Making corpus data visible: visualising text with research intermediaries

William Allen[1]

Abstract

Researchers using corpora can visualise their data and analyses using a growing number of tools. Visualisations are especially valuable in environments where researchers communicate and work with public-facing partners under the auspices of 'knowledge exchange' or 'impact', and corpus data are more available thanks to digital methods. However, although the field of corpus linguistics continues to generate its own range of techniques, it largely remains orientated towards finding ways for academics to communicate results directly with other academics rather than with or through groups outside universities. Also, there is a lack of discussion about how communication, motivations and values also feature in the process of making corpus data visible. My argument is that these sociocultural and practical factors also influence visualisation outputs alongside technical aspects. I draw upon two corpus-based projects about press portrayal of migrants, conducted by an intermediary organisation that links university researchers with users outside academia. Analysing these projects' visualisation outputs in their organisational and communication contexts produces key lessons for researchers wanting to visualise text; consider the aims and values of partners; develop communication strategies that acknowledge different areas of expertise; and link visualisation choices with wider project objectives.

Keywords: Big Data, immigration, impact, knowledge exchange, research intermediaries, visualisation.

[1] Centre on Migration, Policy and Society (COMPAS), University of Oxford, 58 Banbury Road, Oxford, OX2 6QS, United Kingdom.
Correspondence to: William Allen, *e-mail*: william.allen@compas.ox.ac.uk

Corpora 2017 Vol. 12 (3): 459–482
DOI: 10.3366/cor.2017.0128

www.euppublishing.com/cor

1. Introduction

Researchers using text as data can use an increasing number of visualisation tools and techniques. They come in forms built for text, such as tag clouds, Wordles or network diagrams (Brezina *et al.*, 2015; Dörk and Knight, 2015; and Viegas *et al.*, 2009), as part of visualisation features embedded in more comprehensive linguistic software (Kilgarriff *et al.*, 2014), or through general packages like Tableau Public (2016). In some ways, this variety and availability is heartening: visualising aspects of corpus data can be useful for discovery as well as for communicating results. Perhaps more than ever before, visualisation is within the reach of a wide range of researchers and those working with data of many different types (Gatto, 2015).

In other ways, though, much existing work on the subject in corpus linguistics is 'aimed mainly at people with expertise with linguistics' (Dörk and Knight, 2015: 84). There is less reflection on the decision processes involved in creating, designing and justifying visual representations of corpora and corpus analysis for non-expert audiences (Gough *et al.*, 2014). This gap is not necessarily specific to linguistics. Indeed, a great deal of the debate about what makes a visualisation generally 'effective' tends to focus on technical dimensions like user experience or the details of particular designs (Kennedy *et al.*, 2016b).

My aim here is to widen discussions about visualisation in linguistics to include an appreciation of research intermediaries' values, positions, and relationships with academics and visualisers as they interact during the process of making visualisations. Research intermediaries are people or organisations that link academics with end-users. Although they may be part of academic or end-user groups themselves, intermediaries are characterised by their ability to repackage, translate or facilitate understanding of research for use by policymakers, media, civil society organisations or members of the public (Knight and Lyall, 2013; and Tseng, 2012).

Instead of reporting on new tools or techniques for visualising corpora, I identify how different contexts and constraints feature in the process of making corpus data visible for groups outside research settings. I draw upon two corpus-based projects initiated by the Migration Observatory, an organisation at the University of Oxford that engages in intermediary activities on the issue of immigration. By demonstrating how non-technical factors like organisational values and communication among stakeholders featured in the development of the visual outputs, I argue that these dimensions need to be considered alongside the technical 'how-to' of making visualisations. As a result, this research speaks to ongoing discussions about data visualisation occurring not only in the humanities but also in the social sciences. Here it is important to note that I focus specifically on the process of visualising text, not on how users perceived the two visualisations that came out of the research projects. Those findings, which are important to highlight as part of the story of how visualisations communicate different kinds of insights, are available in Kennedy *et al*. (2016b) and Kennedy and Hill (2017).

Section 2 outlines two broader changes in the nature of conducting and communicating corpus linguistic research that make visualisation particularly appealing and worthy of examination. Section 3 explains what visualisations are and their role in current corpus linguistic research. Then, Section 4 considers the empirical data and methods used in two corpus-based projects that resulted in the visualisations, as well as a consideration of the organisational values that informed this work. Section 5 reports on three key lessons that emerged from reflecting on the decisions made during those projects. Finally, Section 6 concludes by drawing parallels between the expansion of corpus linguistic methods and the possibilities of visualisation for enhancing rather than replacing expertise.

