts translated from each source language represented in the Modern English corpus individually and recording the number of CLs for the same keyword search.

A spreadsheet with an entry for each of the 261 files in this subcorpus was created and metadata (the author, the title, the translator and the date) was entered for each file. This was done manually and was time consuming. Henry explained that in this spreadsheet "the information could easily be (re)sorted according to each of these meta-data facets and patterns more easily identified". The number of CLs for each file was found by selecting a subcorpus consisting of a single file and searching for *Statesman*. Performing this action for each of the 261 files was very time consuming. After the spreadsheet has been filled in for each file the information could be analyzed to look for frequency patterns across the metadata attributes (such as author or date). Sorting and visualization (bar charts) were the main techniques used to get an overview of the identified patterns.

Henry described a process of exploration that aimed at determining the use of the term *statesman* and its frequency in comparison to semantically related terms. The main difficulty in this process was the time-consuming nature of the collection of frequency data. As analysis of collocations played a minor role in this context, the visualization tools were not used as frequently as in other cases. The most significant outcome of the two case studies was the emergence of the obvious need for a method to support the analysis of concordance lists through the lens of metadata. This topic had previously been raised in the requirements elicitation process. However, without this observational work, the requirements were too general and it was difficult to understand the low level actions we would need to support. From these two sessions it became clear that filtering the concordance list via metadata facets would be worthwhile. In addition having an instant breakdown of the number of CLs per metadata attribute would be useful. This observation session led to further discussion among team members and the eventual development of a metadata analysis tool which eventually became Metafacet. Another problem identified was that in the version of Mosaic available to the researchers at that time only a single collocation statistic was available and it was based on mutual information scores. Due to a lack of proper documentation, Henry did not know exactly what the scaling scheme was for the collocation strength view of Mosaic and so could not accurately interpret or use it for publication. This led to the writing of a detailed user manual and the addition to the Mosaic tool of optional scaling schemes based on well known collocation metrics. More collocation measures are still being added to the tool.

Discussion

One of the main themes that emerged from the iterative design process presented in this paper is the role of visualization tools as a means of sensitizing the investigator to overall patterns which may not be easily captured by extensive sequential reading. Detecting such patterns is not however, as one might expect, merely a matter of compiling and interpreting text statistics. It blends data representation and statistical elements with aspects of the laborious reading and interpretation process that characterizes more traditional scholarly enquiry. Although certain statistical aspects of analysis have been raised by researchers, especially in connection with the discovery of collocation patterns through Mosaic, they are not seen as the main product, or even necessarily part of the main product of analysis. They rather serve the purpose of guiding the exploration when a corpus of text is represented as graphical summary. The analytical work remains essentially qualitative and interpretative.

The methods detailed by Sinclair are at the core of many research areas. We do not however expect them to capture the full range of tasks in these areas, or in a research project such as GoK. Researchers working with text build upon and sometimes diverge from the core methods as needed. It is not surprising then that we find differences between the methods identified in our hierarchical task analysis of Sinclair's work and our domain characterization efforts with the GoK project. Sinclair's tasks do not feature any comparative methods similar to those we observed. In particular subcorpora are not discussed at all. In our experience working with translation scholars this type of comparison is very common. Since these methods are not explicitly described in foundational texts it is less likely that they would be well supported by tools.

Avoidance of bias is another critical issue. The sparsity of language virtually guarantees that any sample (corpus), however large, will include an element of bias. While this is unavoidable, it is important that the text processing and visualization tools allow the researcher to identify possible sources of bias in text. This is important from two perspectives. First, identification of bias in the corpus, tracing it back to the texts and contexts in which it occurs is often an important element of analysis, and may sometimes be the object of analysis itself. Second, the highlighting of biases through overviews of textual patterns may help the researcher become aware of their own pre-conceptions that adversely affect the interpretation of data.

A use of the visualization tools which has not been explicitly discussed in the previous section but which featured in studies produced by project members (see Baker's and But's contributions to this special issue, for instance) is the use of graphics for communicating conclusions and viewpoints originating from corpus analysis. In communicating one's views to others through visualization, simplicity is imperative (Bertin, 1981). In this sense, the intuitive simplicity of Mosaic, for example, has enabled investigators to present much clearer illustrations of usage patterns than would be possible to do by means of tables and concordance displays. This is an aspect of the use of the GoK visualization tools that we would like to explore further.

