FreeTxt: Analyse and Visualise Multilingual Qualitative Survey Data for Cultural Heritage sites

Nouran Khallaf[4], Ignatius Ezeani[1], Dawn Knight[2], Paul Rayson[1], Mo El-Haj[1,3], John Vidler[1], James Davies[2], Fernando Alva-Manchego[2]
[1]Lancaster University, UK, [2]Cardiff University, UK,
[3]VinUniversity, Vietnam, [4]Leeds University, UK.
freetxt@cardiff.ac.uk

Abstract

We introduce FreeTxt, a free and open-source web-based tool designed to support the analysis and visualisation of multilingual qualitative survey data, with a focus on low-resource languages.[1] Developed in collaboration with stakeholders, FreeTxt integrates established techniques from corpus linguistics with modern natural language processing methods in an intuitive interface accessible to non-specialists. The tool currently supports bilingual processing and visualisation of English and Welsh responses, with ongoing extensions to other languages such as Vietnamese. Key functionalities include semantic tagging via PyMUSAS, multilingual sentiment analysis, keyword and collocation visualisation, and extractive summarisation. User evaluations with cultural heritage institutions demonstrate the system's utility and potential for broader impact.

1 Introduction

Qualitative surveys and open-ended questionnaires are widely used across domains and sectors to collect in-depth textual feedback. While such data provides rich insights, organisations working in these areas often lack the technical expertise to analyse free-text responses effectively. This is particularly challenging in multilingual and low-resource language scenarios, where readily available tools are often limited or non-existent.

Several digital platforms and tools have been developed to support researchers and analysts in working with textual data. Tools such as Voyant (Sinclair and Rockwell, 2016), AntConc (Anthony, 2019), and LIWC (Pennebaker et al., 2015) provide frequency-based and collocation analysis, while Sketch Engine (Kilgarriff et al., 2014) allow for complex corpus queries and linguistic annotations. However, they have limited support for multilingual data (especially low-resource languages), require some level of linguistic or technical expertise, lack interactive visualisations, or are not openly accessible for all users.

In response to these limitations, **FreeTxt** offers a free, open-source online tool for **interactive, multilingual analysis and visualisation of qualitative data**. FreeTxt provides several key functionalities:

- Bilingual support for English and Welsh, including multilingual sentiment analysis based on a fine-tuned BERT model (Devlin et al., 2019) and semantic tagging via PyMUSAS (Rayson et al., 2004);
- Interactive visualisations such as word clouds, collocation networks, and word trees, all of which can be customised and exported;
- Close reading tools including keyword in context (KWIC) displays;
- Extractive summarisation using the TextRank algorithm (Mihalcea and Tarau, 2004); and
- An intuitive user interface, co-designed with non-technical professionals.

In this paper, we present the FreeTxt system, describing its architecture and core functionalities.[2] Feedback received from stakeholders demonstrates how scrutiny of the unique context of a specific minoritised language, and meaningful collaboration with potential user groups, can determine the design and construction of language resources (Knight et al., 2024). Ongoing work extends FreeTxt to additional language pairs, such as English–Vietnamese (*FreeTxt-Vi*), through international collaboration.[3]

[1] https://freetxt.app/

[2] See a short introductory video here: https://youtu.be/eEMYLSQSvhs?si=8GasO40Bex646EK-

[3] https://vinnlp.com/freetxt-vi

Proceedings of Recent Advances in Natural Language Processing,pages 541–545
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-063

2 Design and Implementation

To accommodate needs for scalability, flexibility, and integration with advanced NLP/CL tools, FreeTxt initially began as a Streamlit-based[4] prototype but transitioned to a Flask-based[5] architecture. While Streamlit was sufficient for basic analysis, it struggled with large datasets, lacked adequate security for sensitive data, and provided limited visualisation clarity and analytical flexibility.

FreeTxt accepts input from a variety of file formats, including `.txt`, `.csv`, and `.xlsx`, enabling users to either directly paste text, work with pre-loaded survey files, or upload their own files.

After uploading data, FreeTxt processes the text using configurable options, such as splitting the text into individual sentences or handling it as a single entity. FreeTxt then applies tokenisation and linguistic analysis to extract relevant textual features. Key outputs include word frequency distributions and n-grams (bigrams, trigrams, etc.), enabling users to analyse recurring word sequences and linguistic patterns within the dataset.

