Task-based Quantitative Evaluation of the Concordance Mosaic Visualization

Citation for published version:
Sheehan, S, Masoodian, M & Luz, S 2022, Task-based Quantitative Evaluation of the Concordance Mosaic Visualization. in *Proceedings of the 26th International Conference Information Visualisation, IV 2022.* Proceedings of the International Conference on Information Visualisation, vol. 2022-July, Institute of Electrical and Electronics Engineers Inc., pp. 123-129, 26th International Conference Information Visualisation, IV 2022, Vienna, Austria, 19/07/22. https://doi.org/10.1109/IV56949.2022.00028

Digital Object Identifier (DOI):
10.1109/IV56949.2022.00028

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 26th International Conference Information Visualisation, IV 2022

Task-based Quantitative Evaluation of the Concordance Mosaic Visualization

Shane Sheehan*, Masood Masoodian†, Saturnino Luz*,
*Usher Institute, The University of Edinburgh, UK
{shane.sheehan, s.luz}@ed.ac.uk
†School of Arts, Design and Architecture, Aalto University, Finland
masood.masoodian@aalto.fi

***Abstract*—Researchers working in areas such as lexicography, translation studies, and computational linguistics, use a combination of automated and semi-automated tools to analyze the content of text corpora. Concordancing – or the arranging of passages of a textual corpus in alphabetical order according to user-defined keywords – is one of the oldest and still most widely used forms of text analysis. Concordance Mosaic is an interactive concordance visualization which emphasises quantitative information such as word frequency. While Concordance Mosaic is in active use by humanities scholars, no quantitative evaluation of the technique exists. In this paper, the Concordance Mosaic is quantitatively evaluated in comparison to a typical concordance browser. The comparison is evaluated using speed and accuracy on identified corpus analysis actions.**

***Index Terms*—Visualization in humanities, text visualization, document data visualization, visual knowledge discovery, quantitative evaluation.**

I. Introduction

In many academic fields, corpus analysis is central to the study of texts. Computational tools have long been used for lexicography, corpus linguistics, and corpus-based translation studies [1]–[3], and new methods motivated by such tools have been more broadly applied to the study of policy in areas such as medicine [4] and politics [5]. One of the most popular techniques supported by computation is the indexing, retrieval and display of keyword-in-context. This technique dates back to at least the 1950's with Luhn's work on concordance indexing [6].

Corpus analysis using concordance and collocation (the habitual juxtaposition of a particular word with another word with a frequency greater than chance) provides a data-driven approach for corpus analysis, unlike more traditional scholarly work in these fields which requires close reading, researchers' prior knowledge, and theoretical frameworks to interpret texts. Data-driven corpus analysis techniques are heavily influenced by the work of John Sinclair and Michael Halliday [7], [8], in which 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 close reading of the texts. Visualization tools can aid this process by providing effective overviews and helping to identify patterns in the texts, as well as visual explanation of the analysis outputs [9].

Concordance Mosaic [10] is a visualization tool which has been adopted by members of the corpus linguistic community for corpus analysis, and the presentation of their scholarly work [5], [11] . The visualization provides an overview of the context words within a window of the selected keyword. Quantitative information – most often word frequency – is visually encoded allowing the analyst to identify positional collocation patterns around a single keyword.

II. Related work

Keyword visualisations often take the form of networks or clusters of keywords [12]–[14]. While these visualisations are useful for identifying similar or connected keywords in a corpus, they do not provide any insight into the collocations of the keywords, and as such, they are not directly usefully for comparing the contexts in which these keywords appear.

Corpus Clouds [15] is a frequency-focused corpus analysis tool. A word cloud, based on the tag cloud visualization [16], is used to encode the frequencies of all words returned by a corpus query. For quantitative tasks involving frequency estimation or comparison, the use of font size to encode value in cloud-based visualisations is a limitation [17]. Also, since positional collocations are not encoded in *Corpus Clouds*, the visualization is of an entire context window.

TagSpheres [18] are word cloud based visualization where keyword co-occurrences are encoded using an integral combination of color and radial position from the central keyword. The cloud layout places the same word from different positions close together to help identify strong collocation patterns. However, the linear structure of the text is removed making it difficult to identify multi-word collocation patterns.

Tree-based visualisations of keyword-in-context have therefore been proposed as a technique for encoding quantitative information while maintaining the linear structure of the text [10], [19], [20]. In practice, however, displaying a large number of concordance lines, where font size is used to encode represent positional frequency, requires a trade off between frequency estimation and readability. The variable length of words also makes encoding frequency using font size challenging – since area is not as perceptually efficient as length for visualising quantitative information.