2. The changing nature of conducting and communicating linguistic research

Two broad changes have implications for the way that corpus linguists relate to the wider world, and contribute to the rising importance of visualisation in corpus linguistics. The first is the popularity of Big Data approaches, where the amount of new sources of data introduces opportunities and challenges for building and analysing corpora. The second is the increased role of research intermediaries that link academics with other users – a process expressed through ideas of 'knowledge exchange' and 'impact'. This section explains how these two changes heighten the importance of critically examining visualisations not only as objects themselves but also how they exist within particular contexts and achieve particular ends.

2.1 Big Data and corpora

From a technical perspective, Big Data often refers to datasets that are large enough to require significant computing power in their analysis. Due to their huge size, complexity or speed at which they grow, Big Data are sometimes differentiated from so-called 'small data' by their 'volume, velocity and variety' (Laney, 2001: 1). However, as boyd and Crawford (2012: 663) observe, what actually sets Big Data approaches apart is the 'capacity to search, aggregate, and cross-reference large data sets'. In 2008, the editor-in-chief of Wired Magazine at the time argued that the apparently comprehensive nature of these datasets replaces analysis and theory:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Who knows

> why people do what they do? The point is they do it. [. . .] With enough data, the numbers speak for themselves.
>
> (Anderson, 2008)

Some critical scholars have reacted to this view of Big Data, pointing out the continued need for acknowledging 'the situatedness, positionality and politics of the social science being conducted' (Kitchin, 2014: 10). Numbers rarely, if ever, speak for themselves in straightforward ways. Meanwhile, others have observed that most people do not directly access or encounter data in their original forms anyway. Instead, this access is often mediated through platforms, search engines or corporate interests (Couldry and Turow, 2014). Also, large datasets present issues of provenance and validity, particularly in the case of social media data (Driscoll and Walker, 2014). If sources and the decisions leading to their inclusion or exclusion are opaque, then this has implications for how researchers treat and interpret the resulting analyses. Finally, the use of Big Data has transformative impacts at the societal level. It raises questions about who has access to these data – as well as the required skills and resources to collect, organise and make sense of them – and who does not (Graham, 2011; and Kitchin, 2014).

Big Data approaches have also appeared in discussions about corpora. In the early 2000s, Kilgarriff and Grefenstette (2003: 333) made mention of corpus size when they summarised the then-emerging phenomenon of 'Web as Corpus':

> Language scientists and technologists are increasingly turning to the Web as a source of language data, because it is so big, because it is the only available source for the type of language they are interested in, or simply because it is free and instantly available.

Since then, the development and availability of large reference corpora drawn from diverse sources exemplifies how digital access to more textual data has transformed contemporary study of language. Meanwhile, techniques like webscraping and data mining have enabled innovative study of online language that relies upon corpora containing blogs, tweets, comments or forum postings (Hardaker, 2010). And these textual datasets can be linked with other forms of data – like those held in Geographic Information Systems (GIS) – to extend linguistic knowledge even further (Gregory and Hardie, 2011).

The central ideas of 'volume, variety, and velocity' are just as relevant for lexicographers and linguists who build, maintain and share corpora, especially in cases where corpora are either derived from online sources or archives, or are linked with other kinds of databases. But as the quantity and availability of corpora and textual analysis software flourishes, it is imperative for corpus linguists to consider for whom and for what purposes such data and tools exist. Fruitful, interdisciplinary debates about Big Data have identified limitations and inequalities associated with their uncritical use

(Andrejevic, 2014; and boyd and Crawford, 2012). Similar questions need to be asked of the changing ways that corpora are constructed, analysed and eventually visualised.

2.2 Intermediaries and knowledge exchange

The second change that impacts how corpus linguists present their work involves the growing number of research intermediaries who communicate and produce corpus analysis for non-academic users. Think tanks, pressure groups and other civil society organisations like campaigning or advocacy charities are all examples of research intermediaries (Scott and Jabbar, 2014; and Smith *et al.*, 2013). Broadly speaking, when researchers and end-users share knowledge, sometimes through intermediaries, this can be called 'knowledge exchange' (Ward *et al.*, 2012).

