Computational Challenges, Innovations and Future of Scottish Corpora

David Beavan

University College London

Abstract

This chapter discusses the computational challenges and innovations encountered in the development of the Scottish corpora (the Scottish Corpus of Texts & Speech and the Corpus of Modern Scottish Writing), considers how tools for corpus analysis can encourage new audiences and complement existing resources, and explores possible future technological advances for corpus creation and exploitation.

Keywords: *Scottish corpora*, *corpus tools*, *corpus data, challenges*, *technological innovations*

1 Introduction

Embracing technology has enabled the Scottish corpora (the Scottish Corpus of Texts & Speech and the Corpus of Modern Scottish Writing) to develop into rich resources, targeting

wide-ranging audiences from the general public to scholars interested in the minute details of regional variants of the Scots language. Many challenges and innovations have taken place behind the public interfaces presented by these resources: this chapter explains these developments and discusses their nature in detail.[1] Exciting new analytical tools have been developed to exploit the data of the two corpora, bringing access and interpretation of language data to new audiences. The issues of how corpus data can interface with analytical tools are explored in detail. Looking towards the future, and building upon these successes, various possible trends in the development of corpora can be identified: these relate to infrastructure, common platforms and resource discovery. Ultimately, what can this investment in technology give us and what types of new research question can be enabled? These questions will be explored in this chapter, highlighting the prospects for corpus compilers and users. While the specific shapes of the corpora in question are particular to their time and situation, the challenges and solutions which were encountered resonate with those of other resources big and small, local and global.

2 Scottish Corpus of Texts & Speech

2001 marked the beginning of the first of the Scottish corpora, the Scottish Corpus of Texts & Speech (SCOTS). This resource was well placed to benefit from the increased awareness and recognition of Scotland, its people, and its languages which reached an apogee at the time of the reopening of the Scottish Parliament on 1 July 1999. Corpus-based linguistics was well-established, major national corpora had proved their worth, and there was increasing demand

[1] The author worked as Computing Manager for SCOTS and CMSW from their inception to the end of their funded phases. He is now Research Manager at the Centre for Digital Humanities, University College London. The author would like to thank Jean Anderson and Dr Wendy Anderson for their invaluable inspiration and assistance on prior drafts.

for specialist or localised corpora. The final stimulus was technology: the Internet had been shown to be a viable platform for academic dissemination and research; hardware and software were inexpensive, well-supported, and choices were abundant. Establishing SCOTS was therefore timely: it would provide linguists with a new resource focused on the languages of Scotland, and one which they could use with the minimum of requirements as it was to be Internet-accessible.

Advances were being made in two areas: linguistics and technology. Unlike general corpora, the content of the SCOTS resource was to be tightly geographically bound, focussed solely on Scotland. A desire to be roughly representative of the population led to a need to sample widely across genres, registers, and geography, while also meeting word-count targets.[2] This was coupled with a need to provide a web presence advanced enough to deliver meaningful linguistic information to researchers while also being engaging to the public. These goals gave SCOTS a unique ambition; it also presented many unique challenges.

Understanding the enabling nature of technology led to a project team with an enviable balance of language specialists and IT professionals. From the outset, the project could be seen as an example of Literary and Linguistic Computing, recognised today as a subset of Digital Humanities, in the way that it attempted to push the boundaries of established research through the use of computing and technology to the point where new resources could be imagined, and brand new research questions could be posed and answered.

SCOTS responded to a particular combination of challenges facing its digital outputs: it had to be publicly accessible, which was an expectation of its funding; content addition and changes had to be prompt and inexpensive; it had to remain fresh and demonstrate growth;

---

[2] Fiona Douglas, 'The Scottish Corpus of Texts and Speech: Problems of Corpus Design', *Literary and Linguistic Computing*, 18 (2003), 23–37 (p. 27).

and it had to be useable with the minimum number of software prerequisites, to further widen the audience and encourage novice users.

