

<b>Institution:</b> 10007140 – Birmingham City University (BCU)		
<b>Unit of Assessment:</b> 27 – English Language and Literature		
<b>Title of case study:</b> Developing automated linguistic analysis and annotation tools to support collaborative learning, professional translation, policy making and HE management decisions		
<b>Period when the underpinning research was undertaken:</b> 01/07/2004 to 31/12/2020		
<b>Details of staff conducting the underpinning research from the submitting unit:</b>		
<b>Name(s):</b> Dr Andrew Kehoe Mr Matt Gee	<b>Role(s) (e.g. job title):</b> Associate Professor Research Fellow	<b>Period(s) employed by submitting HEI:</b> 01/07/2004 to present 14/11/2005 to present
<b>Period when the claimed impact occurred:</b> 01/08/2013 to 31/12/2020		
<b>Is this case study continued from a case study submitted in 2014?</b> Y		
<p><b>1. Summary of the impact</b> (indicative maximum 100 words) Our online WebCorp, eMargin and OurSurveySays tools bridge significant gaps in textual analysis, enabling novel teaching practices and enhanced insights from otherwise unmanageably large datasets. With over 5000 monthly users in 190 countries, our software has:</p> <ul style="list-style-type: none"> <li>• facilitated data-driven language teaching in HE institutions around the world</li> <li>• augmented English language teaching in German secondary schools</li> <li>• enriched the teaching of literary analysis and textual interpretation in HE, FE and schools, with particular growth during the COVID-19 pandemic</li> <li>• improved the accuracy and efficiency of professional translation services</li> <li>• enabled the Belgian and Alicantian chapters of Podemos to formulate political policy</li> <li>• informed decision making by university planners and management at five UK institutions</li> </ul>		
<p><b>2. Underpinning research</b> (indicative maximum 500 words) The Research &amp; Development Unit for English Studies (RDUES) is a corpus linguistic research group, based at BCU since July 2004. While alternative branches of linguistics rely on cherry-picked or invented examples of language use, corpus linguistics uncovers real-world linguistic patterns and trends evidenced in a large-scale digital collection of texts (a ‘corpus’). We have built an international reputation through our work in RDUES, developing novel computational, statistical and linguistic methods, with a particular focus on the <b>automatic identification of word meaning, textual topic and language change</b>. To enable others to engage with our research, we have released the free-to-use <b>WebCorp Live, WebCorpLSE</b> and <b>eMargin</b> software.</p> <p><b>WebCorp tools:</b> WebCorp Live (released in December 2008 after several years of prototyping as WebCorp) was designed to test the hypothesis that the web could complement static offline text collections by providing evidence of rare, new and changing language use. Previous linguistic research relied on searches using the web interfaces of commercial search engines such as Google, but this required researchers to expend substantial effort visiting each web page manually to observe the linguistic patterns within which their search terms occur. WebCorp Live streamlines this approach by processing the results of commercial search engines, automatically accessing the web pages and producing examples of words and phrases with the level of detail required for linguistic study [R01, R02]. While WebCorp Live uses commercial search engines as gatekeepers to the web, the goal of its sister project, the WebCorp Linguist’s Search Engine (WebCorpLSE), was to build a bespoke large-scale collection of web-texts, and thus enable advanced linguistic and statistical analysis of the kind only possible in datasets of known size and composition. We developed linguistically-focused web processing, annotation and search tools and used these to build a large-scale representative sample of the web (a ‘miniweb’), capturing the distribution of document formats, subject domains and web-native text-types [R03], as well as constructing specialist datasets of online news and blogs with their associated comments [R04, R05]. WebCorpLSE was supported by EPSRC, HEFCE and AHRC grants totalling £290,090, with the EPSRC project graded ‘outstanding’ through final peer-review. The WebCorp tools have featured in over 1700 publications by researchers across disciplines, with a multitude of users worldwide.</p> <p><b>eMargin:</b> The initial aim of the eMargin project in 2010 was to bridge the gap between two distinct approaches to textual analysis: the top-down, quantitative approach of corpus linguistics and the</p>		

