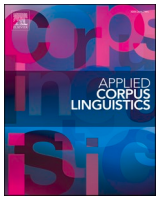
Applied Corpus Linguistics 4 (2024) 100103

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp



Articles

FreeTxt: A corpus-based bilingual free-text survey and questionnaire data analysis toolkit

Dawn Knight [a,*], Nouran Khallaf [b], Paul Rayson [b], Mahmoud El-Haj [b], Ignatius Ezeani [b], Steve Morris [a]

[a] *Cardiff University, Wales*
[b] *Lancaster University, England*

ARTICLE INFO

Keywords:
Corpus tools
Qualitative analysis
Free-text responses
Questionnaires

ABSTRACT

Qualitative free-text responses (e.g. from questionnaires and surveys) pose a challenge to many companies and institutions which lack the expertise to analyse such data with ease. While a range of sophisticated tools for the analysis of text *do* exist, these are often expensive, difficult to use and/or inaccessible to non-expert users. These tools also lack support for the analysis of English *and* Welsh text, which can be a particular challenge in the bilingual context of Wales. This paper details the key functionalities of the first corpus-based 'FreeTxt' toolkit which has been designed to support the systematic analysis and visualisation of free-text data, as a direct response to these two key needs. This paper demonstrates how, by working in partnership, software engineers, natural language processing (NLP) experts and corpus linguists can collaborate with end-users and beneficiaries to provide effective solutions to real world problems. Through the development of FreeTxt (www.freetxt.app), we aimed to empower end-users to *direct* and lead their own analyses of both small-scale and more extensive datasets to maximise the reach and potential impact generated. The approaches reported here, and the bilingual toolkit developed, can be replicated and extended for use in other language contexts and across a range of public and professional sectors. FreeTxt is now available for the analysis of Welsh and/or English, for use by *anyone* in *any sector* in Wales and beyond.

1. Introduction

Texts in a corpus can be derived from a range of different resources, sources and/or contributors. A digitised collection of student essays has the potential to be a corpus, as does a collection of interview transcripts, digitised court room hearings, or business meeting minutes. Corpora are effectively everywhere – they are not restricted to the academic domain – but not everyone necessarily knows i) what a 'corpus' is, or ii) how best to analyse one if a need arises to do so.

Corpora, corpus methods and the analytical tools that facilitate corpus analysis are predominantly designed for, and used by, trained experts/analysts based in academic institutions and/or publishing houses (e.g. the multi-billion-word Cambridge English Corpus (CIC) which has restricted access and is primarily accessible to colleagues affiliated with Cambridge University Press). There is often a distance between these expert-users and the potential future end-users and beneficiaries of analyses and insights derived from corpora. The level of engagement that end-users and beneficiaries typically have with corpus resources is effectively 'indirect', insofar as they do not engage with corpora or corpus query tools themselves (Leech, 2006). The use of corpora in some forms of Data-Driven Learning (DDL - (Johns, 1991)) practices provides the only exception to this tendency for 'indirect' engagement (although not all: approaches to DDL can often be mediated by materials writers and/or teachers, again creating engagement between the learner and corpus resources – known as 'DDL hands-off', (Leech, 2006)). In DDL hands-on, 'direct' engagement with a corpus is understood to enhance learner autonomy (Aston, 2001; Little, 2007); rather than teachers telling students how the language works, students are supported in working it out for themselves.

The methods and tools used by corpus linguists arguably have the potential for wider application in numerous professional and public domains beyond the pedagogic context. However, to fully realise this potential, a paradigmatic shift in the development and use of corpus-based utilities is required. Through the development of FreeTxt, we

* Corresponding author.
E-mail address: KnightD5@Cardiff.ac.uk (D. Knight).

https://doi.org/10.1016/j.acorp.2024.100103
Received 2 May 2024; Received in revised form 7 August 2024; Accepted 16 August 2024
Available online 23 August 2024

aimed to address this need by creating a toolkit which empowers end-users to direct and lead their own corpus-based analyses, and to query/probe their own corpus dataset, with a focus on a specific type of data: qualitative feedback (i.e. often free-text comments/written feedback). The design and construction of FreeTxt focused primarily on repurposing existing open-source tools and approaches as a means of extending their potential reach and relevance, rather than constructing a new tool from the ground up.

The main aim of the FreeTxt project was, thus, to create a bilingual free-text analysis and visualisation tool that i) integrates a number of tools and techniques from a range of projects in corpus linguistics and NLP, ii) is freely available, intuitive and user-friendly, iii) works with Welsh and English language data and iv) has the flexibility to be utilised by a range of different end-users and stakeholders. The first goal of this current paper is, thus, to provide a full walk-through of the FreeTxt toolkit produced in response to this aim.

For the effective transition from academic research to public utility to be fully realised, an iterative, collaborative approach to the design and development of tools and analytical workflows was used by the project team. Robustly reporting these approaches is often challenging, but this is the second key goal of this paper: to chart a problem-based approach for working with non-academic partners, as used in the design and development of FreeTxt. This approach capitalises on user-led introspection and reflection, along with dynamic, full-team engagement with new ways to thinking about language, data and the reporting of feedback. This approach (and the FreeTxt tool itself) can be replicated and extended for use in other language contexts and across a range of public and professional sectors.

2. Context

In a modern consumer-led culture, obtaining and responding to qualitative feedback is embedded in the professional practice of many walks of life. Surveys and questionnaires are used, for example, in staff development, professional training, product design and testing, and in various forms of service provision across the public and private sector. Surveys and questionnaires often produce a combination of quantitative and qualitative forms of data. Quantitative forms, such as rating scales (e.g. Likert scale responses), multiple choice questions and rank order questions can be numerated (i.e. quantified) with ease, and the analysis of which can be conducted in a systematic and often automated way. By contrast, qualitative questions, which prompt open ended, free-text comment responses (e.g. written feedback from exhibitions, events and/or historical sites on social media channels or websites including Trip Advisor and Trustpilot) pose a more difficult challenge for the analyst. Tackling written, text-based feedback often requires a more labour-intensive and manual approach to analysis than quantitative-based data. However, a range of existing digital tools do exist to facilitate the qualitative data analysis process, a sample of such CAQDAS (Computer Assisted Qualitative Data Analysis Software) tools are seen in Table 1.

These tools target both academic (e.g. ATLAS.ti) and commercial markets (e.g. Signal AI), and support several common functionalities, including tagging/coding, linking, mapping, querying, analysing and visualising qualitative data. Some of these tools include specific provisions for automatic sentiment analysis (e.g. Keatext), the generation of bar charts (e.g. Qualtrics), network and tree maps, word clouds (e.g. Displayr), word frequency (e.g. NVIVO), KWIC (e.g. MAXQDA) and thematic analysis (e.g. Dedoose), geospatial visualisations (e.g. Signal AI), and AI techniques (e.g. ATLAS.ti), for example.

Whilst some of these resources enable automatic coding and analysis (e.g. Signal AI), other tools such as NVivo require a more manual (and labour intensive) approach to annotation and analysis. Furthermore, whilst some of these tools have been built specifically to support the analysis of survey data (e.g. Qualtrics), this is not the case for all of them. In addition to this, whilst some of these tools are free to use, others incur high fees and/or subscription costs. Finally, whilst some of these tools have the ability to support the analysis of a number of (typically major) languages (e.g. Qualtrics), they often lack the ability to fully support the task of systematically processing feedback when it is presented in more than one language (particularly under resourced, minority and/or lesser used languages), which is often the case worldwide, and is certainly the case in the specific context of Wales.

Wales represents the largest bilingual community in the UK, with the 2021 census estimating that there are 562,000 speakers of Welsh in Wales (19 % of the total population of Wales – 2955,841 – (ONS 2011)), although annual population surveys consistently indicate a higher

Table 1
Commonly used software for annotating and/or analysing qualitative text data.

Name and accessibility[1]	Key utilities
ATLAS.ti https://atlasti.com/ Free with limited functionalities, £16 p/m individual license	Supports the flexible import and coding/annotation of text, video, audio and images. Includes networks, tree mapping, word frequency tools as well as tables, charts and other visualisations. Integrated with Open-AI to facilitate AI (Artificial Intelligence) driven coding, summarisation and content quoting.
Dedoose www.dedoose.com Circa £14.30 p/m individual license	Supports the management, coding and categorisation of text, image, audio and video files, enabling users to organise and explore data according to pattern, themes and trends (both qualitative and quantitative). Produces visualisations and reports based on the analyses.
Displayr www.displayr.com Free for small data files, £2309 p/a for professional use	Corporate tool for market research and data visualisation, used by several global companies from Meta to Amazon. Includes sentiment analysis, entity extraction, word cloud creation toolkits and visualisations. Data can be easily exported for integration into websites, presentations etc.
Keatext www.keatext.ai/en/ Free for small data files, circa £435 p/m for professional use	Corporate AI based text analysis toolkit which includes text categorisation, trend, sentiment and customer experience analysis tools, along with a range of visualisation charts, graphs and reports. This tool has the flexibility to be customised to specific user domains and supports analysis in multiple languages.
MAXQDA www.maxqda.com/ Circa £39 for six months for a student license	Mixed-methods data analysis tool for researchers which enables the importing, organisation and coding of text, audio, video and image files. Includes query, text analysis (including corpus functionalities such as word frequency and keyword in context (KWIC) tools), visualisation (including charts and network displays) and reporting tools.
NVivo https://lumivero.com/products/nvivo Free 14-day trial, from £97 p/a individual license	Mixed-methods data analysis tool for researchers which enables the importing, organisation and coding/annotation/tagging of unstructured datasets, across a range of formats including text, audio, video and image files. Includes word frequency, word cloud, and text searching functionalities, reporting tools and visualisation tools including diagrams and charts.
Qualtrics www.qualtrics.com Free trial version, £1185 p/a for a licensed version	Supports the development and analysis of online surveys. Includes statistical and sentiment analysis, reporting tools, along with visualisation tools for creating charts and graphs. Supports collaborative analysis.
Signal AI https://signal-ai.com/ Commercial tool, pricing depends on needs	A commercial tool designed specifically for supporting AI media monitoring and business intelligence. The tool uses machine learning to track and analyse news, social media etc. Includes a range of visualisation tools including word clouds, heatmaps, geospatial visualisations, amongst others. Signal AI has been used by our project partners, Amgueddfa Cymru | Museum Wales.