To achieve these capabilities, FreeTxt integrates a range of advanced NLP libraries and models. The platform incorporates key libraries, such as `spaCy`, `NLTK`, and `Scattertext`, to ensure efficient and scalable processing of large datasets.

In developing FreeTxt, one of the core challenges was ensuring the tool would be accessible to users without specialist knowledge of NLP or corpus linguistics (CL). To make FreeTxt accessible, particularly for those working in cultural heritage sectors, it was essential to hide these complexities behind an intuitive and user-friendly interface.

Visualisation played a significant role in this approach. We prioritised clear, engaging graphical representations such as word clouds, collocation networks, and sentiment charts, which allow users to explore and interpret data without needing to understand the underlying computational processes. Furthermore, rather than presenting technical terms like "keyness" or "collocation" directly, the tool focuses on the outcome—showing patterns and trends in ways that are easy to interpret.

One of the key functionalities is the **sentiment analysis** feature, which utilises a BERT-based multilingual sentiment analysis model.[6] This model classifies text into positive, neutral, and negative sentiments, capturing broader contextual information beyond word-level tagging. Trained on multilingual product reviews, the model achieves an accuracy of approximately 95% for English texts.[7] The Welsh sentiment model, fine-tuned and evaluated on manually annotated data provided by one of our project partners, achieved an accuracy of around 73%. This sentiment analysis functionality is particularly valuable for understanding feedback, reviews, and general public sentiment.

FreeTxt provides tools for analysing **n-grams** (word sequences) and **word frequency distributions**. FreeTxt also incorporates a keyword collocation and network analysis tool, which enables users to visualise word relationships through interactive collocation networks. This feature highlights patterns in word usage, allowing users to explore connections between words as they co-occur in the text, providing meaningful insights into how terms relate to one another within the dataset.

The platform offers several visualisation options to help users interpret the results of n-gram and word frequency analysis including word clouds arranged in a bespoke variety of shapes for export into reports, with keyness calculations relative to the British National Corpus (BNC)[8] for English and CorCenCC[9] (Knight et al., 2018) for Welsh.

FreeTxt supports the generation of word clouds based on semantic tags from the USAS semantic tagger (PyMUSAS, Rayson et al., 2004),[10] which is available for multiple languages including English and Welsh. This allows users to visualise thematic trends through semantic groupings, making it particularly useful for cross-linguistic analysis.

FreeTxt also supports users in close reading with Key Word in Context (KWIC) displays, and visualises collocation networks to identify patterns where certain words frequently appear together. The length of the connector lines in these networks represents the proximity of the relationship between the searched word and its collocates, with shorter lines indicating closer associations. Additionally, the thickness of the lines reflects the frequency of the co-occurrence, with thicker lines representing more frequent word pairings.

[4] https://streamlit.io/

[5] https://flask.palletsprojects.com/en/3.0.x/

[6] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[7] As reported on the Hugging Face Model card

[8] http://www.natcorp.ox.ac.uk/

[9] https://www.corcencc.org/

[10] https://ucrel.github.io/pymusas/

Adapting Google Charts word tree,[11] this tool allows users to explore branching relationships between keywords and their surrounding phrases. This is particularly useful for identifying patterns of word usage and tracing how specific terms are connected to various expressions across the text. By providing a hierarchical view of word relationships, the word tree enables users to uncover both direct and indirect connections between words, offering deeper insights into the contextual structure of the dataset, as shown in Figure **??**.

For text summarisation, FreeTxt leverages the TextRank algorithm, a graph-based ranking model for text processing that extracts the most important content from the input text by determining the relative significance of sentences (Mihalcea and Tarau, 2004; Ezeani et al., 2022; El-Haj et al., 2022). Users can also select the compression ratio of the original text, with options ranging from 10% to 50%, allowing them to retain varying levels of detail depending on their needs. This functionality is particularly useful for generating high-level overviews of lengthy documents, offering users a quick way to grasp key insights without losing the essential information.