None of these visualization systems have been quantitatively compared with keyword-in-context concordance browsers, which are the long standing tool of choice for the analysis of linguistic patterns surrounding keywords.

A. *Concordance Mosaic*

An example of a Concordance Mosaic for the keyword *Silver* is shown in Fig. 1. Each column of tiles represents a word position relative to the keyword. The height of each tile is encodes quantitative information, in Fig. 1 tile height encodes word frequency at the position. For example, using the Mosaic we can easily see that, in the corpus under investigation, the word *gold* is the most frequent co-occurrence at the position two to the left of the keyword *silver*. Concordance Mosaic offers four deiffert views, column frequency, coulmn frequency without stopwords, a view scales the tiles using positional collocation strength, and a view which scales the words using windowed collocation strength.

The Concordance Mosaic can be used together with a keyword-in-context concordance browser. One of the key interactions is concordance highlighting via Mosaic selection. In Fig. 2 we see the keyword-in-context browser linked to the mosaic in Fig. 1. The second most frequent word two positions to the left of the keyword is selected. This word is *talents*. The selection causes the keyword-in-context display to be sorted at the selected position, scrolled to the relevant part of the concordance list, and highlighted to show the selected word and concordance lines. When used in combination quantitative information and close reading can be used in combination to understand the context of a keyword.

III. User study

This study was designed to compare the performance of two visualization tools, the Concordance Mosaic and a textual keyword-in-context interface (KWIC). A third option was also tested, in which both the Concordance Mosaic and KWIC were available side-by-side (Juxtaposed Interface). The evaluation was performed on concordance analysis tasks for which the Concordance Mosaic was designed. These tasks have been identified from analysis of the corpus methodology [21], described by John Sinclair [22].

An initial heuristic evaluation and a pilot study were used to refine the visualization tools and test their usability prior to the main user study. During this study we found that tasks requiring analysis of multiple context words or positions were difficult for non-expert users to understand. Based on this we limited the evaluation to five simple quantitative analysis tasks, only one of which required looking at multiple positions.

The null hypothesis, in this study, is that there is no significant difference in performance between the interfaces. Performance was measured by the speed and accuracy with which participants completed corpus analysis tasks. These tasks were created with the quantitative actions found during our task analysis in mind. Each participant attempted to answer five questions using each interface, the order in which the interfaces were presented was randomised and balanced across participants for all possible combination of interface orderings. Each participant used each of the three visualization options, in balanced and randomised order. Similarly, for each visualization option a different keyword was required per question, and the keyword per option was balanced.

For the study we recruited thirty-six participants ($N = 36$) from the student population via an online university notice-board and mailing list. Since the study evaluates performance on quantitative tasks we decided previous experience with concordance tools or corpus analysis would not be a pre-requisite for participation. A pilot study was run with two participants who were not included in the main experiment, this was done to determine which areas of the interfaces, and any terminology, participants may have difficulty with. Informed by this pilot a tutorial was designed. The tutorial took approximately ten minutes to complete, it was given to each participant immediately before they participated. In this tutorial each of the required interface features were introduced and explained, any linguistic concepts required were also clarified and a researcher was available to answer questions.

A. *Software*

The Software we created to conduct the experiment consists of four major elements: the KWIC interface, the Concordance Mosaic interface, the question box and the answer box. The question box appears at the bottom left of the software (Fig. 3) and is simply a text area into which the questions and instructions are rendered. The answer box (bottom right Fig. 3) contains a button for proceeding to the next question, a button for resetting the software to the questions original state and a text box for the participant to enter the answer. Both the KWIC and Mosaic interfaces will be present when the software is displaying the Juxtaposed interface. However, when displaying either interface alone the space where the other usually resides will be blank. The experiment was instrumented to log all participant interactions with the software such as clicks, keystrokes and hovers. Screen recordings were also captured.

B. *Questions*

The participants were asked the same five questions on each of the three interfaces. The keywords about which they were being asked were different for each interface they encountered. We used three sets of keywords. The combination of circularly shifted keyword set orderings and interface orderings was balanced across the participants.

The selection of the keywords for each of the five questions was done in such a way as to standardise the difficulty of the question using the KWIC interface. For example, question two asks "*For the keyword KEYWORD, what is the most frequent word at position keyword - 1*?". The three keywords chosen for this question were *Wealthy*, *Daylight* and *Massive*. These keywords all returned a concordance with approximately three-hundred concordance lines, the most frequent word to the left of the keyword (position $k - 1$) occurs with a frequency of between 26 and 27%, and the second most frequent word at position $k - 1$ occurs with a frequency of 20 to 22%.