[1] Prices obtained on 1st April 2024.

number of speakers, with the June 2023 survey suggesting a figure of 889,700 (ONS 2023). To fulfil the obligations of the Welsh Language (Wales) Measure 2011, which promotes Welsh standards and legislates for the Welsh language to be treated no less favourably than English in Wales, surveys administered in Wales should provide individuals with the option/means to use Welsh as well as English in their responses. Due to limited staff and funding resources, effectively processing bilingual survey data and visitor feedback is a challenge faced by a wide range of businesses and institutions in Wales as there is often not sufficient time or Welsh language expertise to fully process and/or respond to Welsh-language responses effectively. The need to fully attend to Welsh language feedback data was articulated by our project partners (Section 2.1), all of whom are based in Wales and thus regularly receive responses to surveys and questionnaires in both English and Welsh. These partners revealed that prior to the project, the processing of English and Welsh language qualitative data within their own institutions typically involved either i) an *indirect* approach, i.e. paying for an external company to clean, anonymise, process, analyse and report on the results (often using one of the tools above, for example, Signal AI has previously been used by project partners Amgueddfa Cymru | Museum Wales, to tackle the English language content specifically), or ii) a more *direct* manual and labour intensive approach including an individual or group of individuals reading through and manually highlighting/encoding data in order to make sense of the results seen (this was the case for Amgueddfa Cymru | Museum Wales when tackling Welsh language data). The former of these is often expensive, and restricted to English language content, and the latter is time-consuming, convoluted, and prone to human error. For small, underfunded institutions such as Cadw (the Welsh Government historic environment service), however, the latter option is often the only viable solution to this challenge (this is likely to be true for many of the 170 other Welsh Government supported organisations in Wales).

As corpus linguists and NLP experts, the authors of this paper saw the potential for integrating corpus methods and tools into the project partners data analysis practices. Whilst corpus methods and tools *have* previously been used for the purpose of free-text analysis in studies of, for example, the opinions of UK veterinarians (Huntley et al., 2018), the reception of public health messages (McClaughlin et al., 2022) and patient experience comments (Maramba et al., 2015), no widely agreed or documented pipeline for approaching such analyses has been agreed on and/or published. Like the CAQDAS tools in Table 1 (some of which contain corpus-based functionalities), corpus software used to analyse free-text responses is, again, typically targeted at expert users who require training before use, as such software is not necessarily sufficiently intuitive to enable non-expert users to freely engage with them. Such corpus tools have also not been designed with the specific purpose of analysing questionnaire/survey data in mind.

Inspired by these user needs, and gaps in availability, the FreeTxt project (which ran for twelve months from March 2022) was established. The project team aimed to work iteratively in collaboration with the project partners (Amgueddfa Cymru | Museum Wales, National Trust Wales, Cadw, Learn Welsh and the Welsh Joint Education Committee (WJEC)) to create a bilingual free-text analysis and visualisation tool that: i) integrates a number of tools and techniques from a range of projects in corpus linguistics and NLP, ii) is freely available, intuitive and user-friendly, iii) works with Welsh and English language data and iv) has the flexibility to be utilised by a range of different end-users and stakeholders.

To enhance the range of potential users of the proposed tool, it was deemed essential for the tool to contain generic analytical features that enable it to be used by any public and/or professional company and institution dealing with varying qualitative datasets and to have relevance to academic researchers analysing and visualising survey data in English and/or Welsh. Given the richness of insight that free-text comments provide, the utility of the FreeTxt tool therefore seeks to provide an immediate improvement to, for example, the processing of member surveys and on-site visitor feedback, giving users the scope to analyse extensive and detailed responses from the surveyed population, whilst more fairly and consistently meeting the needs of Welsh-language responses.

2. Methodology

2.1. User-driven design

To articulate, then overcome, the real-world challenges faced in the analysis of free-text data, the project team adopted a user-driven approach to the research (i.e. one in which practitioners and end-users co-construct the research design from the start) to ensure that it has 'relevance and application to real-world problems and uses beyond the academic context' (Knight et al., 2021: 44). The end-users of the project represent different professional domains, from those working in the cultural and heritage sector (e.g. Cadw), to education (e.g. WJEC), but they all shared the common issue: tackling qualitative feedback data from surveys/questionnaires (although, of course, the nature, scope and frequency of obtaining this data varied across the different partners). Partners in the cultural and heritage sector, for example, wanted to better understand what people come away with (when they have visited one of their sites) and how they can measure that against what they are trying to achieve in their strategy. Furthermore, the Amgueddfa Cymru | Museum Wales strategy notes that 'We're committed to listening and collaborating with staff and volunteers, people, partners and communities to make Amgueddfa Cymru relevant and welcoming to everyone' (Amgueddfa Cymru – Museum Wales 2023: 2) and, more specifically, aim to 'make sure that everyone is represented'. Exploring bilingual visitor feedback provides a small contribution to meeting this aim. For this partner, then, having the means to obtain broad insights into visitor impressions, as well as specific (case-by-case) reasons for these (i.e. the *why*), in Welsh and English, was seen to be particularly advantageous.

The user-driven design included two main phases. To maximise the relevance of the tool, we aimed to complement what the project partners were already doing themselves, that is, to speed up the analysis rather than cause any additional workload/confusion/convolution. Phase one, therefore, gained a baseline understanding of external partners individual practices as well as needs. Project partners' engaged with initial scoping meetings (comprising unstructured interviews and walk-throughs of current practices), completed open choice questionnaires and participated in online workshops to help achieve this. These scoping initiatives provided the project team with details of, for example: the forms of feedback data received by the institution (including the quantity, frequency and format), file formats they typically deal with, how they might want to subdivide or filter the data they receive, what functionalities, ideally, they would like to see in the FreeTxt toolkit (i.e. how they would like to interrogate the data), as well as what the tool should look like (i.e. the basic requirements of the user interface – existing tools and digital libraries of tools/visualisations were also shared to get a clearer idea of what is possible and well as preferred). Table 2 provides some examples of the verbatim responses from this consultation phase (column one).

Feedback gained from the partners was collated and aligned to provide the foundations for a combined vision for how the tool should operate and what basic functionalities it should include (i.e. offering a 'problem – solution' approach to development). Specific solutions to the needs/problems identified by the partners were proposed by the project team (column two) which fed into the development of provisional plans for the tool's core functionalities. Again, this user needs analysis was an iterative process insofar as, as beta versions of each individual functionality were developed, they were demonstrated to partners and constant checking and feedback provision was sought, which fed into future amendments and developments of the tool.

Project partners also shared details of any digital tools they already use and how they typically process data using these tools. Cadw, for

Table 2
Partner feedback and proposed functionalities for FreeTxt.

Feedback	Proposed functionality
We normally get a *data dump in a database file, like an Excel spreadsheet*	Flexible data input formats, supporting .xls
We would like a tool that allows us to quickly understand *what people are talking about*	Text visualisation; frequency analysis; collocation
It would be nice to have a tool that helps us go *through the positives as well as the negatives. We normally only focus on the negatives, even though last time we looked at Trip Advisor only 8 % were ranked at 1, 2, or 3*!*	Sentiment analysis
We not are not sure how to ask questions of data – ideally, we would like a tool that enables us *to be guided through it* [the data]	Intuitive analysis and visualisation
It would also be useful to be able to click on a node [word] *and get a sense of frequency (an idea of what might be important to the audience) – or to be taken directly to the comment(s) to inspect in more detail.*	Frequency analysis; key word in context (KWIC)
Needs to have capabilities to explore: Welsh and Welsh, Welsh and English, English and Welsh text	Fully bilingual; language recogniser

example, manually extract data from Trip Advisor and use (an unnamed) digital software to create word clouds and word lists as a mean of identifying the most common words used. These terms are then manually grouped into overarching topic-based categories. Columns one and two in Table 3, provide examples of 'trigger' words that Cadw search for/identify when categorising feedback ('key term', in column two) pertaining specifically to the broad topics of '*Arlwyo'* ('catering') and '*Dysgu Gydol Oes'* ('lifelong learning/education') (column one, 'category').

Learning about the use of this approach, for example, provided a justification for integrating a semantic tagging facility to be integrated into FreeTxt: to enable current practice to be supported in an automised way. The team therefore included the USAS semantic tagger (Rayson et al., 2004) - a system for automatically ascribing semantic tags to text, which forms part of the online corpus analysis toolkit Wmatrix (Rayson, 2002) into the plans for the tool. Notably, a Welsh-language version of the USAS semantic annotation system was developed as part of

Table 3
Thematic groupings - an example.