Increasingly, university researchers are encouraged to engage in knowledge exchange as part of their work. Corpus linguists are no exception. In the United Kingdom, a particular kind of exchange activity was recently evaluated in the 2014 Research Excellence Framework (REF) under the rubric of 'impact'. The REF defined impact as 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (REF, 2011: 26). Although this concept is understood differently across several disciplines (Oancea, 2013), examining some of the impact case studies using corpus linguistics that were submitted to the 2014 REF gives a sense of how researchers worked with businesses and public services for the benefit of a wide range of users: language learners, 'at-risk' young people and teachers.[2]

Outside the REF, textual analysis increasingly features in public realms. For example, in early 2014 the Guardian newspaper produced an interactive chart comparing the number of mentions of migrant groups in five British national newspapers with actual immigration statistics.[3] The UK-based think-tank Demos also features a 'Centre for the Analysis of Social Media' that links quantitative textual analysis with ongoing political and policy debates. And the Leveson Report in the UK, produced after a major public investigation into British press practices, considered several pieces of linguistic research as evidence for inaccurate, biased and unfair reporting on immigration (Leveson, 2012: 670–1). They were introduced by representatives of ENGAGE and the Federation of Muslim Organisations, two examples of civil society and community-based groups. So, whether through intermediaries or more direct engagement with non-academic groups, people

[2] Full-text versions of all impact case studies submitted to the 2014 REF are available online at: http://impact.ref.ac.uk/CaseStudies.

[3] The chart is available online at: http://www.theguardian.com/news/reality-check/interactive/2014/jan/06/uk-migration-statistics-v-headlines.

working with text as data are continuing to take their findings outside the academy.

But knowledge exchange is not straightforward in practice: it is contingent on the contexts in which it eventually happens (Sebba, 2013). Several studies have systematically identified factors that limit and enable these processes. Some of the most important ones include the perceived credibility and validity of the knowledge itself, the timeliness of the research, and the clarity or accessibility of its presentation (Oliver *et al.*, 2014). For corpus linguists wanting to harness the earlier-mentioned advantages of Big Data, these factors present additional considerations. How can intermediaries or end-users hope to make sense of masses of data? How can academics communicate their findings in timely ways? And how can they instill confidence in the resulting analysis? Data visualisation can provide powerful answers to these kinds of questions. However, corpus linguists must extend their gaze beyond discussion of tools and techniques among fellow experts to critically consider how and why they are even visualising in the first place.

3. Visualisation and its uses within corpus linguistics

Data visualisation can be thought of as the representation or presentation of data to facilitate understanding (Kirk, 2016: 19). Charts, maps and graphs are all examples of visualisations that people might encounter in their daily lives. They can be static, meaning they are unchanging images, or interactive, meaning that users can modify or change aspects of the visualisation (Kirk, 2016). This definition suggests at least two main uses for visualisation: a mode of communicating analyses and a tool for analysis itself. Both have proliferated in spheres including journalism, business and social media.

Visualisation advances in corpus linguistics have applied existing tools and created new ones to deal with textual data. Some of these are built into existing software like the histogram function in Sketch Engine (Kilgarriff *et al.*, 2014) that shows how a given lexical feature is distributed across text types like genres or time periods. Others use network patterns to reveal collocation patterns among words (Di Cristofaro, 2013). Meanwhile, projects like WordWanderer (Dörk and Knight, 2015) and Compare Clouds (Diakopoulos *et al.*, 2015) harness the advantages of data visualisation by making analysis more interactive, exploratory and customisable. GraphColl (Brezina *et al.*, 2015) is a tool that allows users to visualise their corpora to find new and potentially unexpected collocational relationships, while also preserving a high degree of user control over key parameters like thresholds for statistical significance.[4] Culy and Lyding (2010) present

[4] GraphColl is freely available online at: http://www.extremetomato.com/projects/graphcoll/.

‘Double Tree’ as an innovative way of displaying concordances. Hilpert (2011) deals with the problem of how to show changes in language over time by using scatterplots placed in sequences – a feature that allows users to look at individual ‘slices’ of time as well as dynamically. And Rohrdantz *et al*. (2012) use Tableau to analyse how frequently versions of words ending in – *gate*, as a way of indicating scandals, appeared in English, French and German.

Many of these visualisation tools and techniques include a welcome aim of bringing corpus analysis to students, language learners and researchers from other disciplines who use corpora. And early reports are promising. But at this stage, where users and researchers are confronted with a host of tools and techniques, there is an urgent need within the field to reflect upon these developments. What can be said about how, why, and in what contexts researchers visualise corpora, as well as the wider implications for visualisation practice?

Several contributions from the informational visualisation literature deal with both the process of making text visualisations as well as the contexts in which visualisations are made. Textual data present specific challenges for linguists who use visualisation techniques for different purposes: exploration, explanation and statistical confirmation (Siirtola *et al*., 2014). This is an important distinction because it acknowledges that researchers need visualisations for a range of reasons. In another paper, Gough *et al*. (2014) advance the idea of NEUVis, or ‘visualisations for non-expert users’ to outline some valuable lessons for practitioners: ‘be aware of the impact of your chosen design on the reading of the data’, ‘define the intent of the visualisation’ and ‘consider the intended audience and their context’ (Gough *et al*., 2014: 175–6). These kinds of considerations matter because, as Hullman and Diakopoulos (2011) argue, visualisations can affect how users interpret an issue by prioritising some elements or data over others. Visualisations can persuade (Pandey *et al*., 2014), as well as give impressions of objectivity (Kennedy *et al*., 2016a). But there appears to be a lack of academic attention to this critical aspect of visualisation: ‘there is a need to think more systematically about how values and intentions shape visualization practice’ (Dörk *et al*., 2013: 2190).