During the late 1990s and early 2000s, many corpora depended upon physical methods for distribution, typically relying on CD-ROM. This allowed the distributor a high degree of control (although not complete control, owing to the potential for duplication), which may be desirable for a commercial enterprise, but not for a public resource. Updating content is also complicated by using physical media, forcing corpus compilers into long release cycles incorporating many changes, rather than frequent or more spontaneous updates. In addition, many corpora at this time required specialist or bespoke query software to be installed on the user's computer (e.g. the British National Corpus, the Helsinki Corpus of Older Scots, and the International Corpus of English).[3] This software could be unfamiliar to the user or incompatible with the operating system of their computer; these factors raised the barrier to access, neither encouraging new users nor allowing for easy engagement with a broader public.

It was clear, therefore, that SCOTS had to provide a web based resource with integrated tools. In this way, the corpus would be distributed world-wide with little cost, the content, which was held centrally, could be tightly controlled, and it was platform-neutral and highly accessible.

The scale of the content (a four million word corpus, about twenty percent of which was to be spoken language) also presented a major challenge to the team. Alongside each document extensive metadata was captured, concerning the document itself, its setting,

---

[3] British National Corpus, <http://www.natcorp.ox.ac.uk/>; Helsinki Corpus of Older Scots, <http://www.helsinki.fi/varieng/CoRD/corpora/HCOS/>; International Corpus of English, <http://ice-corpora.net/ice/> [these and all other links in this chapter were last accessed on 20 August 2012].

authors and participants.[4] The time period to be covered (1945 to the present) meant that the entire corpus would comprise in-copyright texts, necessitating comprehensive permission-gathering protocols and documentation. The range of texts sampled was diverse, for example advertisements, personal correspondence, business invoices, creative writing, and spontaneous speech.[5] This proved costly in terms of time and resources, as it involved directly contacting authors and participants as well as other rights holders.

As the project was collecting modern material, including unpublished material such as personal writing, there were few occasions to re-use existing digitised material from potential collaborators. Instead, the vast majority of the corpus was built from scratch, by seeking contributions directly from the public, for example through targeted calls or publicity in the media. While this gave the project an excellent opportunity to gather texts, the make-up of the corpus was dependent on volunteers donating their texts, or taking part in recordings. This was a risky strategy.

Corpora in the early and mid-2000s were designed for academics, and rarely tried to target the public. If SCOTS was to meet its goals, it needed to win over this new audience. It was vital that likely contributors could not only see what they were contributing to, but also that they understood the value of such a resource. The public face of the project, its web site, became a vital tool in this public engagement. It had to be established as early as possible, to articulate its purpose appropriately, and demonstrate as much functionality as possible, all to help persuade the public to get involved.

---

[4] See further Jean Anderson, David Beavan, and Christian Kay, 'SCOTS: Scottish Corpus of Texts and Speech', in *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, ed. by Joan Beal, Karen Corrigan, and Hermann Moisl (Basingstoke: Palgrave Macmillan, 2007), pp. 17–34 (pp. 19-22).

[5] Anderson, Beavan, and Kay, p. 20.

Creating a corpus from first principles was a difficult task, given the demanding total word-count and the challenges of sampling across different genres and registers. Actively soliciting documents from the public, coupled with issues such as copyright and licensing meant the team had numerous administrative requirements. As this workload was spread across multiple team members, an overarching management system was needed. This was achieved through a centralised interface for all corpus compilation tasks: contact management, submission control, metadata capture, and copyright permission recording.

The starting point was a Microsoft Access contact management database created in the initial months of the project to assist the tracking of outgoing mailings and associated requests for texts. Through many iterative cycles of requirements gathering and testing, the final administrative database was created. The front-end continued to use Access, benefiting from its ability to form user interfaces quickly, as well as its powerful tools to connect with other Office software, such as Microsoft Word for mail-merges. Data-storage was entrusted to an open-source Relational Database Management System, PostgreSQL, a powerful and reliable database server.[6] This splitting of interface and storage allowed for concurrent multiple users, and opened up opportunities to probe the database directly, which was important in the formation of the web site. This bespoke solution allowed all aspects of the corpus to be created, from metadata to document text contents, all in one place. It also allowed the team to track progress and obtain reports on word-count subtotals.