3 Evaluation and System Comparison

We evaluate FreeTxt's performance, unique functionalities, and user experience by comparing it with other widely-used text analysis platforms. We focus on key aspects such as multilingual support, visualization capabilities, customization options, and system scalability.

3.1 System Components

One of FreeTxt's main strengths is its ability to handle both English and Welsh text via its NLP/CL methods and tools, but also with a bilingual user interface, both features not commonly supported by existing tools. This is further enhanced by the integration of the PyMUSAS semantic tagger, which allows for detailed semantic analysis across both languages. Compared to tools such as Voyant (Sinclair and Rockwell, 2016) and AntConc (Anthony, 2019), which primarily offer frequency-based analyses and basic concordance functionalities, FreeTxt provides the additional annotation layers of semantic and sentiment analysis, resulting in a richer set of visualisation options.

[11] https://developers.google.com/chart/interactive/docs/gallery/wordtree

FreeTxt supports multiple file formats such as `.txt`, `.csv`, and `.xlsx`, making it versatile for different input sources from user's survey outputs. It allows users to perform customizable analysis through options like sentence splitting, exclusion of specific sentences, and adjusting summarisation ratios. These features enhance overall usability, enabling users to tailor the analysis to their specific needs—offering more flexibility than options available in platforms such as LIWC (Pennebaker et al., 2015) and NVivo (QSR International, 2021).

FreeTxt embodies several key features:

Multilingual Support: FreeTxt's ability to handle both English and Welsh text analysis sets it apart from other tools such as AntConc, Sketch Engine (Kilgarriff et al., 2014), which either do not support Welsh or offer limited bilingual functionality. While Sketch Engine is known for its extensive corpus resources and powerful linguistic tools, it is a subscription-based platform, making FreeTxt's open-source model more accessible, especially for users with budget constraints. Additionally, FreeTxt's integration of PyMUSAS for semantic tagging enhances its capability to analyze thematic content in both languages, a feature not fully developed in other platforms.

Visualisation Capabilities: FreeTxt excels in providing advanced visualisation features, such as frequency-based word clouds, semantic word clouds, and interactive collocation networks. These visualisations, as shown in Appendix D, offer a more intuitive and interactive approach to understanding text data compared to the static visualisations offered by platforms like LIWC, NVivo, or Wmatrix. While tools such as Sketch Engine provide advanced corpus queries, they often lack the same level of dynamic visualisations that FreeTxt offers for exploring collocations and semantic relationships interactively.

Customisation and Flexibility: FreeTxt enables users to customise their analysis through sentence splitting, exclusion of specific sentences, and the generation of downloadable reports in PDF format. In contrast, tools like Sketch Engine, AntConc, and Voyant, while powerful, offer less flexibility for user-defined configurations and report generation. FreeTxt's flexibility in offering these options makes it more adaptable for a wider range of research projects. Furthermore, being an open-source platform, FreeTxt offers an accessible solution to users who might be deterred by the

subscription fees of tools like Sketch Engine.

3.2 Qualitative System Components Comparison

Key technical code and linguistic resources for the bilingual analysis of English and Welsh text were available to the FreeTxt team from the previous CorCenCC and related projects. As explained in section 2, these fall into two main types implementing methods from both NLP and CL. Long standing CL methods such as frequency lists, concordances, collocations, key words and n-grams (sometimes known as lexical bundles or clusters) are ubiquitous in other tools such as AntConc[12], LancsBox[13] and Wmatrix[14], and we follow standard methods, statistical metrics and settings. Such methods are also common in other CAQDAS (Computer Assisted Qualitative Data Analysis)[15] tools employed in other academic disciplines. However, NLP methods such as summarisation, sentiment, part-of-speech and semantic analysis, are much less commonly supported in CL and CAQDAS tools which usually only operate at the lexical level (although this is starting to change in LancsBox and Wmatrix).