fine-grained, introspective approach of literary close reading in a classroom context. With the proliferation of eBooks and online databases, the limitations of the traditional close-reading approach were becoming apparent: notes on physical texts become cluttered and are not easily shared or reused, and recreating class-based close reading for distance-learning students is particularly challenging. Despite the proliferation of Web 2.0 technologies, we found none suitable to resolve these issues. Our solution, eMargin, is a web-based annotation tool which, by moving the annotation process online, enables collaboration and discussion across multiple locations in both synchronous and asynchronous modes, as well as retaining a digital record of students' progress. eMargin was developed through two Jisc grants with a combined value of £59,336. Upon completion, the projects were assessed by the funder as meeting a "clear need" and being "designed around the practices and expectations of learners and educators" [R06].

### 3. References to the research (indicative maximum of six references)

[R01] 2008-20. A. Renouf, A. Kehoe & M. Gee. WebCorp Live, WebCorpLSE and WebCorp Learn software systems: <http://www.webcorp.org.uk>. EPSRC project EP/E001300/1

[R02] 2006. A. Kehoe. 'Diachronic Linguistic Analysis on the Web with WebCorp'. In A. Renouf & A. Kehoe (eds.) *The Changing Face of Corpus Linguistics*, Amsterdam: Rodopi, <http://www.open-access.bcu.ac.uk/10188/> (returned to RAE2008). [https://doi.org/10.1163/9789401201797\\_020](https://doi.org/10.1163/9789401201797_020)

[R03] 2007. A. Kehoe & M. Gee. 'New corpora from the web: making web text more "text-like"'. *Studies in Variation, Contacts and Change in English Volume 2: Towards Multimedia in Corpus Studies*, University of Helsinki: [https://varieng.helsinki.fi/series/volumes/02/kehoe\\_gee/](https://varieng.helsinki.fi/series/volumes/02/kehoe_gee/) (returned to RAE2008)

[R04] 2017. U. Lutzky & A. Kehoe "'I apologise for my poor blogging": Searching for Apologies in the Birmingham Blog Corpus'. *Corpus Pragmatics* 1(1), 37-56, <http://www.open-access.bcu.ac.uk/4046/> (returned to REF2021). <https://doi.org/10.1007/s41701-017-0004-0>

[R05] 2019. A. Kehoe & M. Gee. 'Thanks for the donds: A corpus linguistic analysis of topic-based communities in the comment section of The Guardian'. In U. Lutzky & M. Nevala (eds.) *Reference and Identity in Public Discourses*. Amsterdam: John Benjamins: <https://www.open-access.bcu.ac.uk/8340/> (returned to REF2021). <https://doi.org/10.1075/pbns.306.05keh>

[R06] 2011-20. A. Kehoe & M. Gee. eMargin software system: <http://emargin.bcu.ac.uk> (returned to REF2014). Jisc end of award comments: <https://bit.ly/emargin>. Further software details: <http://www.ariadne.ac.uk/issue71/kehoe-gee>

### 4. Details of the impact (indicative maximum 750 words)

**WebCorp Live:** With the ability to search in multiple languages, WebCorp Live has **augmented language teaching and translation** in 190 countries (see Fig 1, S01). Regarding translation, WebCorp Live has been integrated into the search interface at Proz.com, the world's largest community of translators, and recommended as a terminology checking tool by numerous online translation guides and language support groups, including the Terminology Coordination Unit of the European Parliament, which coordinates over 1200 translators [S01]. The fact that WebCorp Live lets translators check the acceptability of their wording not in static dictionaries but against real texts on the web **facilitates translation into the second language of the translator**, a practice traditionally fraught with danger and thus frowned upon. Feedback from a professional translator [S02] notes this advantage and adds