Online delivery was the chosen mode of dissemination for SCOTS, and the corpus had to target the widest possible audience, while also providing researchers with the data and tools they would need fully to exploit the content. From the initial public launch on 30 November 2004, the web site continued to grow in content and functionality. Since that time the online

[6] PostgreSQL, <http://www.postgresql.org>. On the Open Source Initiative, see <http://opensource.org/>.

corpus has been updated seventeen times, and has increased in size beyond its original target: it now stands at over 4.5 million words.

The corpus infrastructure borrowed heavily from the administrative database, as the online corpus was fed from data here, performing actions such as verifying that all copyright permissions had been obtained, and also censoring sensitive personal data. Further open-source software packages were used, chosen for their flexibility, wide-scale adoption, and performance. For example, the Apache web server was employed to answer all page requests from users interacting with the site, using PHP to drive the dynamic elements of the site, such as the search functions.[7]

Different constituent user groups were taken into consideration: scholars needed tools, statistics, and the possibility of downloading texts, whereas it was expected that the public would focus on the exploration, browsing, and reading of the full text documents. The interface addressed this by providing a number of methods of accessing the content. By browsing, a user could see all documents and easily reach their full textual content. A standard search interface offered the most common search criteria and provided basic statistical data. An advanced search option extended this by introducing a flexible search, incorporating all 250+ possible metadata fields, and added geographic representations via Google Maps.[8] At all times the user could gain access to the full document, its associated metadata, and the function to download all of this for offline use.

To complement the written portion of the corpus, all spoken texts were orthographically transcribed and presented to the user alongside the audio footage. Spontaneous speech, particularly when it involves a group of individuals rather than just one

---

[7] Apache HTTP Server Project, <http://httpd.apache.org/>; PHP Hypertext Preprocessor, <http://www.php.net/>.

[8] Google Maps, <http://maps.google.com/>.

or two speakers, presents its own problems, chiefly that of overlap. During natural speech speakers do not simply take it in turns to talk, they readily overlap their utterances, talking over each other up to seventeen percent of the time.[9] Transcribing content of this nature must take this overlap into account: it must not simply present a sequence of utterances, but instead allow for more than one speaker at the same time. The software chosen for this task was Praat, primarily a tool for phonetic analysis, but one which allows for overlap through its concept of tiers, each attributed to an individual speaker.[10]

The greatest challenge, and where the project really innovated, was to tie this orthographic transcription to the audio/video footage, and to present this over the Internet. The first stage was to convert the transcription from Praat format, while preserving all overlap and timing data. This allowed the web site to present a more traditional linear transcription to the user for ease of reading: this featured colour-coded speaker utterances and the marking of overlap, as well as other features such as false starts, non-lexical items (e.g. laughter, coughs, sighs, etc.) and the marking of inaudible passages.

To provide cross-platform multimedia playback, Apple QuickTime was used in the form of a browser plug-in.[11] This was exploited in such a way as to keep the footage and transcription in synchronisation: scrolling the transcription to keep pace with the footage, or allowing to user to skip to a given passage by re-positioning the player at an arbitrary point,

---

[9] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, 'Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation', in *Proceedings of Eurospeech 2001*, ed. by Paul Dalsgaard, Børge Lindberg, Henrik Benner and Zheng-Hua Tan (International Speech Communication Association: Aalborg, 2001), ISBN 8790834097, pp. 1359-1362

[10] Praat, <http://www.fon.hum.uva.nl/praat/>.

[11] Apple QuickTime, <http://www.apple.com/quicktime/>.

or clicking on an utterance in the transcription to start playback from that point.[12] This solution gave the online users of SCOTS unparalleled ways to navigate its transcribed content.

3 Corpus of Modern Scottish Writing

Following on from the success of SCOTS, the Corpus of Modern Scottish Writing (CMSW, 1700-1945) began in 2007 to fill the chronological gap between the Helsinki Corpus of Older Scots (HCOS, 1450-1700) and SCOTS (1945-present). The addition of this new resource would allow scholars for the first time to study corpora of the language varieties of Scotland covering 650 years. Many of the SCOTS team members continued to work on CMSW, which gave the project an impetus in terms of skills and experience.