Finally, it became apparent that tools need to be documented with the end users in mind, the language should be the language of the domain and not of the designer. Detailed documentation is needed and this documentation must be written collaboratively by developers with the help of users to ensure both its accuracy and its relevance to key research problems.

Conclusion

"Translator invisibility" is an issue with which many language scholars are familiar (Venuti, 1995). Venuti offers a critique of the view of "tranparency" as an ideal in translation, regarding it as an illusory notion that tends to render the work of the translator invisible, in detriment of culture. Although this is not the place for an in-depth discussion of this argument, we note that in many ways software developers working with humanities researchers often find themselves in a similar position of inconspicuousness. Software tools and methods, much like statistical tests, are taken to provide an objective (transparent) means through which analytical work that is often of a markedly qualitative nature is done, and the developer or computational researcher regarded simply as the (invisible) "service provider". The methods and analysis presented in this paper suggest that this is neither an accurate characterization of the developer's contribution nor a desirable situation in interpretation projects. Effective interdisciplinary engagement is necessary if progress is to be made in the digital humanities.

Data availability

All data generated or analysed during this study are included in this published article.

Received: 30 October 2019; Accepted: 3 March 2020;
Published online: 23 March 2020

References

Annett J (2003) Hierarchical task analysis. In: Erik H (ed.) Handbook of cognitive task design. 2. Lawrence Erlbaum Associates, New Jersey, pp 17–35
Anthony L (2004) Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In: Anthony L, Fujita S, Harada Y (eds) Proceedings of IWLeL, pp 7–13
Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. New J Phys 13 (4):043004
Baker M (1993) Corpus linguistics and translation studies: implications and applications. In: Baker M, Francis G, Tognini-Bonelli E (eds) Text and technology: in honour of John Sinclair. John Benjamins Publishing Company, pp 233–250
Baker M (1993b) Corpus linguistics and translation studies: implications and applications, chapter 11. John Benjamins Publishing Company, Netherlands
Baker M (1995) Corpora in translation studies: an overview and some suggestions for future research. Target 7:223–243
Baker M (2020) Rehumanizing the migrant: the translated past as a resource for refashioning the contemporary discourse of the (radical) left. Palgrave Commun 6(1):1–16
Baker P (2006) Using corpora in discourse analysis. Bloomsbury discourse. Bloomsbury Academic
Bernardini S, Kenny D (2020) Corpora. In: Baker M, Saldanha G (eds) The Routledge handbook of translation studies. Routledge, pp 110–115
Bertin J (1981) Graphics and graphic information-processing. de Gruyter
Bertin J (1983) Semiology of graphics. University of Wisconsin Press
Biber D, Douglas B, Biber P, Conrad S, Reppen R, University C (1998) Corpus linguistics: investigating language structure and use, cambridge approaches to linguistics. Cambridge University Press
Bonelli ET (2010) Theoretical overview of the evolution of corpus linguistics. In: Anne O'K, Michael McC (eds) The Routledge handbook of corpus linguistics. Routledge, p 14
Buts J (2020) Community and authority in ROAR Magazine. Palgrave Commun 6 (1):1–12
Cleveland W, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. Science 229(4716):828–833
Coblis (2020) Coblis color blindness tool. https://www.color-blindness.com/coblis-color-blindness-simulator/. Accessed Feb 2020
Culy C, Lyding V (2010) Double tree: an advanced kwic visualization for expert users. In: Banissi E, Bertschi S, Burkhard R, Counsell J, Dastbaz M, Eppler M, Forsell C, Grinstein G, Johansson J, Jern M, Khosrowshahi F, Marchese FT, Maple C, Laing R, Cvek U, Trutschl M, Sarfraz M, Stuart L, Ursyn A, Wyeld TG (eds) Proceedings of the 14th International conference on information visualisation (IV). pp 98–103
Culy C, Lyding V (2011) Corpus clouds—facilitating text analysis by means of visualizations. In: Mariani J (ed.) Human language technology. Challenges for computer science and linguistics, vol. 6562 of Lecture notes in computer science. Springer, Berlin, Heidelberg, pp 351–360