Category (manually defined)	Key term (manually defined)
Arlwyo (Catering)	*bwyty* (café) *lluniaeth* (refreshments) *diod* (drink), *diod boeth* 'hot drink' *cinio* (lunch) *bwyty* (restaurant) *peiriant coffi* (coffee machine) *bwyd* (food) *diodydd* (drinks) *bwyta* (eat) *coffi* (coffee) *siop goffi* (coffee shop) *ystafell de* (tearoom) *byrbrydau* (snacks) *diod boeth* (hot drink)
Dysgu Gydol Oes (Lifelong Learning) / *Addysg* (Education)	*addysg* (education) *ysgol* (school) *addysg gartref* (home education) *actifeddau* (activities) *addysgol* (educational) *addysgwr* (educator) *taith* (trip) *dysgu* (learn) *adnoddau* (resources) *dysg* (learning)

CorCenCC (National Corpus of Contemporary Welsh, www.corcencc.org), which allows for the fully bilingual (Welsh and English) provision of this functionality.

As shared decisions for 'core functionalities' (see Section 2.2) of FreeTxt were established, demonstrations of current corpus-based and NLP tools that offer potential solutions were shared with the project partners to i) ensure there was a clear and shared understanding of the problem/need and ii) to offer an insight into how we might address it. This step was to mitigate the risk of including a functionality/tool that is not fit-for-purpose. For instance, we took the key terms identified in column two of Table 3 and illustrated to the partners how the USAS tags and categories would tackle language of this kind. The results of the automated USAS tagging process are presented in columns three and four of Table 4.

Whilst the manually assigned and USAS category labels are not *identical*, the mapping between them is clear and most of the terms were successfully mapped into semantically related groups automatically assigned by the USAS tagger. The partners were pleased with these results and satisfied that this automated approach would provide them with a more systematic (and much quicker) way-in to grouping feedback responses than is currently afforded when processing data manually.

2.2. Core functionalities

Feedback from the phase one planning meetings was combined with insights from scoping current CAQDAS tools (as seen in Table 1), and functionalities from NLP and corpus linguistics to identify some of the core functionalities that FreeTxt would provisionally include. Whilst there exist some universal QDA (Qualitative Data Analysis) functions in currently available tools, these were to be augmented with additional language-specific functionalities within FreeTxt. Technical functions include, for example, multi-language facilitation (i.e. with a standard approach to coding, enabling the tool to be adapted for use in other languages), flexibility in visualising results (i.e. with graphs, charts, word clouds etc.) and pattern and relationship analysis (i.e. using corpus query tools). The proposed integration of these features in FreeTxt aimed to provide users with the flexibility of working between different interface levels within the tool whilst analysing with bilingual data in a systematic and efficient fashion that is not currently possible.

There was an aim for FreeTxt to be an integrated, web-based platform that meets the following key technical requirements:

- **Installation**: run on most standard web-browsers (Edge, Chrome, Firefox, Safari etc.). It should not require any installation on the user's machine, although should offer a secure and reliable way for users to upload and explore their data, and for the data to be fully deleted from the system at the end of a session.
- **Input specification:** include an easy-to-use and intuitive input option. For example, texts (or surveys) could be entered directly into the text area provided on screen, on a one-per-line basis, or uploaded from file (namely .txt and .xls formats).
- **Output specification:** display on-screen but with the option of being exported or downloaded and saved in appropriate formats for future use.
- **Features**: potentially perform the following functions (based on existing corpus linguistic and NLP tools, such as those offered by AntConc (Anthony, 2023), Sketch Engine (Kilgarriff et al., 2014) and Wmatrix):
  - **Word and n-gram frequency:** enable a basic count of words, n-grams, POS tags and semantic tags in a dataset.
  - **Key word in context (KWIC)**: enable lexical searches of the text to generate a list of all instances of a search term and their immediate co-text in the dataset.
  - **Text visualisation:** present a variety of pictorial representations of word distributions in a dataset. To be implemented using a

Table 4
Manually assigned semantic tags Vs automatically assigned semantic tags.

Category (manually defined)	Key term (manually defined)	USAS tag	USAS category (first listed tag only)
Arlwyo (Catering)	*bwyty* (café)	F1/H1	*bwyd* (food) / *pensaernïaeth a mathau o dai ac adeiladau* (architecture, housing and buildings)
	lluniaeth (refreshments)	F2	*diodydd* (drinks)
	diod (drink), *diod boeth* 'hot drink'	F2	*diodydd* (drinks)
	cinio (lunch)	F1	*bwyd* (food)
	bwyty (restaurant)	F1/H1	*bwyd* (food) / *pensaernïaeth a mathau o dai ac adeiladau* (architecture, housing and buildings)
	peiriant coffi (coffee machine)	F2/O3	*diodydd* (drinks) / *trydan ac offer trydanol* (electricity and electrical equipment)
	bwyd (food)	F1	*bwyd* (food)
	diodydd (drinks)	F2	*diodydd* (drinks)
	bwyta (eat)	F1/B1	*bwyd* (food) / *Iechyd a chlefyd* (health and disease)
	coffi (coffee)	F2	*diodydd* (drinks)
	siop goffi (coffee shop)	F2/H1c	*diodydd* (drinks) / *pensaernïaeth a mathau o dai ac adeiladau* (architecture, housing and buildings)
	ystafell de (tearoom)	F2/H2	*diodydd* (drinks) / *rhannau o adeiladau* (parts of buildings)
	byrbrydau (snacks)	F1	*bwyd* (food)
	diod boeth (hot drink)	O4.6+ | F2	*tymheredd* (temperature) | *diodydd* (drinks)
Dysgu Gydol Oes (Lifelong Learning) / *Addysg* (Education)	*addysg* (education)	P1	*addysg yn gyffredinol* (education in general)
	ysgol (school)	M1/P1	*symud, dod a mynd* (moving, coming and going) / *addysg yn gyffredinol* (education in general)
	addysg gartref (home education)	H4/H1c | P1	*preswylio* (residence) | *addysg yn gyffredinol* (education in general)
	actifeddau (activities)	A1.1.1	*gweithredu cyffredinol, gwneud ac ati* (general action)
	addysgol (educational)	P1	*addysg yn gyffredinol* (education in general)
	addysgwr (educator)	P1/ S2mf	*addysg yn gyffredinol* (education in general) / *pobl:- benyw/gwryw* (people male/female)
	taith (trip)	M1/P1	*symud, dod a mynd* (moving, coming and going) / *addysg yn gyffredinol* (education in general)
	dysgu (learn)	X2.3+	*dysgu* (learn)
	adnoddau (resources)	A9+	*derbyn a rhoi; meddiant* (getting and giving; possession)
	dysg (learning)	X2.3+	*dysgu* (learn)

combination of available Python text and data visualisation libraries like Wordcloud, ScatterText and Matplotlib.

- **Part-of-speech (POS) tagging**: implement available English and Welsh POS taggers including the CorCenCC's rule-based POS tagging and tokeniser software, CyTag. In early evaluations of the tagger, CyTag achieved accuracy levels of over 95 % (see Neale et al., 2018).
- **Semantic tagging:** implement the English USAS tagger and the CorCenCC Welsh semantic tagger, CySemTag (Piao et al., 2018) to semantically tag data.
- **Sentiment analysis:** implement an existing English language sentiment analyser in addition to a basic sentiment analysis tool based on the cross-lingual sentiment analysis resources and data made available via a Welsh Government-funded project on bilingual word embeddings (Espinosa-Anke et al., 2021).
- **Text summariser:** apply an existing summarisation tools for English and Welsh (Ezeani et al., 2022) to generate more concise versions of the survey texts as a collection.
- **Multi-lingual support:** support the analysis of bilingual (Welsh and English) language data, aided by the inclusion of a language identification tool will be included to identify the language of each survey text.

As signposted in this list (and referenced in the main project objectives detailed in Section 2), the FreeTxt tool aimed to reuse some Welsh language resources which were built by members of the project team including, for example, the semantic and POS taggers created as part of the ESRC/AHRC-funded CorCenCC project and Welsh-Government funded Welsh Automatic Text Summarisation tool (see www.digigrid.cymru/analyse), as well as existing English language tools which support these functionalities. The design of FreeTxt, therefore, aimed to build on existing tools and research, taking it in a new direction by enabling non-corpus linguists (and non-academics) to create and use corpora, and generating additional non-academic impact of corpus linguistic tools and research.

2.3. User testing and feedback

The second key phase in the development of FreeTxt was to use the user-defined requirements and core functionalities to create a prototype version of individual features of the tool and to demonstrate it to the end-users to obtain constructive feedback. This included guided demonstrations and updates provided by members of the project team, as well as the use of walk-through worksheets that partners would test on their own and respond to. Walk-through prompts encouraged users to either use a pre-loaded set of data, or to upload their own to see whether they could navigate through the tools with ease. We asked users to comment specifically on what worked well and what didn't work so well, urging them to be as honest and as critical as possible, and to reflect on what specific changes should be made to increase the usefulness of the tool (with a specific consideration of the needs of their own institution).

Input was also sought on the terminology used to label each individual function within the tool, and how the results were presented/described. 'Collocation', for example, did not have an obvious meaning to users, so to increase the intuitiveness (and usability) of the tool, 'relationships' was a more immediately meaningful term to describe this concept. Examples of other preferred terms include 'word use' instead of concordance (or KWIC) and 'meaning analysis' instead of sentiment (or affect/ positive and negative reviews) analysis. In fact, any reference to technical terms from the broad field of corpus linguistics (including 'corpus') required renaming to make the concepts and analytical tools more accessible and meaningful to potential end users. Additions to the core lexicon of the USAS tagger were also made, in both the Welsh and English language versions, to ensure that it effectively captured some of the domain specific terminology associated with partners' sites. These additions included some of the specific locations of the museums/sites relevant to the project partners (e.g. *Amgueddfa Wlân Cymru* (National Wool Museum) – a site belonging to Amgueddfa Cymru | Museum Wales).