*since I recommended WebCorp to my colleague [...] he's lost without it. WebCorp is now a totally indispensable tool for his work and I'd second that. Since I started to use it in 2016, it is absolutely necessary for quality and productivity reasons.*

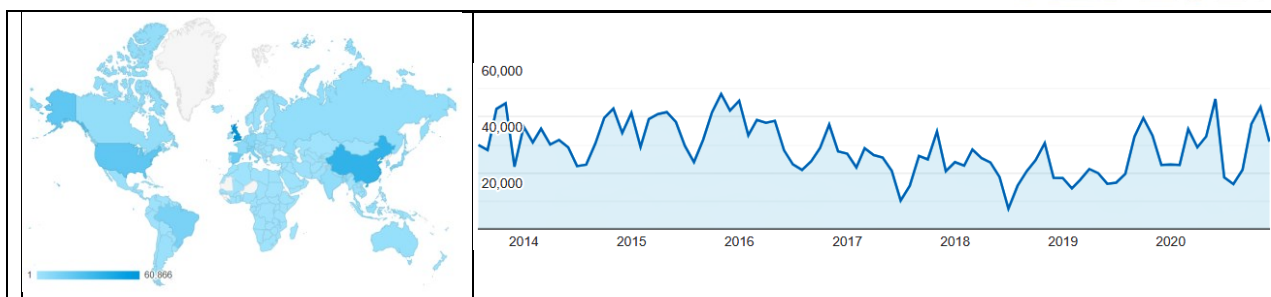


Fig 1 WebCorp Live user locations and monthly page views during REF cycle (Google Analytics)

In teaching, usage records [S01] show that WebCorp Live has been integrated into courses worldwide covering Linguistics, English for Academic Purposes, Teaching English as a Foreign Language (TEFL) and Translation at universities including Sheffield, Nottingham, Essex, Soochow (China), Macquarie (Australia), Washington, Warsaw, Nantes, and the Open University. WebCorp Live facilitates data-driven learning in the language classroom, allowing students to become active researchers, finding and evaluating examples of words and phrases on the web. Gatto [S03] highlights WebCorp Live's value to TEFL students:

*its user-friendliness and ease of access have made it a valuable resource from the start. [...] using the web as a ready-made corpus through WebCorp can immediately improve students' language awareness, and [...] result in an extremely rewarding learning experience that students can easily reproduce outside the classroom.*

**WebCorp Learn:** Expanding on this approach to data-driven language learning, we have further developed the WebCorpLSE technology during this REF cycle with the launch of WebCorp Learn, a tool optimised for interactive English language learning. WebCorp Learn has been integrated into courses at German secondary schools through a collaboration with the Teaching Solutions (TS) language consultancy. According to TS, this collaboration allowed them to **“diversify our service and broaden our customer base”** and **enhanced English language learning in schools**, helping teachers *“realise the language learning goals set out by the government”* for the use of digital language reference tools [S04]. In addition to providing the technology, we created teaching materials (videos and exercises) based on our corpus linguistic research. This enabled TS to deliver seminars and workshops to 20 schools and 50 individual teachers in 2020. WebCorp Learn provided teachers with experience of linguistic tools, examples of real language use, and skills in data-driven learning, in turn enabling them to introduce their students to new methods and technologies. Such technologies were previously *“not used by teachers due to lack of knowledge, access, and ease of use”*, meaning that WebCorp Learn exposed students to linguistic theories and analysis techniques previously *“almost never done in Year 10”* [S04]. Our software provided a new way of teaching vocabulary and improved *“the delivery and discussion of [...] societal and media issues that [...] are the focus of German secondary school English cultural education”* [S04]. One teacher noted that WebCorp Learn *“bridges the gap between school and university”* [S04]. At a time when more teaching is happening online than ever before, there are considerable benefits for students and teachers in using WebCorp Learn:

*The pre-formulated exercises and clear instructions are key benefits, as is the fact the software runs equally well on all systems ([...] mobile phones are often the only device that many students have access to at home).* [S04]

**eMargin:** Between August 2013 and December 2020, 12,629 people registered to use eMargin, uploading and annotating over 7000 texts [S05]. The wide applicability of the tool across disciplines at universities, colleges and schools worldwide is evident in the names of the 2558 groups created by users during this period, e.g. 'Goldsmiths: *Sociology – Marxism*', 'University of California, Santa Barbara: *Spanish and Portuguese*', 'Shandong Normal University: *Chinese as a Second Language*', 'River Valley High School [Singapore]: *English Literature – Year 4*', 'Tolleson Union High School [US]: *Science – Anatomy & Physiology*', 'Bury College [UK]: *Health and Social Care Level 3*'. In addition, eMargin provides solutions for streamlined Virtual Learning Environment (VLE) integration. Over 20 institutions have taken advantage of this by embedding eMargin in their VLEs, including Roehampton, Edge Hill, Minnesota, and Goethe University Frankfurt. A

particularly prolific user of this feature has been Vrije Universiteit Amsterdam, where staff across all faculties have created over 1000 eMargin groups.

Users tell us [S06] that they typically adopt a blended-learning approach, where eMargin **facilitates a seamless transition between face-to-face seminars and collaborative activities outside the classroom**. For instance, a teacher from Saint-Sernin High School, France reports that textual annotations made by students in eMargin before each lesson have allowed them to “[a]nticipate class activities and collective in-depth analysis of texts” [S06]. Furthermore, students and teachers worldwide found eMargin’s blended-learning approach invaluable when face-to-face seminars became impossible during the COVID-19 pandemic. There were 1850 new user registrations between March and December 2020 [S05] with feedback received during that period including:

*I use eMargin with my university Classics students: whenever they have to read longer texts without seminars, they can ask their questions here. In the current situation (covid-19) it is again proving an ideal tool for teaching at a distance. Thank you for this tool!*  
(Assistant Professor, University of Groningen) [S06]

*It’s been invaluable in lockdown as an electronic version of texts with the ability to recap and have a copy of the text even if they missed large periods of teaching during the lockdown. [...] We also used it recently as a department to annotate poetry together as a form of CPD [...] We were able to do this remotely as social distancing in a department of 20 is difficult!* (English teacher, The Streetly Academy, Sutton Coldfield) [S06]

In addition, a lecturer at Goethe University Frankfurt told us that, after her course was forced to move online due to COVID-19, she noticed that all students were taking an active part in discussions of core texts via eMargin, even those who were usually quiet in the classroom [S06].

eMargin was designed as a teaching tool but has been extended to areas of collaborative textual annotation and interpretation. An example comes from the Belgian and Alicantian chapters of the Spanish political party Podemos [S07]. Podemos was founded in 2014 and immediately employed a democratic process in the development of its manifesto. The party uploaded all sections of the draft document to eMargin for party members to discuss. Over 2100 people accessed the texts, with 584 commenting [S05]. Combining so many views would usually be extremely complex but, using eMargin, Podemos was able to **fully engage political party members in the construction of a manifesto**. The party continues to grow and now has over 500,000 members (Feb 2021). In the November 2019 Spanish general election, Podemos secured 12.84% of the vote and a place in the coalition government.

**OurSurveySays (OSS)**: OSS is an application of our WebCorpLSE research in the form of a management tool for open-text survey analysis and insight. It distils our corpus linguistic knowledge and techniques into a web-based visualisation package for use by non-specialists, including marketing strategists, academic planners, and course directors. We designed OSS in 2016 after being asked by BCU’s Planning and Performance Department to assist in analysing responses to the text-based questions in the National Student Survey (NSS). Since then, OSS has become a key tool in **informing policy interventions and identifying priority areas for investment at five UK HEIs**, as demonstrated by user feedback [S08]. For example, the system identified lack of available computers as a significant issue for BCU students in 2016 and provided evidence of which courses were suffering most. Following targeted investment, the proportion of negative comments about computer access halved in 2017, decreased again in 2018 and 2019, and remained low in 2020. The BCU Policy & Strategy Manager notes that OSS has changed the institutional response to NSS, allowing “*teams to make tactical interventions at a local level to improve the academic experience of students on their programme. It was previously impossible to get this level of insight into student feedback due to the prohibitive amount of time [...] taken*” [S08].