Similarly, question one is phrased in exactly the same way as question two but the chosen keywords change the distribution of the words at the position of interest. In this case there are again three-hundred lines in the concordance but the frequencies of the most common and second most common

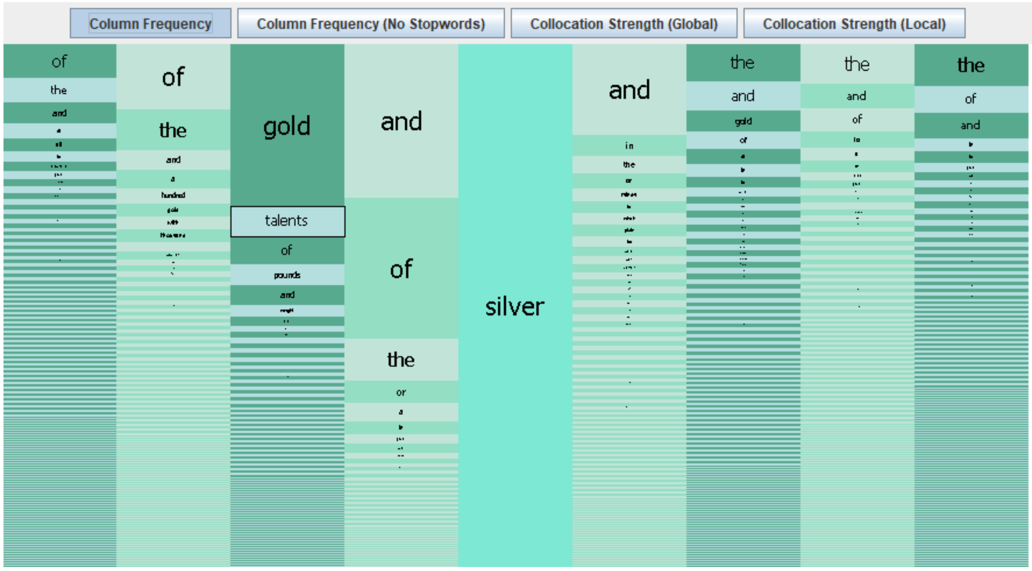

Fig. 1. Concordance Mosaic for the keyword *Silver*. The word *talents* at position keyword minus 2 is selected.

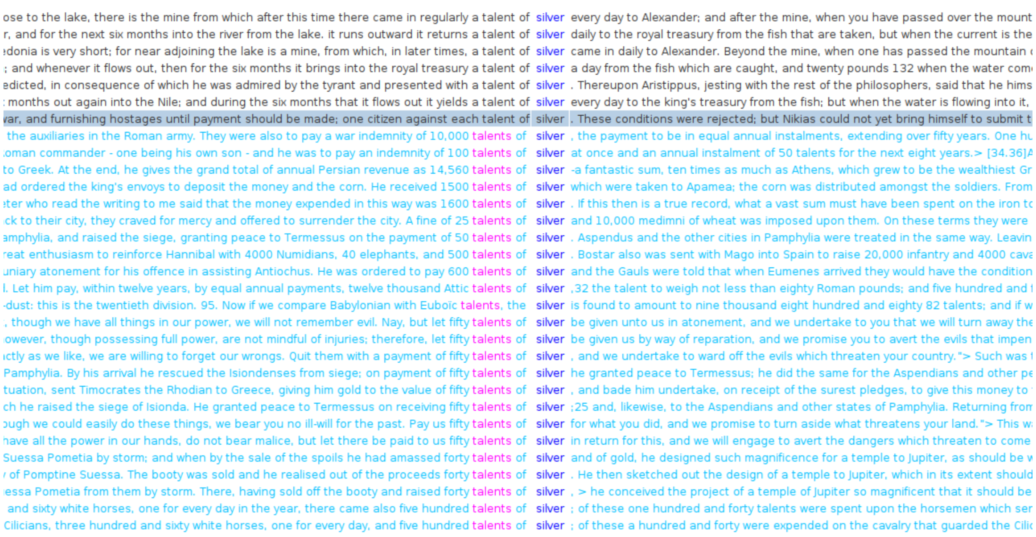

Fig. 2. Textual keyword-in-context linked to Concordance Mosaic. The Mosaic for the keyword *silver* is selected and the word *talents* at position keyword minus 2 was chosen using the Mosaic.

words at position $k-1$ are approximately 40% and between 5 and 10% percent, respectively.