Culy C, Lyding V, Dittmann H (2011) Structured parallel coordinates: a visualization for analyzing structured language data. In: Pastor MC, Trelis AB (eds) Proceedings of the 3rd international conference on corpus linguistics, CILC-11. pp 485–493
Frank AU, Ivanovic C, Mambrini F, Passarotti M, Sporleder C (eds) (2018) Proceedings of the second workshop on Corpus-based Research in the Humanities CRH-2, vol. 1 of Gerastree proceedings
Green M (1998) Toward a perceptual science of multidimensional data visualization: Bertin and beyond. ERGO/GERO Hum Factors Sci 8:1–30
Hareide L, Hofland K (2012) Compiling a norwegian-spanish parallel corpus. In: Andersen G (ed.) Quantitative methods in corpus-based translation studies: a practical guide to descriptive translation research, Studies in corpus linguistics. John Benjamins Publishing Company, pp 75–114
Jänicke S, Scheuermann G (2017) On the visualization of hierarchical relations and tree structures with tagspheres. In: Braz J, Magnenat-Thalmann N, Richard P, Linsen L, Telea A, Battiato S, Imai F (eds) Computer vision, imaging and computer graphics theory and applications. Springer International Publishing, pp 199–219
Jones H (2019) Searching for statesmanship: a corpus-based analysis of a translated political discourse. Polis 36(2):216–241
Jänicke S, Blumenstein J, Rücker M, Zeckzer D, Scheuermann G (2018) Tagpies: comparative visualization of textual data. In: Telea A, Kerren A, Braz J (eds) Proceedings of the 13th international joint conference on computer vision, imaging and computer graphics theory and applications, vol. 2: IVAPP. INSTICC, SciTePress, pp 40–51
Jänicke S, Franzini G, Cheema MF, Scheuermann G (2015) On close and distant reading in digital humanities: a survey and future challenges. In: Borgo R, Ganovelli F, Viola I (eds) Proceedings of the Eurographics conference on Visualization (EuroVis)—STARs. The Eurographics Association
Kilgarriff A, Baisa V, Bušta J, Jakubíček M, Kovář V, Michelfeit J, Rychlý P, Suchomel V (2014) The sketch engine: ten years on Lexicography 1(1):7–36
Kilgarriff A, Rychly P, Smrz P, Tugwell D (2004) Itri-04-08 the sketch engine. Inf Technol 105:116
Kocincová L, Jakubíček M, Kov V, Baisa V (2015) Interactive visualizations of corpus data in sketch engine. In: Grigonyte G, Clematide S, Utka A, Volk M (eds) Proceedings of the workshop on innovative corpus query and visualization tools at NODALIDA 2015. pp 17–22
Léon J (2007) Meaning by collocation. In: History of linguistics 2005. John Benjamins, pp 404–415
Luhn HP (1960) Key word-in-context index for technical literature (kwic index). Am Doc 11(4):288–295
Luz S (2011) Web-based corpus software. In: Kruger A, Wallmach K, Munday J (eds) Corpus-based translation studies—research and applications, chapter 5. Continuum, pp 124–149
Luz S, Masoodian M (2004) A mobile system for non-linear access to time-based data. In: Costabile MF (ed.) Proceedings of Advanced Visual Interfaces AVI'04. ACM Press, pp 454–457
Luz S, Sheehan S (2014) A graph based abstraction of textual concordances and two renderings for their interactive visualisation. In: Proceedings of the international working conference on Advanced Visual Interfaces, AVI '14. ACM, New York, pp 293–296
Lyding V, Nicolas L, Stemle E (2014) interhist—an interactive visual interface for corpus exploration. In: Calzolari N, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA)
Mackinlay J (1986) Automating the design of graphical presentations of relational information. ACM Trans Graph 5(2):110–141
Marai GE (2018) Activity-centered domain characterization for problem-driven scientific visualization. IEEE Trans Vis Comput Graph 24(1):913–922
McEnery T, Wilson A (2001) Corpus linguistics: an introduction. Edinburgh University Press Series. Edinburgh University Press
Miller A (2018) Text mining digital humanities projects: Assessing content analysis capabilities of Voyant Tools. J Web Librariansh 12(3):169–197
Modnlp (2020) ModNLP software repository. http://modnlp.sf.net. Accessed Feb 2020
Moretti F (2005) Graphs, maps, trees: abstract models for a literary history. Verso
Munzner T (2009) A nested model for visualization design and validation. IEEE Trans Vis Comput Graph 15(6):921–928
Newman W, Lamming M (1995) Interactive system design. Addison-Wesley
Oelke D, Kokkinakis D, Keim DA (2013) Fingerprint matrices: uncovering the dynamics of social networks in prose literature. Comput Graph Forum 32(3 part 4):371–380