Feedback on the specific visual presentation of prototype tools was also garnered. An example of an early screenshot is provided in Fig. 1.

Fig. 1 provides a proposed collocation tool, presented in the

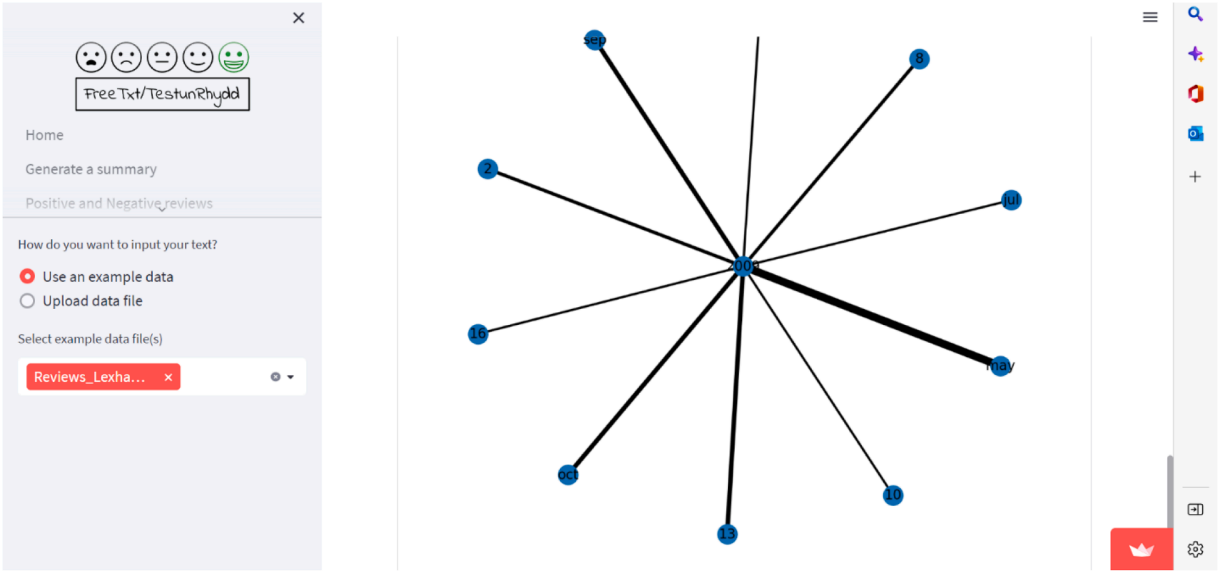


Fig. 1. Screenshot of a prototype version of FreeTxt, created in Streamlit.

prototype environment of FreeTxt. This initial version of FreeTxt utilised Streamlit, an open-source Python library used for building web-based applications. Here, the search term is positioned in the centre of the visualisation (note that this is very unclear due to its positioning in the blue node), with each spoke presenting the collocates of the term. The thickness of the line indicates the frequency of co-occurrence of the search term. Feedback obtained for this particular visualisation led to a number of problems being identified. This included the specific problem related to its layout (i.e. the difficulty with reading the text). Also, it was not immediately clear to users what the sizing and positioning of the spokes represented. Feedback of this nature was taken on board for the next iteration of development.

3. FreeTxt: functionality pipeline

In the third phase of development, the full FreeTxt toolkit was created (in Flask, a Python-based microframework), drawing on all needs and feedback gained from the project partners. A system diagram for the final version of FreeTxt is provided in Fig. 2 and the main functionalities of the toolkit are outlined in the sections below. Note that a full list of the specific digital libraries that FreeTxt draws on can be accessed via the 'About' tab on the tools' website: www.freetxt.app – the descriptions provided below are targeted at a more general (non-technical) audience. A full, plain English and plain Welsh user guide is also provided on the website, in addition to security information (namely that data is temporarily uploaded onto a secure server whilst FreeTxt is being used and is immediately deleted once the analysis session ends).

On the left side of Fig. 2 is the input into the system, which is the text-based sentences/reviews. Once uploaded, a series of analyses are carried out including sentiment analysis (Sections 3.2 and 3.3), which draws on the BERT (a state-of-the-art NLP model) sentiment analyser from Hugging Face (see https://huggingface.co/nlptown). This model uses wider contextual information when undertaking sentiment analysis, rather than tagging at the word level alone. The model is trained on product reviews in multiple languages (English, Dutch, German, French, Spanish and Italian). As per the information on the Hugging Face model page, the accuracy of this model for sentiment analysis on English text is approximately 95 %. We also undertook experiments on manually annotated Welsh reviews and determined that the accuracy for Welsh was approximately 73 %, thus had scope for future improvements. The text is also automatically POS and semantically tagged using CyTag, spaCy, CySemTag and the USAS pymusas-models (see Sections 2.1-2.2). The

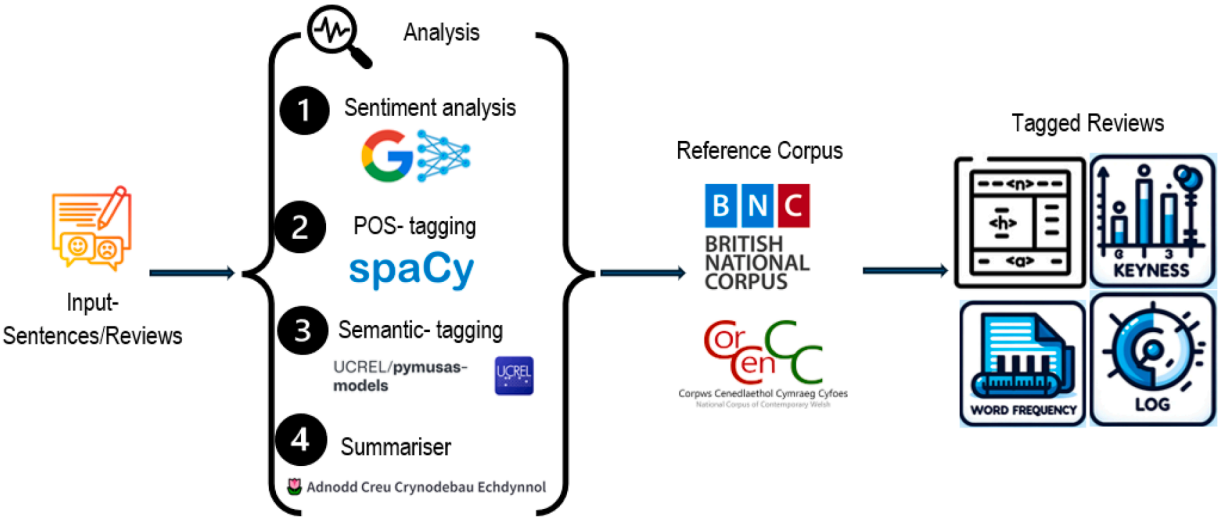


Fig. 2. FreeTxt system chart.

POS and semantically tagged texts are compared with wordlists from the original BNC (1994) and CorCenCC to enable keyword analyses, and the tagged text is analysed using a range of computational methods, including Keyness and frequency analysis, log-likelihood analyses, and semantic and sentiment classification. Each of these is discussed in more detail in Sections 3.1-3.7.

FreeTxt is open source and can be adapted or extended to support additional languages by integrating language-specific resources and models into the freetxt.app website tools (see the FreeTxt GitHub page for more information: https://github.com/UCREL/FreeTxt-Flask). The key steps involve utilising the USAS pymusas-models for semantic analysis in the target language, integrating an appropriate part-of-speech (POS) tagger from available NLP resources, and fine-tuning the adopted sentiment analyser BERT model on datasets specific to the target language. This process includes identifying and incorporating linguistic resources, modifying text pre-processing steps, training, or fine-tuning models on relevant datasets, and adjusting analysis workflows to integrate the new tools and models based on the added language. Additionally, the user interface should be fully translated to accommodate the new language, ensuring a smooth and intuitive experience for users. Certain tools, such as the summariser, do not require language-specific adjustments and can be utilised directly, simplifying the adaptation process. The visualisation aspects of FreeTxt are language independent other than labelling changes as part of the interface customisation

Below includes a detailed walk-through of FreeTxt v1.0.0, using the first 130 entries of an open-source dataset taken from Yelp, an online user reviews and recommendations website for restaurants, shopping, food, entertainment and so on. This dataset includes review IDs, site IDs, date of posting and other numerically based feedback scores (including how many stars) and is available on the DAT7 Course Repository page on GitHub (https://github.com/justmarkham/DAT7). This dataset is one of only a few widely available feedback datasets available for others to use, which is why it has been selected here. The walk-through includes both English and Welsh language examples, to illustrate the bilingual functionalities offered by the toolkit. As there are no freely available datasets of this nature in Welsh, the Welsh language examples here are derived from automatically generated translations of the English-based Yelp dataset (using ChatGTP), which were subsequently checked by a fluent Welsh speaking member of the project team. Whilst it would have been preferable to use data from the project partners to demonstrate the tool, unfortunately confidentiality precludes this.

3.1. Data upload

To increase the usability of FreeTxt, users are not required to login to the website – by clicking the 'analyse' tab they are taken directly to the file upload page where they have the option to paste in a short text (this was a functionality requested specifically by the project partners) or upload a more extensive file in a .txt, .tsv or xls format (up to 400,000 words).