3.3 Integration of Stakeholder Requirements

As prototype versions of FreeTxt were developed, the project team sought feedback from the project partners to ensure that the functionalities of the tool were fit-for-purpose within the context of their own institutional needs/practices. This process of feedback was iterative. Guided demonstrations and verbal updates were provided by members of the project team, along with walk-through task sheets which required partners to upload their own data and evaluate the ease with which the toolkit could be navigated, the results of which would be fed back to the project team. A crucial aspect of the stakeholder engagement was input on terminology used in the interface, to ensure that the naming or labelling of tools was logical and intuitive. For example, 'linguistic annotation' and 'collocation' are unknown terms to non language specialists. We also needed to carefully define and explain terms that are potentially known to non experts e.g. 'sentiment', 'summarisation', and 'word cloud', in order to ensure that end users are not making unexpected assumptions about how the NLP or CL methods worked, and how explainable or accurate they are in order to correctly set expectations for their results.

4 Conclusion

FreeTxt offers three key advantages over existing tools. First, multilingual semantic and sentiment analysis. FreeTxt's ability to seamlessly analyse both English and Welsh text provides advanced functionalities such as semantic tagging and interactive visualisation tools. This makes it particularly suitable for bilingual datasets, unlike other existing tools. Second, FreeTxt provides interactive and advanced visualisations. FreeTxt's visualisation features, powered by libraries such as PyVis and NetworkX, offer a more engaging user experience compared to the static output formats found in many other tools. Its word clouds and collocation networks provide dynamic ways to explore linguistic patterns, while platforms like Sketch Engine or Wmatrix typically rely on static graphs or simpler visualisations. Third, it offers customisation and ease of use for non specialists. FreeTxt's highly customizable options, such as excluding irrelevant sentences or generating detailed downloadable reports, provide a level of flexibility that is not commonly seen in other platforms. This makes FreeTxt a more adaptable tool for users with specific research needs.

However, FreeTxt also has areas that could benefit from further development. For instance, the accuracy of sentiment analysis for Welsh texts, currently at 73%, could be improved through the incorporation of additional training data. Moreover, while FreeTxt performs well for small to medium-sized datasets, further optimization could be implemented to enhance its performance when processing very large corpora.

In conclusion, FreeTxt provides a comprehensive and flexible solution for text analysis, especially for users working with bilingual datasets. Its advanced visualisation tools, semantic tagging capabilities, and customizable analysis options make it a competitive platform in the landscape of text analysis tools. Compared to existing tools, FreeTxt offers significant advantages in terms of multilingual support, flexibility, and interactivity, making it an ideal choice for researchers and professionals in need of a versatile text analysis solution.

[12] https://www.laurenceanthony.net/software/antconc/

[13] https://lancsbox.lancs.ac.uk/

[14] https://ucrel.lancs.ac.uk/wmatrix/

[15] https://www.surrey.ac.uk/computer-assisted-qualitative-data-analysis

Acknowledgements, Ethics, Source Code and Visualisations

Supplementary appendices containing further information can be found in the full version of this paper at https://github.com/UCREL/FreeTxt-Demo-Paper-RANLP2025.

References

Laurence Anthony. 2019. Antconc (version 3.5.8) [computer software].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris, and Dawn Knight. 2022. Creation of an evaluation corpus and baseline evaluation scores for Welsh text summarisation. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*. European Language Resources Association (ELRA).

Ignatius Ezeani, Mahmoud El-Haj, Jonathan Morris, and Dawn Knight. 2022. Introducing the Welsh text summarisation dataset and baseline systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5097–5106, Marseille, France. European Language Resources Association.

Adam Kilgarriff et al. 2014. The Sketch Engine: ten years on. In *Lexicography*, pages 7–36. Springer.

Dawn Knight, Nouran Khallaf, Paul Rayson, Mahmoud El-Haj, Ignatius Ezeani, and Steve Morris. 2024. FreeTxt: A corpus-based bilingual free-text survey and questionnaire data analysis toolkit. *Applied Corpus Linguistics*, 4(3):100103.

Dawn Knight, Steve Morris, Paul Rayson, Jonathan Morris, Mared Williams, and Tess Fitzpatrick. 2018. CorCenCC: The National Corpus of Contemporary Welsh. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

QSR International. 2021. Nvivo (version 12) [computer software].

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. In *Proceedings of the Beyond Named Entity Recognition Semantic labelling for NLP tasks at LREC 2004*, Lisbon, Portugal. European Language Resources Association (ELRA).

Stéfan Sinclair and Geoffrey Rockwell. 2016. Voyant tools: A web-based reading and analysis environment. In *Digital Humanities 2016*.