CMSW was not to be merely an adaptation of SCOTS, as new challenges faced the team at every stage. The content of the resource would have a very different profile: none of the documents would have been born digital, meaning that document capture would be a large and important task. Achieving balance or sampling across many factors would be difficult, as the availability of documents was limited, and while copyright did not apply to many of the published works, it was still a relevant issue for unpublished personal writing.

In particular, the creation of the resource would be driven by the need to answer new research questions, a demand which SCOTS did not have to the same extent. This research would focus on the development of Scottish Standard English during and after the Scottish Enlightenment, tracing the interaction between Broad Scots and written Standard English. Alongside the main four-million-word corpus, CMSW includes one million words of

[12] For more on this, see Wendy Anderson and David Beavan, 'Internet Delivery of Time-Synchronised Multimedia: The SCOTS Corpus', in *Proceedings from the Corpus Linguistics Conference Series* 1(1) (Birmingham, 2005), ISSN 1747-9398.

orthoepist material, an invaluable resource consisting of contemporary records of language commentary and examples of usage, including such texts as *Rules to be Observed by the Natives of Scotland for Obtaining a Just Pronunciation of English* (Francisque Xavier Michel, 1882) and *Rules to be Observed by the Natives of Scotland for Obtaining a Just Pronunciation of English* (John Walker, 1791).

Establishing partnerships with libraries, archives, and other institutions would prove to be essential for the project. These links provided access to many of the documents to be included in the corpus, whether personal writings or well-known novels by authors such as Robert Burns, James Hogg, and Allan Ramsay. Chief amongst the collaborators were the University of Glasgow Library (including Special Collections),[13] the University of Glasgow Archive Services,[14] the National Library of Scotland (NLS), especially the John Murray Archive (JMA),[15] and the Mitchell Library.[16]

With a number of partners came a number of different agreements and a range of different combinations of access to physical and (where it existed) digital material. To be efficient in terms of time and resources, the project team had to make the best use of existing digitised or transcribed material where possible. In many cases there was no room to influence the processes, protocols, or formats used by the host institutions, meaning that the part of CMSW's task would be in the aggregation of data. Ideally, the project team would have preferred the highest quality camera imaging and its raw format digital negatives for all

---

[13] University of Glasgow Library, <http://www.lib.gla.ac.uk/>; University of Glasgow Special Collections, <http://www.gla.ac.uk/services/specialcollections/>.

[14] University of Glasgow Archive Services, <http://www.gla.ac.uk/services/archives/>.

[15] National Library of Scotland, <http://www.nls.uk/>; John Murray Archive, <http://digital.nls.uk/jma/>.

[16] Mitchell Library, <http://www.glasgowlife.org.uk/libraries/the-mitchell-library/>.

images, coupled with quality-assured transcriptions in Text Encoding Initiative (TEI) format to capture both the presentation and the semantics of the document.[17]

The John Murray Archive at the NLS possessed many digital images, with accompanying transcriptions in TEI format, expressed in Extensible Markup Language (XML).[18] These were very close in specification to the outputs that CMSW wished to offer. The JMA material was also comprehensively catalogued, assisting the process of selecting a spread of documents across their collection.

Low-cost digitisation methods were used by the team at the Mitchell Library, because the material was not able to leave the reading room or be processed by third parties. This presented many challenges for image capture due to the lack of a controlled environment; this, however, was balanced by ready access to all relevant documents and collections.

At the University of Glasgow, flatbed scanning was used by CMSW for many documents and books, seeking specialist imaging solutions through the Photographic Unit in cases where individual items were fragile or otherwise needed special treatment. Archive Services used the opportunity provided by our interest in their documents to perform extensive curation on them prior to digitisation.

As can be seen, there were many different technologies used in the capture of the digital images needed for the corpus. The CMSW internal work-flows in turn had to be flexible enough to absorb material of different specifications and stages of readiness: from digital negatives taken with professional medium-format cameras under controlled lighting, to scanned images captured in office-like environments.

To build the corpus and to allow for searching and exploitation, electronically readable transcripts of the source material had to be produced. This machine-encoded text was

---

[17] Text Encoding Initiative, <http://www.tei-c.org/>.