Feedback from early adopters of FreeTxt has commented that, beyond the actual analysis of data, the tool may prove advantageous in their future questionnaire design (i.e. that they can favour formats that they know are likely to be well-analysed by the tool), whilst the specific design of the questionnaire can also inform the 'best' way to navigate the tool during analysis. For example, if a question asks a respondent to provide three words/terms to describe their feelings about an experience, exhibition and so on, the utility of a frequency-based word cloud (see Section 3.4) is likely to be more informative/useful when interpreting the data than, for example, a Keyness based word cloud (which, in the FreeTxt environment, uses the 100-million-word BNC or 11-million-word CorCenCC as its reference).

Typically, corpora are not edited or modified in any way as there is an emphasis on data being as authentic as possible. Any form of editing and/or modification of data is effectively seen to 'distort' the language and is at odds with the main advantage of using a corpus approach: exploring language as it is *actually* used. Exceptions to this rule do, however, exist. Data cleaning is in fact common in many studies of questionnaire/survey analysis, so there is precedence for it in this context. For example, in Ferrario and Stantcheva's study of public policies of income and estate taxation, they removed 'punctuation, excess spaces, numbers, misspelled words, and so-called "stop words," which are common words that carry no intrinsic meaning such as "and" or "the."' (Ferrario and Stantcheva, 2022: 3). As our project partners are mainly concerned with the general patterns of meaning developed from what is being said (including how often a specific word or sentiment is mentioned), rather than how it is being said, they opted to undertake some basic cleaning of data before using FreeTxt, to maximise the effectiveness of the toolkit. Such data cleaning included the standardisation of spelling and removal of bullet points, numbered lists and encoding characters (e.g. â€¢), for example.

Before the uploaded data is visualised in FreeTxt, specific columns can be selected and or filtered (e.g. to focus on a specific time period, exhibition or museum site) and selected for subsequent analysis. It is at this stage that a language recogniser can be run to determine (and allow the separation of) Welsh and English text to ensure that Welsh language data is processed using the Welsh language tools, and English language data with English language tools within the FreeTxt environment. If Welsh language data is detected (by clicking 'Check Language'), the tool automatically creates separate files of English and Welsh content which can be uploaded back into the system for individual analysis. Once selected for analysis, text is automatically tagged using POS and semantic taggers (see Section 2.2).

3.2. Meaning analysis

The first analytical and visualisation tool in FreeTxt is meaning analysis (i.e. sentiment analysis), which enables users to determine how respondents 'feel' about what they are commenting on, based on either three-class (i.e. positive, negative, neutral) or five-class classifications (i.e. as with the three-class sentiment, with the addition of very positive and very negative). Figs. 3 and 4 present screenshots of the English Yelp sample data in this functionality in FreeTxt. The results are presented in an interactive pie chart (left, Fig. 3), interactive bar chart (right, Fig. 3) and a sentiment table (Fig. 4). Each visualisation (here and as described in the sections below) has integrated options to enable the users to download or screenshot the results for future offline use. Users can also, for example, click on specific 'slices' of the pie chart or columns in a bar chart to deselect them, if they, for example, only want to focus on the results for the very positively or positively-only coded results.

In Fig. 3 we see that 40 % of the English Yelp sample data reviews were classed as 'very positive'. The overall sentiment score is indicated to be 22, which means that there are 22 more positive than negative statements in the given text, suggesting that the overall sentiment is positive. The sentiment table in Fig. 4 presents a sample of individual reviews, their sentiment labels and the confidence scores (i.e. the confidence with which this label has been accurately ascribed), sorted by confidence score in this case. Specific words can be searched within this output, and the full table can be downloaded for future use (in an .xls format).

By reading the reviews presented in Fig. 4, we can see that the sentimental labels ascribed provide an accurate reflection of the basic 'feeling/affect' derived by reading the text (e.g. *That darn Smokehouse burger is awesome. 5 stars.* would certainly be conceived as being a 'high positive review').

To increase their confidence in utilising the FreeTxt tool, project partners from Amgueddfa Cymru | Museum Wales undertook their own evaluation of this particular part of the toolkit. Specifically, they compared results from prior sentiment analyses (carried out manually by an external consultancy company) of 774 feedback comments written in English, comparing the results received with results generated by

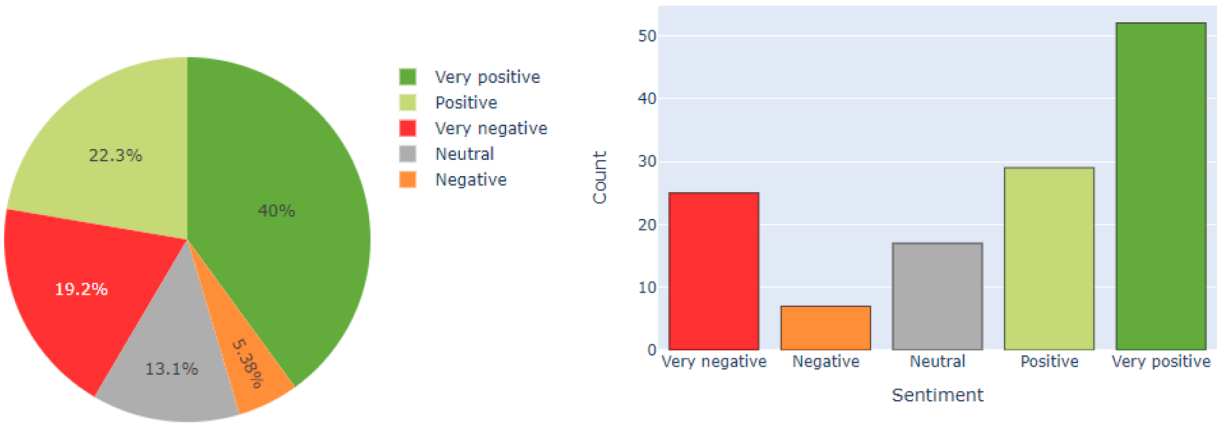


Fig. 3. Meaning analysis graph and chart visualisations, based on 5 class sentiment (created using the Yelp sample data).

Review	Sentiment Label	Confidence Score
That darn Smokehouse burger is awesome. 5 stars.	Very positive	0.98
Best food, super friendly staff, and great prices. Love it!	Very positive	0.92
U can go there n check the car out. If u wanna buy 1 there? That's wrong move! If u even want a car service from there? U made a biggest mistake of ur life!! I had 1 time asked my girlfriend to take my car there for an oil service, guess what? They ripped my girlfriend off by lying how bad my car is now. If without fixing the problem. Might bring some serious accident. Then she did what they said. 4 brand new tires, timing belt, 4 new brake pads. U know why's the worst? All of those above I had just changed 2 months before!!! What a trashy dealer is that? People, better off go somewhere!	Very negative	0.9
Great food and awesome service! Even better that the Chef came out and personally checked on our experience. I will be back for more delicious BBQ	Very positive	0.89
Christy is an amazing cake artist. She has an impressive portfolio as she has a flair for creativity. Her cakes are amazing and are truly one-of-a-kind. She also has several delicious cake flavors to choose from or she can create one for you. She has made several cakes for our family and I have been impressed with everyone of them. I was also excited to see her compete on TLC's Ultimate Cake Off show early this spring. If you're looking for a sculpted cake or one that will be the center of attention at your next party, then I recommend you call Phoenix Cake Company.	Very positive	0.85
Another night meeting friends here. I have to laugh. Waited another 20 minutes for my beer to be refilled at the bar. A girl even took my empty without even asking if I wanted a refill. A new brunette girl that I don't recognize left the bar and sat down with her guy friends on the customer side AT 9:25 ON A FRIDAY NIGHT. Another bartender had to ask her to come back and work. Management.... Pull your head out of your ass! Sad to watch.... I need to talk my friends into another place!	Very negative	0.83
Awesome subs clean and friendly well priced.	Very positive	0.83

Fig. 4. Sentiment table, based on 5 class sentiment (created using the Yelp sample data).

FreeTxt. The manual classification indicated that 429 (67.5 %) comments were positive, 190 negative (15.5 %) and 126 neutral (17 %). The results derived from FreeTxt provided only a small variance to this, with a difference in 1.3%, 6.2 % and 4.9 % for comments coded as positive, negative and neutral respectively (Fig. 5, chart produced in Excel). The partners were happy with this result and felt it underlined the potential cost-benefit of using FreeTxt, insofar as it saved time and money when undertaking the analysis.

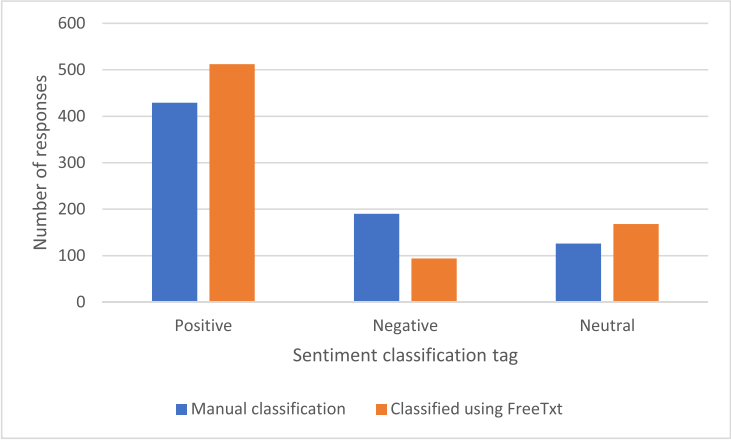


Fig. 5. Manual Vs. automated sentiment classification using FreeTxt.

3.3. Sentiment chart

The second stage of the analytical pipeline, as seen in Fig. 6, is the meaning chart functionality.