Both questions one and two involve tasks which focus on frequency estimation. Question three has the same focus but, additionally, the participant needs to identify the part of speech of the words at the position of interest. This question asks "*For the keyword KEYWORD, what is the most frequent descriptive adjective at position keyword - 1?*" and a clarifying statement and example is given "*A descriptive adjective is a word which describes a noun (KEYWORD is the noun in this case). e.g. In the text fragment "an old book" old is an adjective describing the noun book*". For this question the chosen keywords return a concordances with approximately 1,000 lines where the correct adjective is the fifth most common word at position $k-1$ with a frequency of about 7%.

The fourth question asks the user to identify a frequent combination of words. Specifically, they are asked to identify the most frequent word at position $k-2$ when another specified word occurs at position $k-1$. An example is "*For the keyword "standing", focusing only on concordances that contain the word "still" at position keyword - 1, what word is most frequent at position keyword - 2?*". This question becomes much easier to answer if a filtering interaction is used, so a hint was provided telling the user to do so for all interfaces. They had been previously instructed during the tutorial on how to perform this interaction. The frequency of the answer after the filter interaction was approximately 50% of 200 occurrences.

Finally, question five asks the participant to identify the

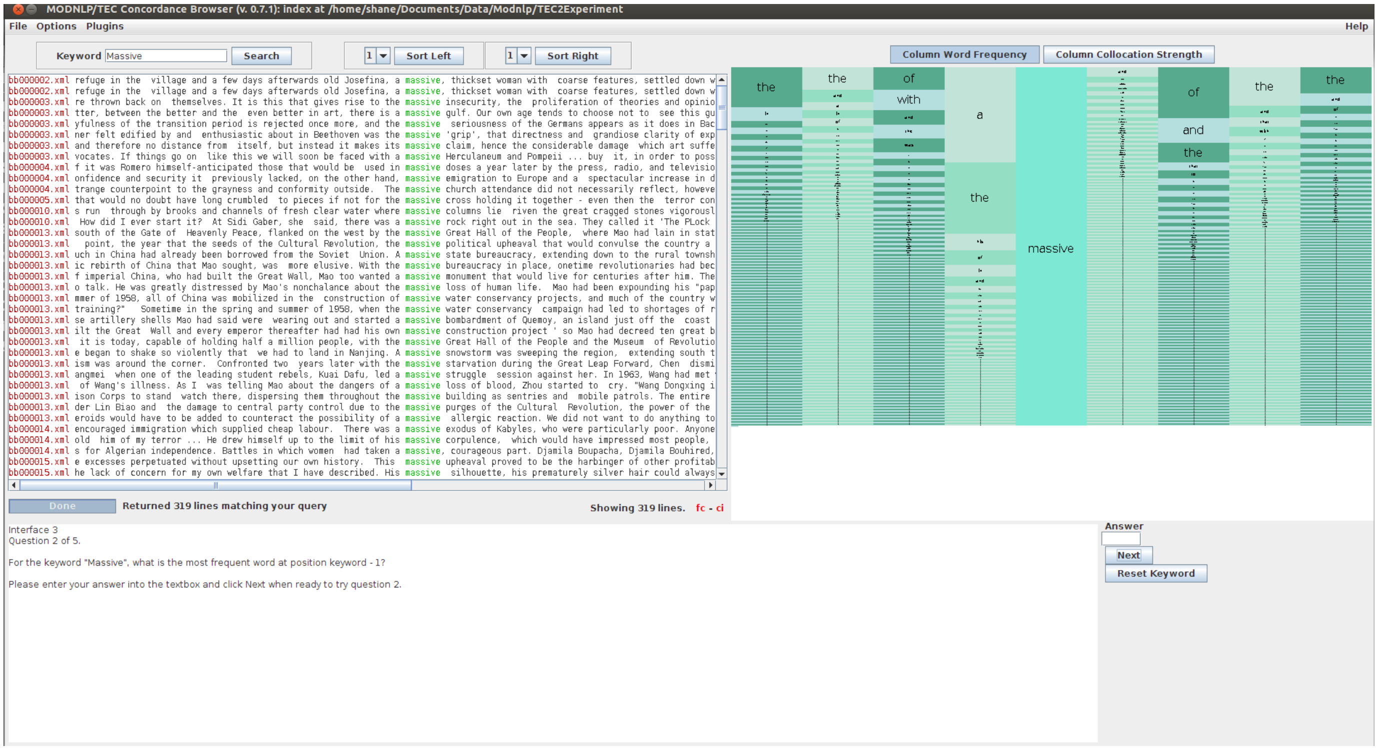

Fig. 3. Experimental Setup

word with the highest collocation strength at position $k - 1$. The correct answers have a positional collocation strength score of 50%, that is, the collocation strength of the word at that position is as strong as the combined collocation strength of all other words at that position. We expect this task to be the most difficult of the five when using the KWIC interface.