Given that the composition and public communication of linguistic research is changing, there are important questions for linguists that go beyond the technical, ‘how-to’ of visualising textual analysis. Some of these questions have already been usefully picked up by critical scholars in diverse strands of social science and information visualisation (for a review, see Kennedy and Allen, 2017). However, there is still a need to see how these questions are actually handled in practice. Furthermore, many of the examples above take the perspective of visualisation as a tool for analysis that is used primarily by academics for other academics. What issues arise when large corpora are visualised by and for non-academic groups? What should researchers be aware of as they decide how to represent their textual data?

4. Data and methods

To address these questions, I draw upon two research projects conducted by the Migration Observatory (MigObs), an organisation that transmits academic research on migration to non-academic stakeholders including journalists, non-governmental organisations (NGOs), civil society organisations and Parliamentarians. Other academics, educators and students also use MigObs resources for research, teaching or self-study. These projects aim to analyse how the British national press has portrayed migration issues across many publications and several years, and then relate these to changes in how the public perceive immigration issues. They developed from a straightforward rationale: before making assertions about how or why the press should cover immigration in particular ways, it is necessary to understand as fully as possible what the press has actually said in the first place.

The following sections detail the justification and design of two linked case-study projects that will serve as data for the subsequent analysis. These examples illustrate key decisions made during the process of visualising a corpus from the perspective of intermediaries, with the aim of suggesting lessons for researchers in the humanities and social sciences as they think about visualising their analyses.

4.1 Two research case studies featuring corpus visualisation

These cases developed from an observation that, while a number of studies had shown how certain sub-sections of the UK national press covered international migration issues in a largely negative light, there was not a great deal of systematic or comprehensive evidence that could demonstrate these statements either over a longer period of time or across multiple sets of publications. A scoping study (Allen, 2012) found a notable exception of the RASIM project conducted by researchers at Lancaster University who used corpus linguistics to document portrayals of refugees, asylum seekers and immigrants in British newspapers (Gabrielatos and Baker, 2008). In response, MigObs has aimed to fill this gap through several projects. Two of these projects had significant visualisation components.

4.1.1 Project 1: Migration in the News, 2010–2012

Since MigObs did not have prior experience in conducting corpus analysis, the first task was to pilot the data collection, methods and analysis. This took the form of a study that was limited in scope to three years. It aimed to answer two questions: (*1*) what kinds of language have different sub-sections of the

Tabloids	Midmarkets	Broadsheets
The Sun, *The Sun on Sunday*	*The Express,* *The Sunday Express*	*The Times,* *The Sunday Times*
Daily Mirror, *Sunday Mirror*	*The Daily Mail,* *Mail on Sunday*	*The Guardian,* *The Observer*
Daily Star, *Daily Star Sunday*		*The Independent,* *Independent on Sunday*
The People		*The Daily Telegraph,* *The Sunday Telegraph*
		The Financial Times

Table 1: National UK publications included in Project 1.

UK national press used to describe immigrants, migrants, asylum seekers and refugees over the 2010–2012 period; and (*2*) how do these portrayals differ among subsections of the press (i.e., tabloids, midmarkets and broadsheets).[5]

The project aimed to collect, as far as possible, all items in all UK national newspapers that mentioned immigrants, migrants, asylum seekers or refugees from 1 January 2010 to 31 December 2012. This three-year period covered an important time in British politics and migration policy change, including a General Election. Nexis UK, a database service that archives many kinds of international periodicals, was queried using the following search string: (refugee! OR asylum! OR deport! OR immigr! OR emigr! OR migrant! OR illegal alien! OR illegal entry OR leave to remain NOT deportivo NOT deportment).[6]

All national UK publications that had continuously published over the three-year period were included in the search, and then divided into tabloids, midmarkets or broadsheets. The *News of the World* was not included in this study because it ceased publication in 2011. These divisions correspond to the 'popular, midmarket, quality' labels used by the Audit Bureau of Circulations (ABC), a national body for the media industry. Table 1 shows the twenty titles included in the study.

In addition, all sections of each publication were searched, including sports, arts and letters to the editors. This was done because readers may

[5] Although a presentation of the results is outside the scope of this paper which focusses on the visualisation component, they can be found in Blinder and Allen (2016).