Fig. 6 presents a scatter plot, displaying the single words with the highest sentiment association. The x- and y-axes show the usage of these words in positive vs. negative and neutral sentiments, respectively. The 'top positive' and 'top negative/neutral' words are listed in columns to the right of the plot (in this case *awesome* and *amazing*, respectively). Hovering over a word in these lists highlights where it is positioned in the scatter plot on the left. Hovering over a specific word in the scatter plot itself provides information of its frequency in the dataset, along with its sentiment score (ranging from −1 to 1, e.g. *amazing* in Fig. 6). Clicking on an individual word within the scatter plot reveals the sentences in which the specific word is used and its frequency of positive, negative or neutral sentiment uses. For example, clicking the word *awesome* in the sample Yelp data, you see instances of *awesome* as used in sentences that have been analysed as being positive, as seen in Fig. 7. It is also possible to search for specific words in this functionality to view their frequency of use in sentences of different sentiments.

3.4. Summarisation

The summarisation tool enables users to create extractive summaries of text data. This functionality operates most effectively on single texts, such as texts pasted in at the initial input phase, although it can provide a general summary of all feedback. The tool allows users to select their summary level, from 50 % to 10 % of the original wordcount/size, creating short versions of complex information to better understand the general gist of information. Project partners found this functionality useful as a means of obtaining reliable and representative examples of the feedback, although, granted, other parts of the toolkit also provide the means for this. The project partners also tested this functionality with other forms of documents, including their own policy and other institutional documents, indicating the potential usefulness to them of using this part of the tool for these forms of data, rather than simply free-text responses. Whilst this particular use of the summariser did not align to the original aims for the toolkit, the partners were keen for the functionality to remain in the final toolkit as they could see a potential use for it in their work.

3.5. Word cloud

Word cloud functionalities are commonly used by institutional communications teams, including by our project partners, because they provide a quick, visual indication of some of the most pertinent topics/issues that arise in text. The FreeTxt word cloud function provides frequency-based, keyness (using the BNC and CorCenCC as reference corpora) and log-likelihood based clouds. Users have the option to focus specifically on individual content words (the utility automatically removes function words because these were seen as 'uninteresting' to the partners), clusters (from 2–4 grams), individual parts-of-speech tags (e.g. to enable users to explore the most common adjectives used to discuss a particular topic) and semantic tags (left, Fig. 8– the most frequent tag here is *Lleoliad a chyfeiriad* (location and addresses)). There is also the option to personalise the word cloud according to its shape (from a pre-defined list – see the right image in Fig. 8– Sherlock Holmes in this case – options for visualising via the Cadw, Learn Welsh and National Trust Wales logos are also available, as per the request of the project partners), and whether it has an outline (and what colour), as well as to deselect and regenerate the cloud if it includes a word that is perhaps redundant (e.g. for Cadw, it is likely that respondents will mention the word *castell* (castle) in their feedback, since many of their sites are castles: the use of this word is, therefore, not necessarily interesting, so is removed in their word clouds) or sensitive/controversial.

3.6. Word use and relationships

The next stage of the pipeline allows users to drill down into the information seen into the word cloud in more detail using frequency counts, KWIC and collocation information. Users can select either a specific word, POS tag or semantic tag and see its use in context. For example, Fig. 9 shows one of the most frequent words in the English language dataset, *good* (with a raw frequency of 39) in context. As with conventional corpus tools, the window size can be adjusted here, and

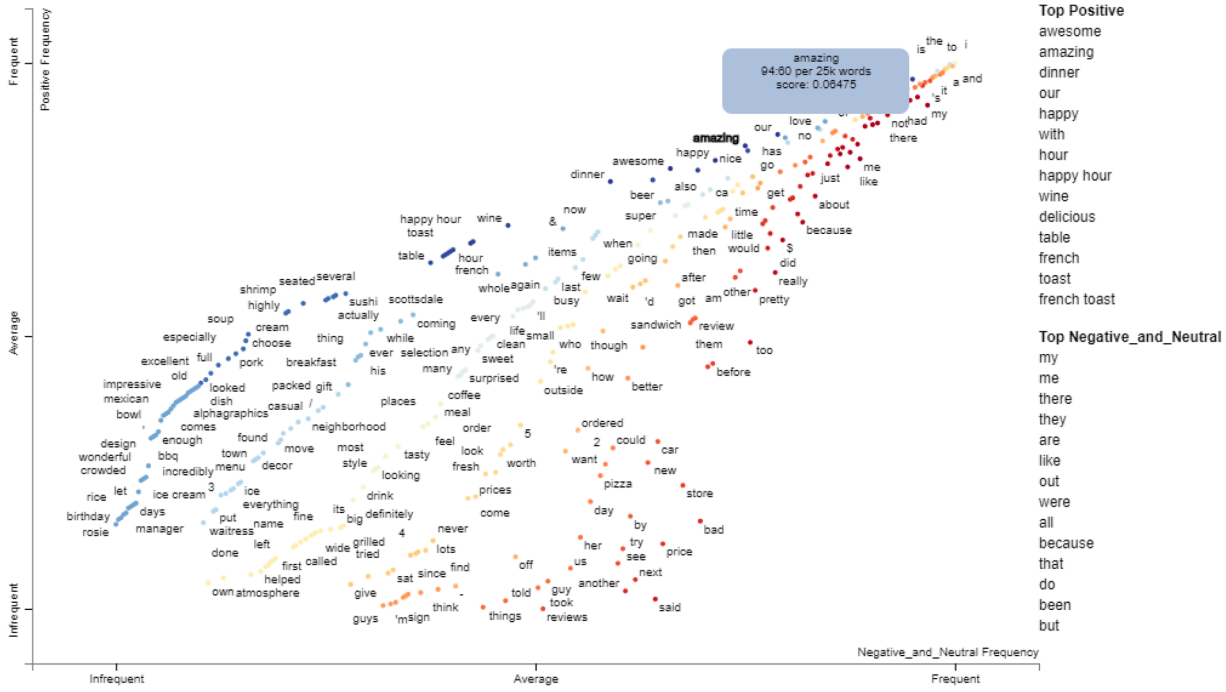


Fig. 6. Sentiment chart visualisation in FreeTxt (created using the Yelp sample data).

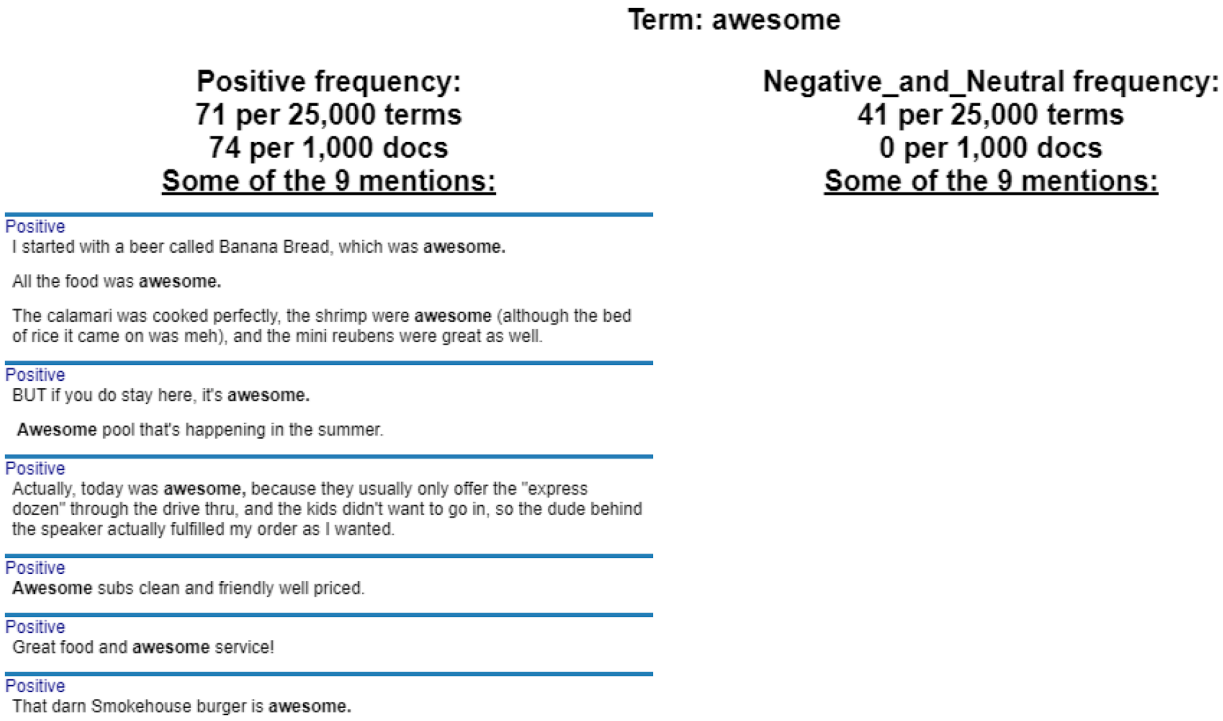


Fig. 7. Sentiment scores of sentences including *awesome* in the Yelp sample data.