Since 2019, Cardiff, Leeds, Edinburgh Napier and Glasgow Caledonian Universities have taken part in a free OSS beta-testing initiative. Leeds used OSS to analyse several thousand comments by respondents to applicant surveys, having trialled but not continued with other qualitative analysis tools in the past [S08]. Analysts at Glasgow Caledonian tested OSS extensively before sharing it with academic leads across Schools, whose feedback noted that the software is tailored



exactly to their needs without being overly complicated, unlike anything they had used before [S08]. Following a trial in 2019, Cardiff adopted OSS for all its qualitative NSS analysis in 2020, leading to the creation of actionable Student Experience Enhancement Plans for each of its Schools. Such analysis would previously have taken two weeks but could be done in under ten minutes with OSS and then shared easily with colleagues drawing up the enhancement plans. For example, academics in one School were able to gain insight into a previously unexplained issue, finding weaknesses in support provided to those students not taking industry placements to be the root cause and employing new initiatives to combat the issue at a localised level. The Student Engagement Officer at Cardiff describes OSS as “ground-breaking” and adds

*On a university level, the question is always “so what” when we analyse data; at the moment, we are stuck with a lot of data and I use OSS to share the insights and the “so what” with key stakeholders. [S08]*

The Head of Market & Student Intelligence at Edinburgh Napier echoes these comments, reporting that OSS

*allows us to support multiple departments with the timely provision of data – ensuring they can feed it into their plans [...] days after we receive the information rather than waiting longer while we go through a fuller data analysis process. [S08]*

OSS thus meets a clear management need, where other tools and methods are lacking, and is ensuring that the student voice is heard clearly across multiple institutions nationally.

#### 5. Sources to corroborate the impact (indicative maximum of 10 references)

##### S01 WebCorp Live usage and referral details

Google Analytics report 01/08/13 to 31/12/20 (location, page views, referrals)

[Web Term Search interface](#) at Proz.com

European Parliament Terminology Coordination Unit pages: [Terminology search tools](#) and [Free term extractors](#)

##### S02 WebCorp Live testimonial from a translator

Email and interview notes from a UK-based translator [Named Corroborator 1]

##### S03 Chapter 4 of the book *Web as Corpus: Theory and Practice*

Gatto, M. 2014. Bloomsbury, 105-118. DOI: [10.5040/9781472542182](https://doi.org/10.5040/9781472542182)

##### S04 WebCorp Learn feedback

Teaching Solutions co-owner testimonial letter [Named Corroborator 2]

Anonymised teacher feedback from workshop with Teaching Solutions

##### S05 Statistics from eMargin user database and server logs

User registrations, texts, groups and VLE referrers (total and for Podemos)

##### S06 eMargin user feedback

Email from teacher at Saint Sernin High School, France

Feedback form from an Assistant Professor, University of Groningen, Netherlands

Email from a lecturer at Goethe University Frankfurt, Germany

Feedback from a teacher at The Streetly Academy, UK [Named Corroborator 3]

##### S07 Podemos party member instructions for using eMargin

Instructions by the regional Podemos groups in [Alicante](#) and [Belgium](#)

##### S08 OurSurveySays university testimonials

BCU Policy & Strategy Manager email

Leeds Senior Market Research Executive interview summary

Cardiff Student Engagement Officer interview summaries [Named Corroborator 4]

Edinburgh Napier Head of Market & Student Intelligence email

Glasgow Caledonian Analyst interview summary