Looking again at these five questions, questions one and two both evaluate frequency estimation actions, and we expect question two to be more difficult as the two most common words have similar frequencies. Question three combines frequency estimation and a qualitative task of identifying parts-of-speech. This should also be more difficult than questions one and two due to the answer being the 5^{th} most frequent word. Question four is a mix of a frequent-combinations action and a filter action; the filter action is not required but makes the task much easier and is recommended to the participant. Question five is a collocation strength action which can be performed accurately using the KWIC interface by using frequency actions for each word or by using expert knowledge to evaluate only the most likely candidates.

C. Results

First let us look at the results of an ANOVA for the dependant variable t, time to complete a task, measured in seconds, with respect to the categorical variables; the question being answered (q), the interface being used (i), the participants assigned interface ordering (iOrder), the participants assigned keyword set ordering (qOrder) and a binary variable representing a correct or incorrect answer (isCorrect). The results of the ANOVA where a significant difference ($p < .05$) was found are given in Table I.

TABLE I
ANOVA RESULTS FOR THE DEPENDANT VARIABLE TIME t, WHERE P $< .05$

Independent Variable	F Value	P Value
q	26.9388	$< 2.2e^{-16}$
i	135.5089	$< 2.2e^{-16}$
isCorrect	4.3170	0.038894
q:i	8.0946	$1.428e^{-9}$
i:iOrder	2.8620	0.002261
i:qOrder	3.5232	0.008239
q:isCorrect	2.4258	0.048970

Since the main effects q, i and isCorrect all feature in significant interactions we focused our post-hoc analysis on these interactions instead of the main effects. We conducted Tukey's post-hoc tests (HSD) to analyse the different groupings of each interaction effect, using $p < .05$ to test for significance.

The result of the HSD test for the i and qOrder interaction (i:qOrder) showed a significant difference between two groupings. In this case the dataset was split into 9 groups by the combinations of the three interfaces and the three circularly shifted keyword set orderings. The HSD groupings simply combined these groups into data points where the KWIC interface was being used and a grouping of all data points where either the Mosaic or Juxtaposed interfaces were being used. This indicates that the interaction can be interpreted as i, and that qOrder can be safely ignored as it does not feature in any other significant interactions or as a main effect. This result shows, as expected, that our choice of keywords has not had a major effect on time to complete within each question.

The mean response times of the i:qOrder groups in which the KWIC interface was used were all greater than 67s, while the remaining groups containing the Mosaic and Juxtaposed data all had mean response times under 30s. This is evidence of interface choice having a large effect on response time.

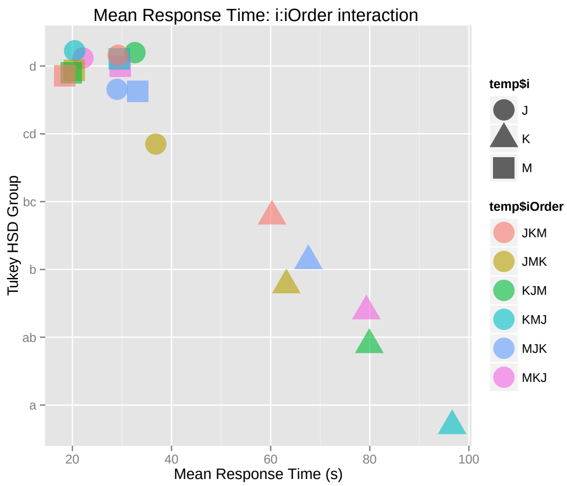

Fig. 4. Mean Response time of i:iOrder data groups. Clustered by Tukey HSD score

The HSD test for the interaction between i and iOrder (i:iOrder) examines the data set split into 18 groups on the combination of the three interfaces and the 6 possible interface orderings. The results of the HSD test found 6 significantly different groupings. We examined the result of the test as a scatter plot of these HSD groupings (Fig. 4). This scatter plot shows the mean response times of the eighteen i:iOrder groups, a slight jitter from the grouping lines was applied. Looking at this scatter plot we can see that 11 of the 12 groups which used the Mosaic (M) or Juxtaposed (J) interfaces are grouped together and all 12 have a mean response time of less than 40s. The remaining groupings all have mean response time greater than 60s and are the cases where the KWIC (K) interface was being used. Looking at the groups where the KWIC was in use, a learning effect can be observed in situations which the participant had used the Juxtaposed interface before KWIC faster response times were recorded. Interestingly, it appears that no significantly large learning effect takes place between the Mosaic and Juxtaposed interfaces. The only data point indicating a difference between the two interfaces when changing the ordering is the J:JMK data point.