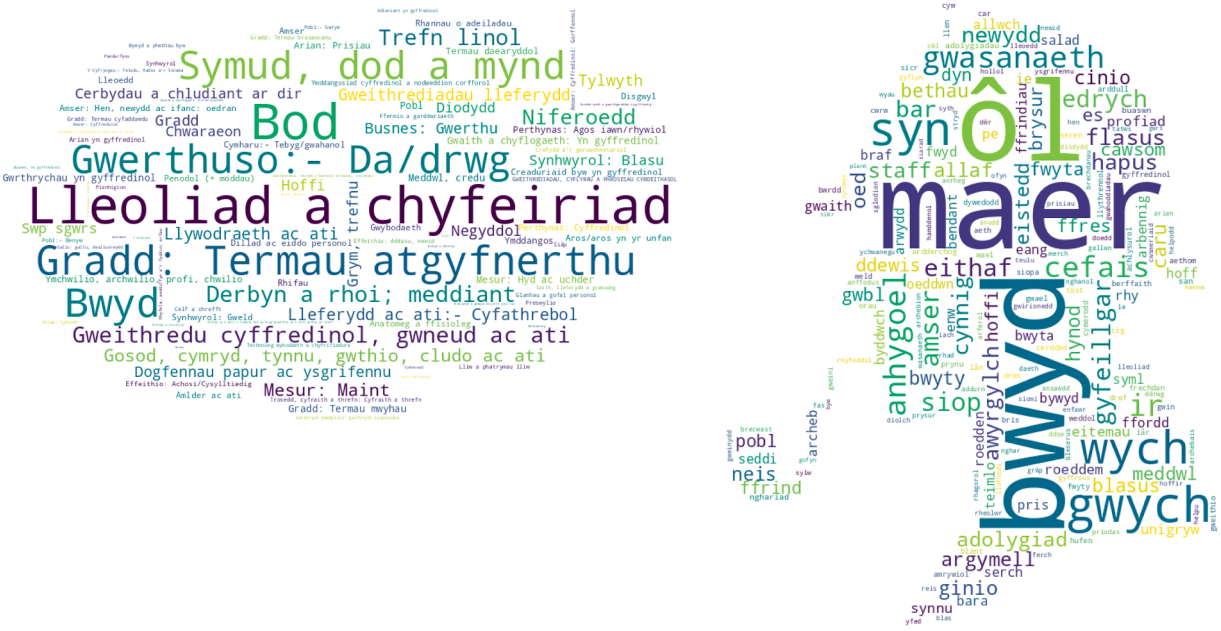


Fig. 8. Frequency-based word and semantic clouds, generated using the translated sample Yelp data.

individual words can be searched for within the concordance outputs. The output can also be sorted alphabetically according to its left or right context.

Beneath this visualisation, users are presented with two additional visualisations in sequence. The first visualisation, as seen in Fig. 10, depicts the frequency of most common collocates of the search term (*good*), along with its mutual information (MI) and log-likelihood scores (the MI score is a statistical measure that shows the strength of association between words; LL is a probability statistic that compares the frequency of co-occurrence of two words).

Next, an interactive network graph (similar to Gephi, https://gephi.org/ and GraphColl, which is available as part of the #LancsBox corpus toolkit, Brezina et al., 2020) is presented (Fig. 11), enabling users to view relationships between words, but also click on items (nodes) and move them around to present them, visually, in whatever way they prefer. The node in green indicates the most common collocate of the word (as also indicated by the thickness of the connecting line), the size of each node indicates the frequency of the collocate.

3.7. Word tree

The final stage of the pipeline is the word tree functionality, which is

Left Context	Keyword	Right Context
	good	tattoo shop clean space multiple artists
	good	for an easy and leisurely business
general manager scott petello is a	good	egg not to go into detail
pleasantly surprised the store had a	good	amount of parking and the store
relief and although there were a	good	amount of people the next day
is getting quality food at a	good	price the other is getting a
as well and it was a	good	sandwich although the bread was a
was super crowded which is a	good	sign on a wednesday i started
photos of vietnam service is always	good	here and ive been here when
combination once and it wasnt as	good	i love the decor here lots

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

Fig. 9. The use of *good* in context.

Word	Frequency	MI	LL
very	3	4.73	19.45
egg	2	6.58	15.61
amount	2	6.58	15.61
too	2	4.88	10.08
pretty	2	4.78	9.77
so	2	3.68	6.72
i	2	0.76	0.49
tattoo	1	7.58	10.54
informative	1	7.58	10.54
reliably	1	7.58	10.54

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

Fig. 10. Frequent collocates of *good* in the Yelp data.

based on Google charts (see: https://developers.google.com/chart/interactive/docs/gallery/wordtree). The word tree is an alternative approach to visualising a specific search term in context (with words that commonly precede and follow the search term represented accordingly). This is visualised in Fig. 12, with the frequency of specific terms indicated by its size (with frequency of use indicated in tool tip when you scroll over the individual word). By clicking on a word, the word tree reloads with the selected word featuring as the search term in the middle of the screen. The searched word here is one of the most frequent words from the Welsh language dataset (see Fig. 8), *gwych* (great).

Like WordWanderer (a navigational approach to text visualisation, see Dörk and Knight, 2015), this functionality supports a more playful approach to language interrogation, which has the potential for increasing engagement with text (something that may provide potential future uses in, for example, teaching and learning contexts).

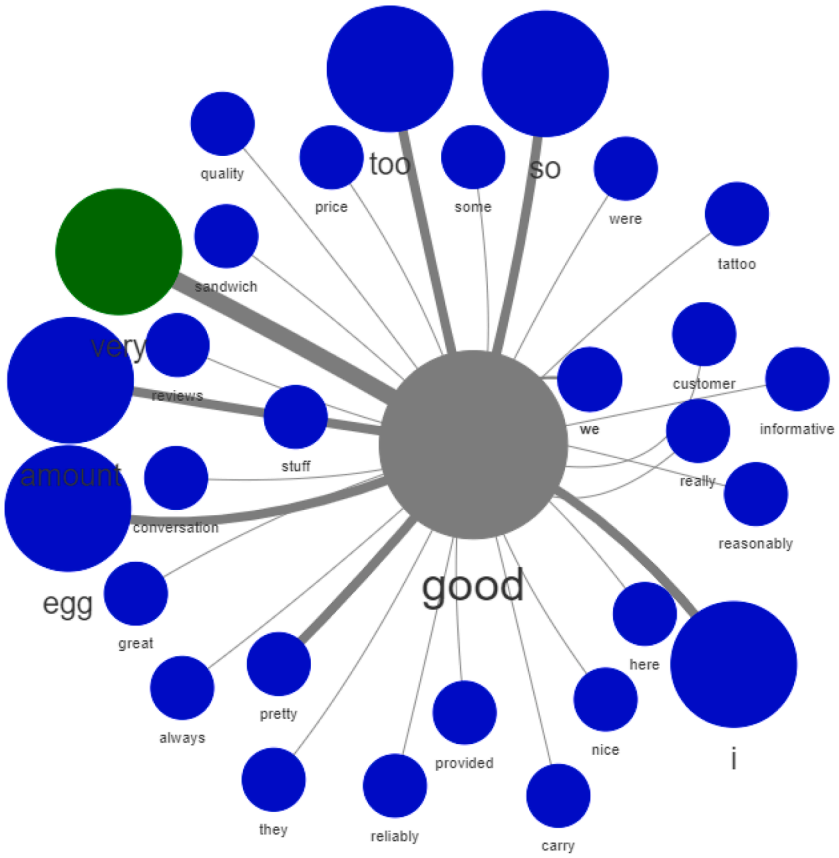


Fig. 11. Interactive network graph illustrating the collocates (and their frequency of cooccurrence) of *good* in the sample Yelp data.

3.8. PDF creator and downloads

The final stage of the FreeTxt pipeline enables the creation of PDF reports based on the analyses undertaken in the prior tabs of the tool. As with individual tool downloads and screenshots, these can be directly integrated into reports, presentations, etc.

4. Reflections and future directions

The released version of FreeTxt has been well received by the project partners, who continue to provide regular (and ongoing) insights and updates on how they are using the tool and what the overall benefits of its use are. One project partner reflected that: *The main advantage of FreeTxt for the respondents is to summarise survey feedback/reviews and identifying common themes, without the need for them to sift through individual responses themselves*. Other reflections also included: *It seems intuitive; The visuals on this make me confident; I like the little logo – positive to negative.* One partner has also estimated that, for example, it facilitates £2.5k per year in time saving alone, as well as more tangible benefits such as:

- providing consistency of interpretation, and potentially removing bias in categorisation compared to manual analysis.
- having built-in provisions for visualisation, to create more compelling impact when reporting to stakeholders.
- bringing a different, linguistically backed approach to analysis which provides a methodical way to report and increases confidence in reporting.
- enabling the facility in Welsh allows the analysis to be undertaken by non-Welsh-speakers (one partner receives c.2–4 % of responses in Welsh generally, so the ability to be able to analyse in both languages was thought to be really helpful – and unique – the partners are not aware of any other tool of this kind – commercial platforms or otherwise – that enables this functionality).

Significantly, the formal release of FreeTxt allowed all partners to fully appreciate its potential. Throughout the project, while some partners initially found it challenging to envision the tool's capabilities, they grew increasingly confident in its ability to meet their needs. Regular and clear demonstrations of 'live' functionalities were crucial in maintaining their interest and commitment, especially as discussions and demonstrations of other existing corpus tools sometimes created confusion. The partners' patience and dedication to the workplan were instrumental, especially given the project's 12-month duration. Their continued support was a testament to the project's promising outcomes, even in the face of potential technical challenges and delays.

In addition to the benefits of using the tool, the project partners also identified some shortcomings of the current version of FreeTxt, including:

- the semantic tags/classification are often not immediately clear/interpretable (i.e. a certain amount of engagement/familiarity is needed to understand the groupings and the words contained within them),
- the accuracy of the sentiment analysis for Welsh language data is only at 73 %, which is significantly lower than the accuracy for English,
- the tool is unable to accurately classify, for example, some proper nouns (e.g. the name of an artist or specific placenames, particularly in Wales), and,
- the summarisation tool proved more useful for long form documents rather than survey/feedback data as intended (see Section 3.4).