[18] Extensible Markup Language, <http://www.w3.org/XML/>.

to be provided to the user, together with the images of the source document. It was important that the quality and durability of all textual depictions were of the highest quality: it was this aim that drove the work-flows and procedures that the project adopted.

For printed material, the project harnessed Optical Character Recognition (OCR) technology to produce the machine-encoded text of each document. To achieve the best quality OCR, the source images were extensively processed prior to using OmniPage.[19] Post-capture splitting of images was performed in situations where the supplied image contained both recto and verso sides: custom batch scripts using ImageMagick were employed to detect and crop the image along the binding.[20] To further provide an image amenable to OCR, unpaper was used to create a greyscale representation of the page, straightening the text and removing image artefacts.[21] The automatic OCR process was run, then the resulting data corrected and proof-read by at least two different project staff before being quality approved.

Unlike printed material, there were no reliable or established means to automatically read handwriting, therefore the creation of the electronic textual depiction of handwritten material had to be entirely manual. Images were processed to achieve correct colour balance, and the greyscale manipulations, unnecessary with texts which had not undergone optical character recognition, were omitted. Transcription of the documents was performed using TEI mark-up, expressed in XML, working alongside images of the source document. Again, two team members proof-read and quality checked each document before it went live.

Both work-flows were time consuming: the manual transcription of handwritten letters was painstakingly slow, especially where the documents had degraded or the writing was barely legible; the OCR process generated its own errors, often changing Scots forms to

---

[19] OmniPage, <http://www.nuance.com/for-business/by-product/omnipage/>.

[20] ImageMagick, <http://www.imagemagick.org/>.

[21] unpaper, <http://unpaper.berlios.de/>.

Standard English ones, or misidentifying the long-s (ſ) and instead littering documents with a lowercase f.

The process of building CMSW called for different administrative functions compared to SCOTS, as more effort was focussed on identifying potential document sources, on seeking individual documents from collections, and on the more intensive tracking of image manipulation and transcription tasks. A shared Microsoft Access database was developed to record all the document metadata, and this was allied to a file-based structure which contained the document contents (source images, manipulated images, transcriptions, etc.). To prevent these two data stores from operating in complete isolation from each other, there were formal protocols to identify and join records. This was a solution which gave the project team the flexibility to handle the wide variety of files from partners and the different work-flows which this required, but one which ultimately required greater effort and resources to keep in synchronisation, particularly as the number of documents grew.

The demands of the CMSW web site were also different to those of SCOTS. For example, the need to deliver audio and video was replaced by the need to deliver document images to the user. A larger change, however, was taking place under the surface, shifting the balance of computational effort: as the administrative functionality was not as extensive, more work had to be done by the web facing components to prepare the source material for user delivery. Documents could not simply be copied from an administrative database to an online database as before, instead they had to be compiled from both administrative data sources (database and file structure). Files of all the varying formats and specifications were seamlessly identified and converted, using different work-flows, to form a consistent representation of the corpus. Once in this unified state, the corpus could be indexed, providing the basis for the search tools and statistical measures.

In essence the corpus had been divided into four increasingly independent sections or stages: the administration and raw data; the functions to normalise this data; the corpus in its indexed form; and finally the online delivery and its associated tools. The structure that CMSW took can be seen as modular, rather than the comparatively monolithic form of SCOTS. The overwhelming benefit of taking this approach was flexibility: development of the distinct stages could take place with little impact on the other processes, new data sources could easily be accommodated, and the data could be converted and output to the online corpus structure. On the other hand, a disadvantage of this approach was the problems which arose in the critical interface between each pair of steps; the expected inputs and outputs had to be precisely defined, and the data had to adhere to this specification.

A more generic and flexible way to deal with corpus data had emerged with CMSW, one which could assist with the building of new linguistic resources, and one which could play a part in the bringing together of existing corpora. A longer-term opportunity to merge SCOTS and CMSW was also becoming a possibility. To work towards this and to explore more generic ways to access and manipulate linguistic data, a series of experimental tools was developed alongside SCOTS and CMSW, using those corpora and others to offer new visualisations of linguistic data.