The discovery of an interaction between q and i (q:i) is of great interest since our null hypothesis states: there are no significant differences between the interfaces. Analysing the groups created by splitting the data by interface and question we found a number of significant groupings (Fig. 5). For each question there is a significant difference between the KWIC interface and both the Mosaic and Juxtaposed interfaces. With the null hypothesis rejected we still look further at the results to investigate these differences.

Fig. 5. Mean Response time of i:q data groups. Clustered by Tukey HSD score

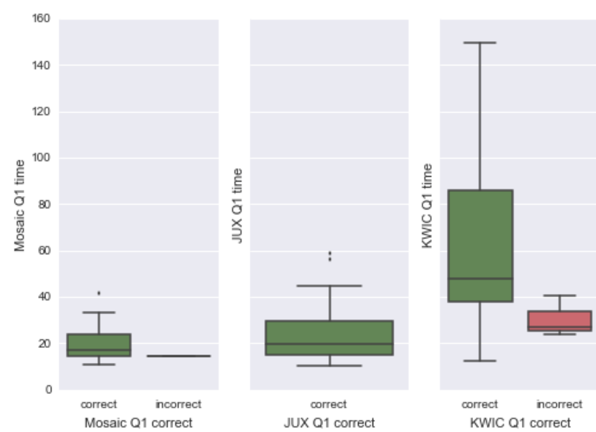

Fig. 6. Box plots of question one response times across the three interfaces

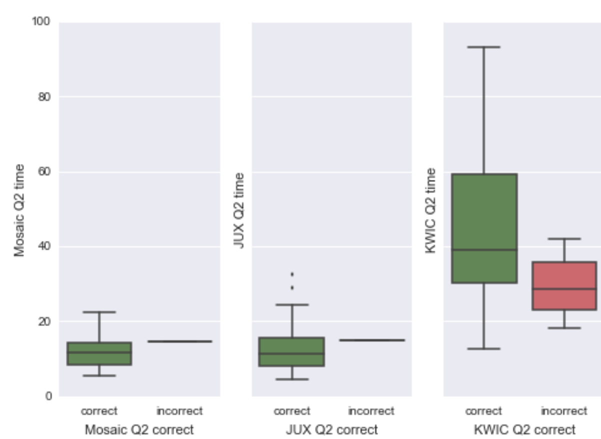

Fig. 7. Box plots of question two response times across the three interfaces

The Tukey HSD groupings of the i:q interaction show (Fig. 5) that for all questions (with the exception of question one Fig. 6) there was no significant difference between the response times per question for the Mosaic and Juxtaposed interfaces. For these interfaces question two (Fig. 7) was the quickest to complete, while question one, three (Fig. 8) and five (Fig. 10) took slightly longer and question four (Fig. 9) took even longer still. In comparison, on the KWIC interface question four was the third fastest to complete, while five takes the most time by a large margin, again question two is the quickest to complete. These plots show the 36 data points for each question and interface combination, the split of correct and incorrect answers can also be seen from these plots. Note the differing ranges on the y-axis between the plots. These plots show the difference between the KWIC response times and the other two interfaces.

In questions two and five we find a much larger number of incorrect answers when using the KWIC interface. Table II shows the number of incorrect answers per interface and question. Question three had the most incorrect answers but they were approximately evenly distributed among the interfaces. In the case of question five there were zero errors using the Mosaic or Juxtaposed interface while twenty-three of the thirty-six attempts using the KWIC interface were incorrect.

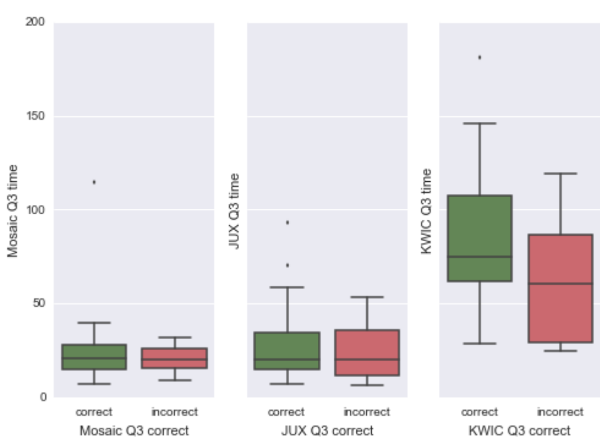

Fig. 8. Box plots of question three response times across the three interfaces

TABLE II
INCORRECT ANSWERS PER QUESTION AND INTERFACE

	Q1	Q2	Q3	Q4	Q5	Sum
J	0	1	9	2	0	12
K	3	7	8	4	23	45
M	1	1	11	6	0	19
Sum	4	9	28	12	23	76