This feedback will help to inform future iterations/improvements of

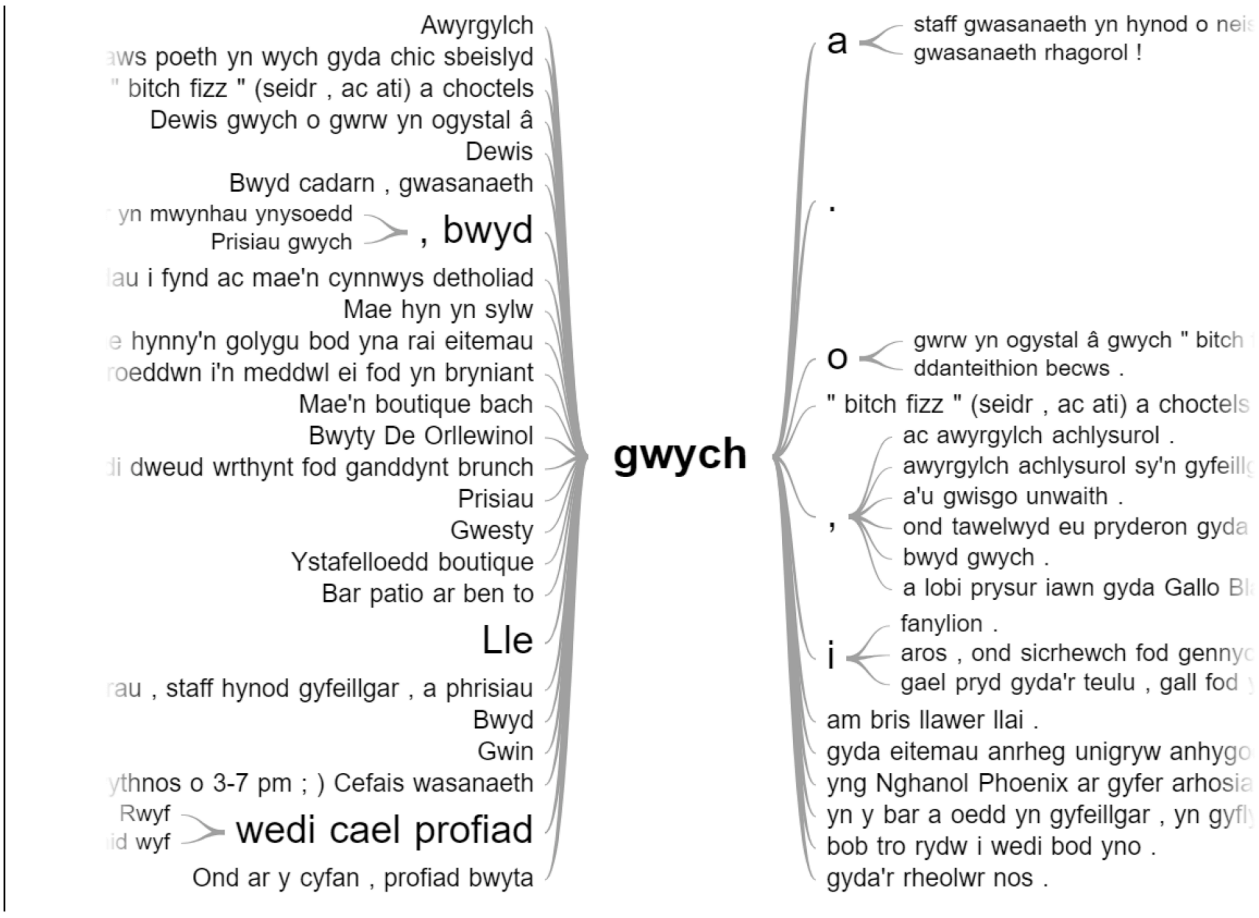


Fig. 12. G*wych* (great) and its co-text in FreeTxt's Word tree visualisation.

FreeTxt. As the available version of FreeTxt is the first release of the tool, there are also many potential areas for extending the tool beyond its current functionalities. An emerging wish-list for development is included below – is it hoped that future iterations of the tool will:

- enable direct comparisons within/between dataset, similar to keyness analysis across texts/corpora,
- map patterns over time (i.e. changes in topic, associations and priorities over time),
- extend the word cloud personalisation (e.g. to enable users to upload their own logos to be used as the word cloud template),
- include a Trip Advisor and/or social media plugin, enabling partners to directly access and analyse online feedback data relevant to their sites/organisation,
- include support for other languages (minoritised and otherwise),
- utilise the toolkit in teaching and learning contexts (for language learning and ICT), and,
- enable the integration of the toolkit into local content management systems, to embed it fully into the common practice of an organisation.

This paper has provided a comprehensive overview of the main features of the first corpus-based toolkit design to facilitate the systematic analysis and visualisation of free-text data. The paper has demonstrated how close and active collaboration between academics and end-users can help provide effective solutions to real world problems. The approach reported here, and the bilingual toolkit developed, can be replicated and extended for use in other language contexts and across a range of public and professional sectors.

CRediT authorship contribution statement

Dawn Knight: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Nouran Khallaf:** Visualization, Software, Resources, Methodology, Data curation, Conceptualization. **Paul Rayson:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Mahmoud El-Haj:** Writing – review & editing, Software, Resources, Methodology, Investigation, Conceptualization. **Ignatius Ezeani:** Visualization, Software, Resources, Methodology, Investigation, Conceptualization. **Steve Morris:** Validation, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The FreeTxt project was funded by AHRC (Arts and Humanities Research Council) follow-on funding for impact and engagement (grant number AH/W004844/1). This work also feeds more broadly into the work carried out as part of the ESRC (Economic and Social Research Council) and AHRC funded Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction project (grant number ES/M011348/1). No new data was generated as part of this paper.

References

Amgueddfa Cymru – Museum Wales, 2023. Amgueddfa Cymru – Museum Wales Strategy 2030 [Online]. Available from: https://museum.wales/about/policy/strategy-2030/.

Anthony, L., 2023. *AntConc (Version 4.2.4)* [software]. Waseda University. Available from, Tokyo, Japan. https://www.laurenceanthony.net/software.

Aston, G., 2001. Learning With Corpora. Open Library, Athelstan.

Brezina, V., Weill-Tessier, P., McEnery, A., 2020. *#LancsBox v. 5.x* [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Dörk, M., Knight, D., 2015. WordWanderer: a navigational approach to text visualisation. Corpora 10 (1), 83–94.

Espinosa-Anke, L., Palmer, G., Filimonov, M., Corcoran, P., Spasic, I., Knight, D., 2021. English–Welsh cross-lingual embeddings. Appl. Sci. 11 (14) article number: 6541.

Ezeani, I., El-Haj, M., Morris, J., Knight, D., 2022. Introducing the Welsh text summarisation dataset and baseline systems. In: Presented at *13th ELRA Language Resources and Evaluation Conference (LREC 2022)*. Marseille, France, 20-25 June 2022.

Ferrario, B., Stantcheva, S., 2022. Eliciting People's First-Order Concerns: text Analysis of Open-Ended Survey Questions. AEA Papers Proc. 112, 163–169.

Huntley, S.J., Mahlberg, M., Wiegand, V., Gennip, Y.v., Yang, H., Dean, R.S., Brennen, M. L., 2018. Analysing the opinions of UK veterinarians on practice-based research using corpus linguistics and mathematical methods. Prev. Vet. Med. 150, 60–69.

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, 7–36.

Knight, D., Morris, S., Arman, L., Needs, J., Rees, M., 2021. Building a National Corpus: A Welsh Language Case Study. Palgrave, London.

Johns, T. 1991. Should you be persuaded: two examples of data-driven learning. In T. Johns & P. King (Eds.), Classroom Concordancing. English Language Research Journal 4, 1–16.

Leech, G., 2006. Teaching and Language Corpora: a Convergence. Routledge, London.

Little, D., 2007. Language learner autonomy: some fundamental considerations revisited. Innov. Lang. Learn. Teaching 1 (1), 14–29.

Maramba, I., Davey, A., Elliott, M.N., Roberts, M., Roland, M., Brown, F., Burt, J., Boiko, O., Campbell, J., 2015. Web-based textual analysis of free-text patient experience comments from a survey in primary care. JMIR. Med. Inform. 3 (2).

McClaughlin, E., Vilar-Lluch, S., Parnell, T., Knight, D., Nichele, E., Adolphs, S., Clos, J., Schiazza, G., 2022. The reception of public health messages during the COVID-19 pandemic. Appl. Corpus Ling. 3 (1), 100037.

Neale, S., Donnelly, K., Watkins, G., Knight, D., 2018. Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. In: Proceedings of the LREC (Language Resources Evaluation) 2018 Conference, pp. 3946–3954. Miyazaki, Japan.

ONS, 2011. Census: Digitised Boundary Data (England and Wales) [Computer File]. Retrieved from: https://borders.ukdataservice.ac.uk. /[Accessed 01/12/2023].

ONS, 2023. Welsh Language Data from the Annual Population Survey: June 2023 [online]. Retrieved from: https://gov.wales/welsh-language-data-annual-population-survey [Accessed 01/12/2023].

Piao, S., Rayson, P., Knight, D., Watkins, G., 2018. Towards a welsh semantic annotation system. In: Proceedings of the *LREC (Language Resources Evaluation) 2018 Conference*. May 2018Miyazaki, Japan.

Rayson, P., 2002. Matrix: a Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison [Unpublished PhD Thesis]. Lancaster University, Lancaster.

Rayson, P., Archer, D., Piao, S., McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the Language Resources and Evaluation (LREC), pp. 7–12. Lisbon, Portugal.