4 Analytical tools

Both SCOTS and CMSW offered users a suite of tools to inspect and analyse each corpus, often providing the same functionality, for example, basic statistics such as normalised word frequency, as well as key word in context (KWIC) concordance lines. The way these tools were implemented was unique to each resource, due to the particular nature of how the linguistic data was modelled and stored. As a result, this tight integration of data and tools was not as efficient as it could be.

Although compared to SCOTS, CMSW moved towards a more generic model of collecting, manipulating, and storing linguistic data, it did not fully separate the data from the tools. This would require a further step, and to explore this, a series of experiments were initiated to allow the SCOTS and CMSW teams fully understand how to repurpose their data and tools.

With a well described and consistent representation of the corpus data in place, other tools, such as advanced visualisations, could be developed with comparative ease. These tools, like the main corpus search facilities, could rely on the infrastructure to perform the advanced computational manipulations of linguistic data. Common tasks, such as statistical measures or tests on the corpus data could be provided to these tools through functions or APIs (Application Programming Interfaces) providing a level of abstraction. In turn, the infrastructure could contain multiple corpora, allowing them to be queried using common methods, sharing the tools to allow the user to interrogate them.

As part of these experiments a third, externally developed corpus was included: the British National Corpus (BNC). The BNC was used alongside SCOTS and CMSW to test and demonstrate the applicability of the new methods to other data sets. For simplicity at this early stage, the lowest common denominator of corpus mark-up present was used, in this case, plain text. The aim of the experiments was to provide visualisation tools, offering new and intuitive methods to explore large-scale linguistic data. This was a task which would forge technical and structural developments towards a generic corpus infrastructure.

4.1 Collocate Cloud

To make these new tools attractive to the widest possible audience, no assumptions could be made about the users' prior knowledge of corpora, existing corpus tools, or analyses. They were designed to be instantly engaging, and to provide insights into language without the

need for additional knowledge, resources, or tools. The first of these tools to be released was the Collocate Cloud.[22]

Linguists often use collocational information as an instrument to examine language use, that is, to inspect the words which cluster with significant frequency around other words. Collocation is summed up by Firth's memorable phrase: 'You shall know a word by the company it keeps'.[23] Collocates can beautifully illuminate how a word is used, by providing invaluable contextual information. This is information rarely present in dictionaries, but present in corpora, as they are based upon a sampling of real-life language in use. Traditionally, collocates are displayed in tabular format: however, the Collocate Cloud visualisation was designed to bring the information to life and to a new public.

[22] BNC Collocate Cloud, <http://www.scottishcorpus.ac.uk/corpus/bnc/>.

[23] John Firth, 'A Synopsis of Linguistic Theory, 1930–1955', in *Studies in Linguistic Analysis. Special Volume, Philological Society* (1957), 1-32 (p. 11).

Figure 1: Collocate Cloud showing the words co-occurring significantly with *bank*

The Collocate Cloud visualisation for the node word *bank* in the BNC is shown in Figure 1. The display takes cues from well-established tag clouds, popularised by the likes of Flickr and Amazon.[24] The top one hundred words co-occurring with *bank* are shown in alphabetical order, allowing quick discovery of a known word. The font size of collocates reflects their frequency; the larger the size, the more often that pair is found. Their brightness signifies collocational strength, that is, how often these words are exclusively found together.

The Collocate Cloud promotes serendipitous discovery through the possibility for the user to click on any word in the cloud to instantly generate a new cloud with the selected word as the new node word. This near-unending browsing experience is ideal for non-corpus linguists, affording any interested user the opportunity to feel their way around a body of language.[25]

The Collocate Cloud was designed to complement and extend existing tools. Using established statistics, it performs calculations on the linguistic data as other collocate lists do; however, the presentation is very different. Users who find value and interest in a particular Collocate Cloud visualisation are therefore free to use their tried and trusted tools or software packages to repeat the search, in order to access additional data otherwise not exposed in the cloud.

---

[24] Flickr Explore Tags, <http://www.flickr.com/photos/tags/>; Amazon Most Popular Tags, <http://www.amazon.co.uk/gp/tagging/cloud/>.