[6] This string, replicating the work of Baker *et al.* (2007), captures variations of each of the four main groups while making two exclusions for 'Deportivo' – a football club – and 'deportment' which refers to etiquette.

encounter information about migration in many different forms and contexts: from traditional reporting about conflicts that generate flows of refugees, to mentions of athletes who may have migrated to Britain, to reviews of recent plays that feature asylum seekers as main characters. In total, the dataset contained 58,351 items comprising 43,966,872 words.

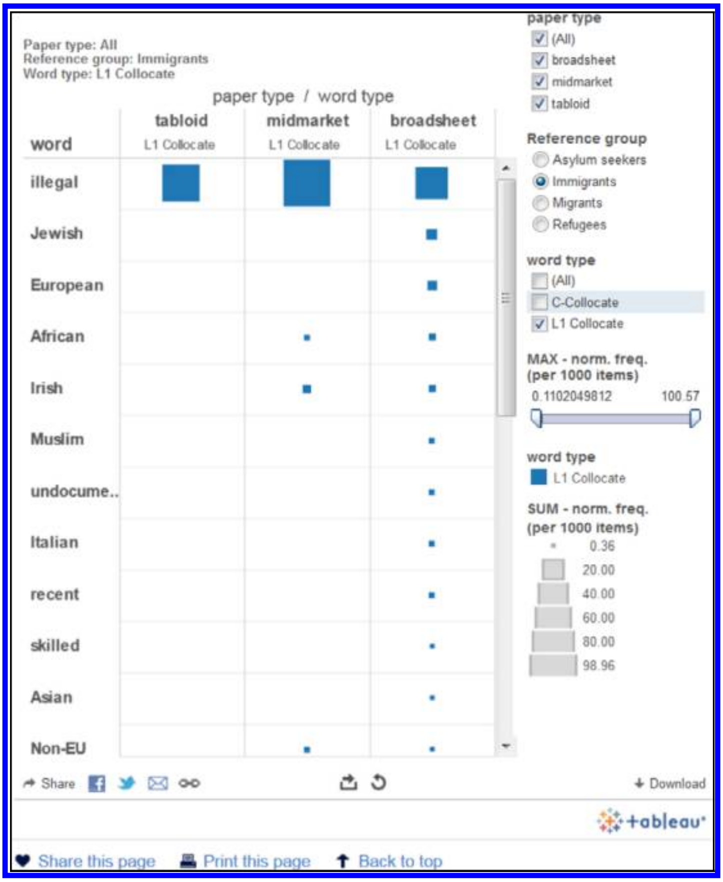
This study used collocational analysis based on a combination of Mutual Information and log-likelihood tests to identify the kinds of modifiers used to describe immigrants, migrants, asylum seekers and refugees. By focussing attention on those words that immediately appeared before mentions of each migrant group in the L1 position as well as within five words to the left or right, the study identified the main descriptors used to reference different migrant groups. Although this is not a perfect rule – some modifiers may appear outside the L1 position, as in 'an IMMIGRANT from BULGARIA' – it does reflect a tendency in written English to use adjectives immediately before the nouns they modify.

However, newspaper coverage can be episodic: some events may generate a great deal of unique coverage that does not reappear. Since the research aimed to examine how the press had consistently portrayed each migrant group, the collocational results were filtered to only show those results that were statistically significant in every annual sub-corpus. This is called a 'consistent collocate' or c-collocate (Baker *et al.*, 2007). So, in order to be reported as c-collocates of the word IMMIGRANT, candidate terms would have to be statistically linked to IMMIGRANT in 2010, 2011 *and* 2012. This additional criterion ensured that the study would only identify descriptors that were regularly associated with mentions of migrant groups over the entire 2010–2012 period. In short, the design of this project was orientated towards looking for consistency rather than differences over time.

Results from the collocation analysis were visualised using Tableau Public (2016). Tableau Public is a visualisation package that is popular in business analytics among other social science fields. It offers one kind of approach to visualisation, although Gatto (2015) explores other available options. Figure 1 shows a screenshot of the visualisation in its published form.[7]

The interface allows the user to display results along different dimensions: by publication type (tabloids, midmarkets and/or broadsheets), reference group (immigrants, migrants, asylum seekers or refugees), and by type of collocate (L1 only, or any c-collocate within five words to the right or left). Hovering over each square reveals both the raw and normalised (per 1,000 items) frequencies for each collocate. As illustrated by the size of each square, Figure 2 shows that the word ILLEGAL was the most frequent modifier of IMMIGRANT across all three publication types during 2010–2012.

[7] This visualisation is available online at: http://migrationobservatory.ox.ac.uk/resources/charts/migration-news-interactive-chart.