Olohan M (2002) Corpus linguistics and translation studies: interaction and reaction. Linguist Antverp 2002(01):419–429
Paley W (2002) Textarc: showing word frequency and distribution in text. In: Wong PC, Andrews K (eds) Proceedings of IEEE symposium on information visualization. Poster compendium. IEEE CS Press
Pecina P (2010) Lexical association measures and collocation extraction. Language Resour Eval 44(1–2):137–158
Rabadán R, Labrador B, Ramón N(2009) Corpus-based contrastive analysis andtranslation universals: a tool for translation quality assessment english and spanish Babel 55(4):303–328
Schmied J (1993) Qualitative and quantitative research approaches to English relative constructions. In: Souter C, Atwell E (eds) Proceedings of the International Computer Archive of Modern English, conference. pp 85–96
Scott M (2008) Wordsmith tools version 5. Lexical Analysis Software, Liverpool, p 122
Scott M et al. (2001) Comparing corpora and identifying key words, collocations, and frequency distributions through the wordsmith tools suite of computer programs. In: Ghadessy M, Henry A, Roseberry RL (eds) Small corpus studies and ELT. John Benjamins Publishing Company, pp 47–67
Sheehan S, Luz S (2019) Text visualisation for the support of lexicography-based scholarly work. In: Proceedings of the eLex 2019 conference on electronic lexicography in the 21st century, Sintra, Portugal. pp 694–725
Sheehan S, Masoodian M, Luz S (2018) COMFRE: a visualization for comparing word frequencies in linguistic tasks. In: Catarci T, Leotta F, Marrella A, Mecella M (eds) Proceedings of Advanced Visual Interfaces AVI'18. Association for Computing Machinery (ACM), pp 36–40
Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualizations. In: Green TRG (ed.) VL '96: Proceedings of the 1996 IEEE symposium on visual languages. IEEE Computer Society, Washington, pp 336–343
Sinclair J (1991) Corpus, concordance, collocation. Oxford University Press
Sinclair J (2003) Reading concordances: an introduction. Pearson/Longman
Svartvik J (2011) Directions in corpus linguistics: proceedings of nobel symposium 82, vol. 65, Stockholm, 4–8 August 1991. Walter de Gruyter
Tufte ER (1990) Envisioning information. Graphics Press, Cheshire
van Ham F, Wattenberg M, Viegas FB (2009) Mapping text with phrase nets. IEEE Trans Vis Comput Graph 15(6):1169–1176
Venuti L (1995) The translator's invisibility: a history of translation. Routledge
Viégas F, Wattenberg M (2008) Tag clouds and the case for vernacular visualization. Interactions 15(4):49–52
Voyant (2020) Voyant tools. https://voyant-tools.org/. Last accessed Feb 2020
Wattenberg M, Viégas FB (2008) The word tree, an interactive visual concordance. IEEE Trans Vis Comput Graph 14(6):1221–1228
Zanettin F (2001) Swimming in words: corpora, translation, and language learning. In: Aston G (ed.) Learning with corpora. Athelstan, p 177
Zanettin F (2013) Corpus methods for descriptive translation studies. Procedia - Soc Behav Sci 95:20–32

Acknowledgements

This research was supported by the Arts and Humanities Research Council, UK (Grant number: AH/M010007/1), and by the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.