[25] For more information on Collocate Cloud, see David Beavan, 'Glimpses Through the Clouds: Collocates in a New Light', in *Digital Humanities 2008* (University of Oulu, 2008), 25-29.

This aggregation of data, distilling millions of words of content into a mere hundred or so, is both an enabler and a barrier. The visualisation provides an excellent overview, but as the above example suggests, this process also conceals some of the raw data from the user. As a tool for the distant reading of an entire corpus it is very successful, but it is not to be seen as a substitute for traditional corpus linguistic practices, more as a gateway to them.

Examples of the Collocate Cloud were made publicly available via the Scottish Corpora web site, for SCOTS as well as the BNC, demonstrating proof of concept. As evidence of support for the corpus-agnostic nature of the visualisation, the Collocate Cloud won Best Idea for Improving an Existing Tool in the 2008 TADA (Text Analysis Developers' Alliance) Research Evaluation eXchange (T-REX).[26] Moreover, the lessons learned in structuring the underlying data model and maintaining fast levels of performance were invaluable to the team.

4.2 ComPair

The experience of developing the Collocate Cloud was a building block for further experiments. After feedback from the linguistic community at large, the ability to make comparisons between clouds was sought. Rather than crudely comparing two Collocate Clouds side by side, a new visualisation was developed to show the range of collocates relating to the selected two node words. This had the potential to expose varying cultural aspects of language by illuminating differing attitudes, as well as demonstrating degrees of synonymy.

[26] 2008 TADA Research Evaluation eXchange (T-REX), <http://tada.mcmaster.ca/trex/>.

This was an evolution of the Collocate Cloud, extending the techniques while retaining the tool's ease of use, stimulation of discovery, and clarity in the information displayed. To this end, ComPair was developed, as seen in Figure 2.[27]

Figure 2: ComPair showing the collocates between words *utterly* and *absolutely*

The user enters two node words, in Figure 2 these are *utterly* and *absolutely*. The collocates of both node words are calculated in accordance with the routines the Collocate Cloud relies

[27] BNC ComPair, <http://www.scottishcorpus.ac.uk/corpus/bnc/compair.php>. For more information, see David Beavan, 'ComPair: Compare and Visualise the Usage of Language', in *Digital Humanities 2011* (Stanford University, 2011), 19-22.

upon. Instead of forming two separate clouds, however, the collocates are distributed along a continuum between the two node words, their closeness to each node word indicating collocational strength with that particular word, and a lack of connection with the other node word. Colour (absent from Figure 2, but available via the web based tool) is inherited from the node words, and as before brightness is indicative of the degree of collocational strength. For example, Figure 2 indicates that the word *ridiculous*, through its brightness and location toward the centre of the plot, is used often in conjunction with both nodes (*utterly* and *absolutely*).

Louw in 1993 introduced us to semantic prosody, which describes how synonymous words can take on positive or negative connotations.[28] In Figure 2, the ComPair visualisation demonstrates this clearly. Negative words such as *disgraceful*, *condemn*, and *ruthless* cluster towards the node word of *utterly*, while positive words such as *marvellous*, *delighted*, and *brilliant* are found with the node word *absolutely*. To the novice, the two node words *utterly* and *absolutely* may at first glance appear to be completely synonymous and interchangeable: however, this is certainly not the case when corpus data is scrutinised. The ComPair visualisation is clearly a candidate resource for language learners, allowing an inspection of language beyond dictionaries or thesauri. Of course, ComPair can be used to contrast any two words, not just (near) synonyms. In fact it can just as easily be used to explore attitudes towards any subject matter, be it *football* vs *rugby* or *cats* vs *dogs*.

Like the Collocate Cloud, ComPair provides a distant view of the data, showing general trends in the corpus, filtering and smoothing out specific details. It is at its best when used for quick investigative work, or to provide the opportunity for users without a

---

[28] Bill Louw, 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in *Text and Technology*, ed. by Mona Baker, Gill Francis, and Elena Tognini-Bonelli (Amsterdam: John Benjamins, 1993), pp. 157-176.

background in text analysis to make casual enquiries on an ad hoc basis. It will not replace established analysis methods, but it is designed to provide a new starting point for discovery.