Paper type: All
Reference group: Immigrants
Word type: L1 Collocate

paper type
☑ (All)
☑ broadsheet
☑ midmarket
☑ tabloid

Reference group
○ Asylum seekers
● Immigrants
○ Migrants
○ Refugees

word type
☐ (All)
☐ C-Collocate
☑ L1 Collocate

MAX - norm. freq. (per 1000 items)
0.1102049812 100.57

Figure 1: Screenshot of Project 1 visualisation.

4.1.2 Project 2: An Expanded View of British Newspaper Coverage, 2006–2013

After completing that pilot study, MigObs wanted to expand its view of the British newsprint media to a wider time period. Accompanying this aim was a secondary objective of improving its understanding and use of data visualisation to share results with end-users. As a research intermediary, MigObs actively promotes its analysis and research among a host of non-academic users. For reasons outlined in Section 2.2, visualisation was an important mode of communication that demanded greater attention.

So, in Project 2, MigObs extended the previous corpus of newspaper items to include items published between January 2006 and December 2013. Derived from the same search string, this second corpus eventually contained about 90 million words from twenty-one titles – the additional publication being the News of the World. Also, a major difference was the manner in which the corpus was stored and analysed. This time, MigObs used Sketch

Engine (Kilgarriff *et al.*, 2014) to conduct more detailed frequency and collocational analysis similar in type to, but in greater depth than, the pilot study.

Then, as part of a larger project called Seeing Data,[8] MigObs visualised this corpus. The project team consisted of Helen Kennedy, Rosemary Lucy Hill, Andy Kirk and myself. We enlisted Clever Franke, a leading European design firm based in the Netherlands, to build a visualisation based on the textual dataset. The resulting interactive visualisation, which is still a work-in-progress as analysis of the expanded corpus continues, combined different kinds of charts to show four key analyses: word frequencies of *immigrant*, *migrant*, *asylum seeker* and *refugee*; collocations of each of these words; frequencies of people or organisations mentioned; and co-occurrences of people or groups in the items.[9] It was built using open-source software D3.js (Bostock *et al.*, 2011) along with jQuery, JavaScript, and HTML5 and CSS3 for styling. These are well-known tools for designing and building different kinds of visual outputs for online use, based on a variety of data types.

4.2 Values and objectives

These two projects that form the empirical basis of this article exist within a particular set of values and objectives. Before exploring the practical lessons learned throughout the process of creating these visualisations, it is necessary to set out briefly the guiding principles and aims that inform how MigObs generally makes decisions across all of its activities. Of course, this is not to suggest that these values are necessarily the best for all situations. Indeed, part of my argument – as well as those found in Hullman and Diakopoulos (2011) and Dörk *et al.* (2013) – is that there are potentially multiple values and objectives at play when researchers visualise their analyses. So, being clear about the context in which visualisations are made is essential as part of good visualisation practice (Kennedy and Allen, 2017).

MigObs holds five key organisational values. 'Authoritativeness' relates to having high standards of academic rigour and integrity. In terms of data, this involves stating what different data sources can and cannot say – being upfront about their strengths, weaknesses and provenance. 'Independence' means avoiding partisanship or pushing for achieving particular policy goals. This is especially critical on the polarised issue of immigration. 'Comprehensiveness' refers to a careful consideration of all aspects of immigration, the links among them, and the potential costs and

[8] See: www.seeingdata.org.

[9] This visualisation is available online at: www.compas.ox.ac.uk/migrationinthenews. Since MigObs is continuing to expand and develop the corpus for its research activities, this visualisation should be considered as a work-in-progress while the project is ongoing.

benefits associated with the full range of policy actions – not just a few pre-selected topics. The value of 'clarity' guides both the presentation of outputs as well as how analyses are written: they should be appealing to a wide range of users, from senior policymakers to journalists to students. Furthermore, this value demands identifying where gaps in evidence and data exist. Finally, 'engagement' involves enabling users to probe and explore the complexities of immigration in as self-directed a manner as possible. It also means introducing previously unfamiliar topics in accessible language.

These values guide MigObs as it seeks to achieve its three main objectives: first, to provide independent and evidence-based analysis of available migration data; second, to use these analyses to inform public debate as it happens in media and policy; and, third, to produce its own high-quality research on migration and policy issues. In Sections 5.1 to 5.3, the practical discussion draws upon observations of how these values and objectives informed decisions that occurred throughout the processes of making the two visualisations.

5. Lessons for corpus linguists

Visualisations play important roles as modes of communicating and facilitating analyses (Kirk, 2016). Within corpus linguistics, they make patterns in texts visible in ways that go beyond conventional techniques like concordances (Culy and Lyding, 2010). From the perspective of MigObs, visualisations are especially useful for two reasons. First, they enable users who may possess different expertise or skills to interact with an analysis and generate insights that are useful to them. Second, they can display key findings in ways that may be more accessible than conventional written outputs like reports or briefings. Since the point of the original study in Project 1 was to shed light on the actual language that the British national press had used to describe different migrant groups, and given organisational values such as engagement and clarity as discussed in Section 4.2, visualisations were a useful means of achieving that goal.