IV. DISCUSSION

The null hypothesis, that there is no significant performance difference between the interfaces per question, has been firmly rejected. For the designed tasks the Mosaic and Juxtaposed interfaces have shown to be equivalent while the KWIC interface has a significantly worse performance on each of the five tasks. We could speculate that this may indicate that participants using the Juxtaposed interface, which combines both the Mosaic and KWIC, have a preference for the Mosaic interface as indicated by the similar performance statistics. Preliminary investigation of participant interaction logs and screen recordings also show a preference for using the Mosaic interface for these tasks when the choice is available.

The five questions cover a broad section of the quantitative actions we identified in our task analysis, and each of these actions feature in many corpus analysis tasks most often performed by text analysts. The Mosaic and Juxtaposed interfaces offer a large and significant performance increase over the standard method in the field should be seen as a significant contribution to corpus analysis.

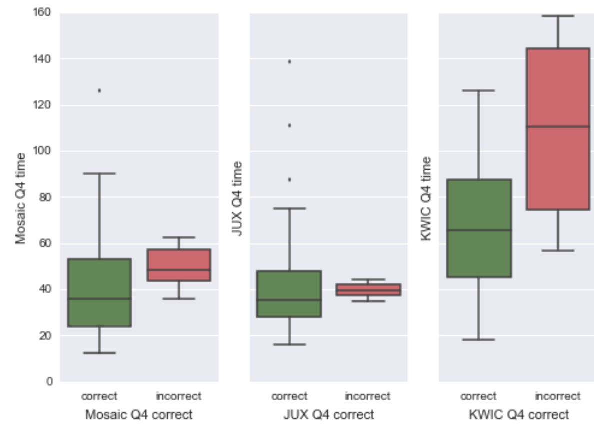

Fig. 9. Box plots of question four response times across the three interfaces

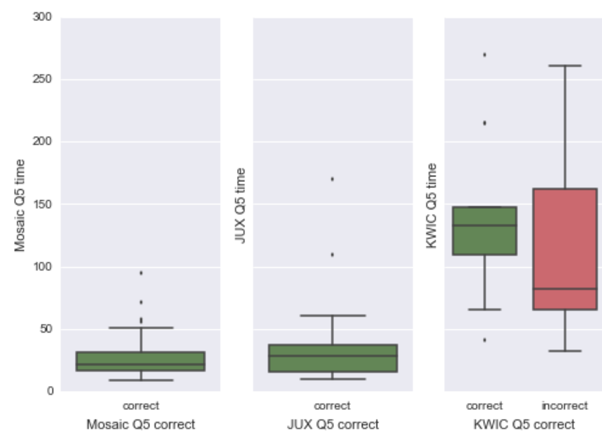

Fig. 10. Box plots of question five response times across the three interfaces

TABLE III
CORRECT ANSWERS PER QUESTION AND INTERFACE

	1	2	3	4	5	Sum
J	36	35	27	34	36	168
K	33	29	28	32	13	135
M	35	35	25	30	36	161
Sum	104	99	80	96	85	464

Looking at the performance between questions, we expected question two to be more difficult than one but the opposite appeared to be true, this is most likely due to participants learning from question one since both questions are the same, only the keywords and frequencies involved are different.

Using the Mosaic and Juxtaposed interfaces participants had the worst performance on question four. However this performance was still much better than the KWIC interface, where this question ranked third in performance. Question four involved the frequencies at two word positions, and a filter interaction much simplified the task. Since the other questions involve observing a frequency or collocation strength at a single position and no interaction is required, the performance decline using the Mosaic makes sense. Also, since the filter interaction was not available for the KWIC interface, its level of performance against questions three and five was expected.

On the Mosaic and Juxtaposed interfaces questions three and five had equivalent performance, the performance of the KWIC interface in terms of both time to complete and error rate on question five is worse by a large margin. From tables II and III we see that 23 of the 36 attempts of question five using the KWIC were incorrect. However, this task is less representative of a common KWIC analysis. These collocation or other statistics are usually calculated by an external tool and returned as a list with no reference to word position. What this question does show is that using the Mosaic interface and the concordance graph we can calculate positional statistics and can include them into our visual representation of the concordance in an easy to understand manner.

V. Conclusion & Future Work

In this paper we presented an assessment of the Concordance Mosaic, a space-filling visual technique for corpus analysis and exploration, in quantitative tasks of corpus analysis. The design and goals of this visualization are spelled out in terms of the common tasks performed using the traditional keyword-in-context (KWIC) concordance visualisation.