The technical challenges of delivering these comparisons in real-time to the user further tested the corpus infrastructure, and provided a test-case beyond that of Collocate Clouds. As before, instances of ComPair were made publicly available, operating on both the BNC and SCOTS corpora. The lessons learned from these experiments underlined the advantages of separating the data from the computational measures and tools. While extra effort was needed to break the corpora down to a consistent and level state, the benefits of a suite of tools which could operate upon these, or on any other corpora, was persuasive.

5 Future directions

As illustrated above, there are benefits of scale and reusability when existing tools are able to operate with a great number of corpora. This has been commonplace practice on the desktop, with popular, current examples including the likes of WordSmith, AntConc, and Xaira.[29] In these examples, both the corpora and the software must exist on the user's computer, which provides barriers to access, whether of availability, cost, or the demand for prerequisite skills or knowledge.

The almost polar opposite exists in the case of web based resources, where the corpus data and tools must exist together, or at least give the logical appearance of co-existing. Notable exceptions do exist, such as Wmatrix, Sketch Engine, Voyant Tools, and indeed the

[29] WordSmith Tools, <http://www.lexically.net/wordsmith/>; AntConc, <http://www.antlab.sci.waseda.ac.jp/software.html>; Xaira, <http://projects.oucs.ox.ac.uk/xaira/>.

Enroller project.[30] However, none of these truly provides the opportunity to mix and match tools and corpora on an ad hoc basis.

Recent large-scale endeavours such as Digital Research Infrastructure for the Arts and Humanities (DARIAH), Common Language Resources and Technology Infrastructure (CLARIN), Project Bamboo and Text Grid aim to extend the methodology attested by ENROLLER, they plan to enable a distributed network of corpora (alongside other resources) and tools.[31] Those individuals, institutions or countries that have access, will for the first time be able to interoperate their corpora and tools for analysis in an open framework. The availability, permissions permitting, of a large number of corpora allows researchers the ability to combine corpora to study problems for which there is a no substitute for a large data-set, e.g. the study of very infrequent linguistic occurrences. Likewise, being able to access a large catalogue of tools promotes the analysis and exploration of any corpus, and provides trusted side-by-side comparison of outputs.

It is essential that every corpus that exists in these infrastructures is well described, in both technical and non-technical senses. There may be a temptation to dispose of project-specific metadata or mark-up, particularly if cognate resources do not describe themselves in those ways, or if there are no tools that take full advantage of the specific details available. To combat this perceived lack of utility, it is possible to transform project-specific information into more generally expressed and used forms, or to make use of extensions to established formats. This would certainly be the case for instance in relation to the large number of

---

[30] Wmatrix, <http://ucrel.lancs.ac.uk/wmatrix/>; Sketch Engine, <http://www.sketchengine.co.uk/>; Voyant Tools, <http://voyant-tools.org/>. On the *Enroller* portal, see Jean Anderson, 'Enroller: an Experiment in Aggregating Resources', this volume.

[31] DARIAH, <http://www.dariah.eu/>; CLARIN, <http://www.clarin.eu/>; Project Bamboo <http://www.projectbamboo.org>; Text Grid <http://www.textgrid.de>.

personal metadata fields SCOTS holds, likewise in the linking of images to their textual counterparts that CMSW employs.

If not directly hosted by these infrastructure projects, corpora wishing to operate in their environments should be automatically discoverable by computer. They need to be able to technically describe themselves, their provenance, their data structure and their possible interactions in order to co-operate with other resources and tools. When operating in the relative isolation of a traditional corpus web site, these needs are not as pressing or altogether absent. This is perhaps the biggest challenge facing established corpora wishing to embrace these new technologies, and one which may be resource intensive to solve.

While the division of corpora and tools makes computational and research sense, it establishes a potential new inter-dependence and reliance on these frameworks. Future corpus compilation projects should still embrace their ownership of the data they produce, and plan to be flexible with how their resources are shared. They should also continue playing a leading role in the shaping of the tools that operate upon them and other resources.