Studies such as Gough *et al*. (2014) suggest some guidance for good visualisation practice, notably to consider values, contexts and audiences. My aim here is to examine how these suggestions actually unfold in the course of making visualisations borne out of academic research. Drawing upon examples from the projects, the following sections identify three key lessons that illustrate how data visualisation is a contingent process shaped by specific contexts of values, communication among different expertises, and the chosen features for display. First, consider the aims and values of partners. Second, develop communication strategies that acknowledge different areas of expertise. And third, link visualisation choices with wider project objectives.

5.1 Consider values and aims before design begins

Visualisation involves a range of choices on the part of designers, researchers and intermediaries. The outcomes of these choices are partly determined by the rationale of whoever is making them. For what reason is the visualisation being created? Is a particular approach important – and if so, why?

Values and aims are important to bear in mind because intermediaries translate and repackage research outputs for particular purposes and towards certain ends. These ends simultaneously exist within social and political contexts: in polarised issue areas, such as debates about immigration in the UK (Threadgold, 2009), intermediaries are particularly prone to use research for strategic reasons (Boswell, 2009). Whether to inform policy or persuade key stakeholders, they exert influence over the information and data that flow through them.

In a similar way, visualisations perform different kinds of work. Kirk (2016) insightfully explains how they can either draw attention to particular findings or enable users to identify their own highlights. Equally, they can present information in ways that avoid making calls for partisan action, or omit potentially emotive and persuasive elements that might demand social or political change. Kennedy *et al.* (2016a) link these choices of presentation to an overarching sense of ‘objectivity’ or scientific soundness. So, a clear sense of why a visualisation should exist in the first place – an understanding of its broader aims and purposes for the intermediary or organisation making it – is a vital prerequisite for subsequent design decisions.

MigObs publicly espouses particular values, including comprehensiveness and independence. What these mean in practice is that analysis should come from as full a picture as possible, with all decisions and limitations accounted for. These values influenced many decisions from the research design to subsequent visualisation. Which publications should be included: only those that are perceived as particularly ‘interesting’ because they hold particularly strong left or right wing views? Which time periods should be examined: those which are popularly thought to feature immigration stories, such as around elections? How much explanatory guidance should accompany the visualisation – and where should it appear?

Given the limitations of previous studies into UK press portrayals of migrant groups, both Projects 1 and 2 aimed to be as comprehensive as possible in both research design and analysis. It would be wrong to assume that certain kinds of publications or time periods would produce more ‘interesting’ stories about migration before any data collection had even occurred. So, when corpus linguists consider possible ways of communicating their work through intermediaries, it is important to reflect upon why they wish to do so in the first place.

5.2 Develop clear communication among all team members who have different areas of expertise

Assessing these values and aims is one of several important steps. In a context where multiple team members contribute to the development and design of a visualisation, it is vital to ensure that all participants understand why these values are so important for the project. Also, designers, academic researchers and intermediary organisations possess different kinds of expertise and skills. While this may appear to be a source of strength, in practice it requires time and energy to ensure that diversity of perspectives does not lead to divergence of paths.

Three examples of collaboration among the Seeing Data research team, MigObs, and Clever Franke while developing the corpus visualisation illustrate this point well. MigObs hosted a half-day workshop with the designers and research team where each stakeholder group explained its goals and provided information about its audiences and working methods. This was a key moment where all participants could develop a sense, face to face, of one anothers' communication styles and personal motivations. Then, to display how certain words were collocated with each target word (*immigrant*, *migrant*, *asylum seeker* or *refugee*), the designers had to become familiar with the outputs produced by Sketch Engine. This required regular contact between them and MigObs which had already used the tool. And, throughout this collaboration, MigObs staff had to use their own expertise about immigration debates in the UK to disambiguate meanings. This was especially the case in identifying key issues, people or organisations that the design firm did not recognise because its Dutch staff were less familiar with British politicians or policies.

Although it seems self-evident to include time for mutual exchange and developing professional relationships, its role cannot be overstated in complex visualisation projects involving intermediaries and designers. In the literature about knowledge exchange, it is well-documented how social factors like trust and respect among intermediaries and researchers impact on the perceived quality of the subsequent exchanges (Oliver *et al.*, 2014). Likewise, if the process of visualisation is thought of as a particular kind of knowledge exchange, then it is clear that making concerted efforts to identify and harness different expertise is likely to enhance the final outcome.