The user study had 36 participants perform five tasks using each visualization while we monitored accuracy and time to complete the tasks. This study revealed that, for each of the questions asked, the Mosaic and Juxtaposed Interfaces had large and significant performance increases when compared to the KWIC interface. The tasks included in the study at their core involved word frequency estimation, recognition of frequent word combinations, and investigation of corpus statistics. These tasks were identified during a task analysis as fundamental quantitative actions which form a part of most corpus analysis using a concordance browser. The Mosaic visualization technique offers a significant performance increase over the traditional KWIC display for these fundamental actions is a significant contribution.

Concordance Mosaic is under active development in collaboration with humanities scholars. Future work includes colouring Mosaics using corpus metadata and investigating interactive techniques for comparing multiple keywords simultaneously.

References

[1] J. Svartvik, *Directions in corpus linguistics: proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Walter de Gruyter, 2011, vol. 65.

[2] S. Bernardini and D. Kenny, "Corpora," in *The Routledge Handbook of Translation Studies*, M. Baker and G. Saldanha, Eds. Routledge, 2020, pp. 110–115.

[3] M. Baker, "Corpus linguistics and translation studies: Implications and applications," *Text and technology: In honour of John Sinclair*, vol. 233, p. 250, 1993.

[4] J. Buts, M. Baker, S. Luz, and E. Engebretsen, "Epistemologies of evidence-based medicine: a plea for corpus-based conceptual research in the medical humanities," *Medicine, Health Care and Philosophy*, pp. 1–12, 2021.

[5] J. Buts, "Community and authority in ROAR Magazine," *Palgrave Communications*, vol. 6, no. 1, pp. 1–12, 2020.

[6] H. P. Luhn, "Key word-in-context index for technical literature (kwic index)," *American Documentation*, vol. 11, no. 4, pp. 288–295, 1960.

[7] J. Léon, "Meaning by collocation," in *History of Linguistics 2005*. John Benjamins, 2007, pp. 404–415.

[8] J. Sinclair, *Corpus, Concordance, Collocation*. Oxford Un. Press, 1991.

[9] E. R. Tufte, *Envisioning information*. Cheshire: Graphics Press, 1990.

[10] S. Luz and S. Sheehan, "A graph based abstraction of textual concordances and two renderings for their interactive visualisation," in *Procs of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '14. New York: ACM, 2014, pp. 293–296.

[11] M. Baker, "Rehumanizing the migrant: the translated past as a resource for refashioning the contemporary discourse of the (radical) left," *Palgrave Communications*, vol. 6, no. 1, pp. 1–16, 2020.

[12] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller, "Visualization as seen through its research paper keywords," *IEEE Trans on Visualiz and Comp Graph*, vol. 23, no. 1, pp. 771–780, 2017.

[13] H. Li, H. An, Y. Wang, J. Huang, and X. Gao, "Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network," *Physica A: Statistical Mechanics and its Applications*, vol. 450, pp. 657–669, 2016.

[14] J. Choi and Y.-S. Hwang, "Patent keyword network analysis for improving technology development efficiency," *Technological Forecasting and Social Change*, vol. 83, pp. 170–182, 2014.

[15] C. Culy and V. Lyding, "Corpus clouds - facilitating text analysis by means of visualizations," in *Human Language Technology. Challenges for Computer Science and Linguistics*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6562, pp. 351–360.

[16] F. Viégas and M. Wattenberg, "Tag clouds and the case for vernacular visualization," *Interactions*, vol. 15, no. 4, pp. 49–52, 2008.

[17] C. Felix, S. Franconeri, and E. Bertini.

[18] S. Jänicke and G. Scheuermann, "On the visualization of hierarchical relations and tree structures with tagspheres," in *Computer Vision, Imaging and Computer Graphics Theory and Applications*, J. Braz, N. Magnenat-Thalmann, P. Richard, L. Linsen, A. Telea, S. Battiato, and F. Imai, Eds. Springer, 2017, pp. 199–219.

[19] C. Culy and V. Lyding, "Double tree: An advanced kwic visualization for expert users," in *Information Visualisation (IV), 2010 14th International Conference*, July 2010, pp. 98–103.

[20] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1221–1228, 2008.

[21] S. Luz and S. Sheehan, "Methods and visualization tools for the analysis of medical, political and scientific concepts in genealogies of knowledge," *Palgrave Communications*, vol. 6, no. 1, pp. 1–20, 2020.

[22] J. Sinclair, *Reading concordances: an introduction*. Pearson, 2003.