5.3 Link choices of linguistic features with intended purposes and design options

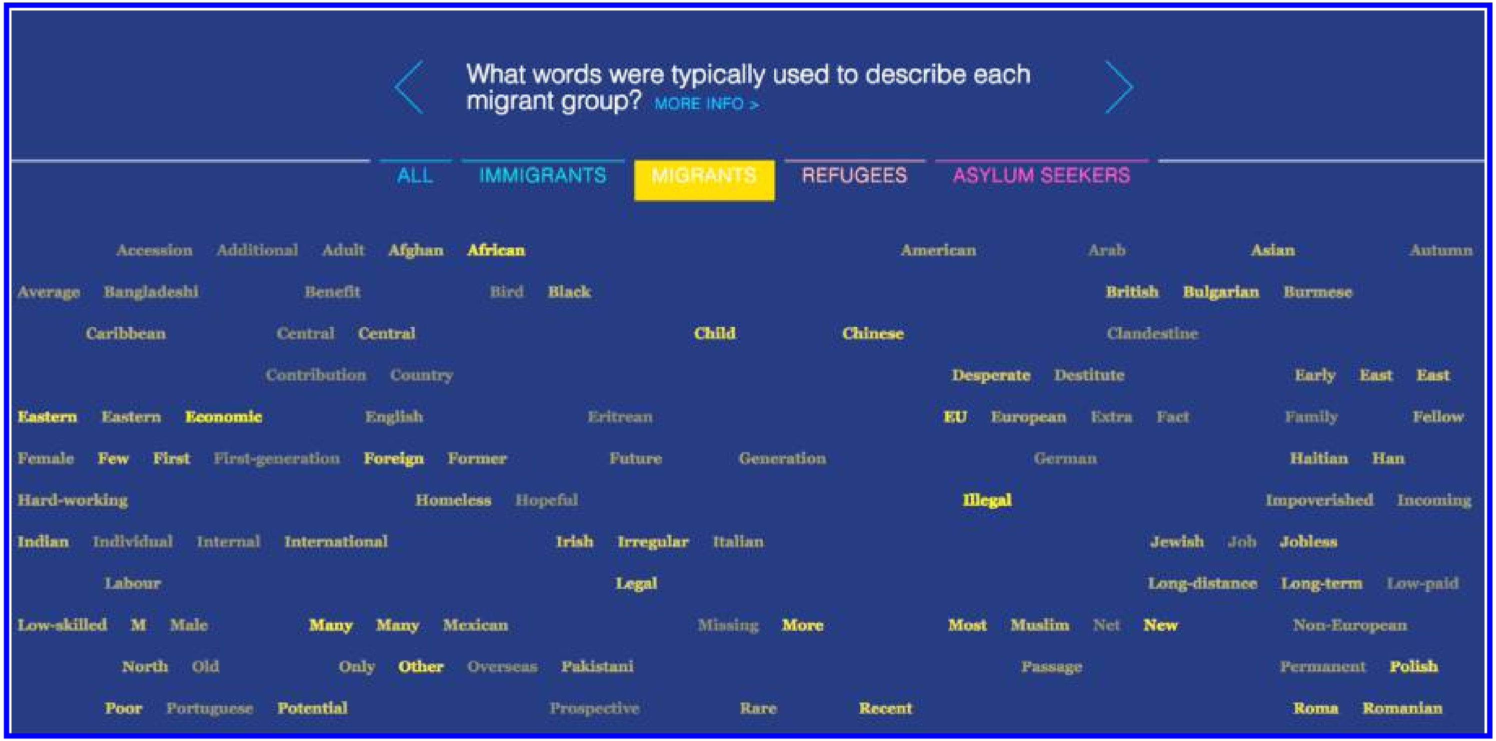
Even with effective communication channels set up, and all team members understanding the values and aims of the visualisation, the question of what to visualise needs careful consideration. There are a number of linguistic features that can be displayed graphically: collocations, frequencies, concordances (Wise *et al*., 1995). These can be differentiated by time period or types of sources. Researchers can also add annotations, such as extra information about the author and context in which the item was produced. Given this range of possibilities, how can researchers make sense of which ones to include?

Comparing the visualisations generated from Projects 1 and 2, as well as the decisions made while producing them, suggests some pointers. Specifically, three dimensions are worth discussing here: collocation, frequency and available depth of analysis. First, the outputs displayed collocational relationships and strengths in different ways. In the visualisation produced through Project 1 (Figure 1), larger squares represented stronger collocates in terms of their appearances per 1,000 words. Then, by hovering over each square, the user could read off precise frequencies. This was an important feature informed by the values of comprehensiveness, independence and engagement: users could explore their own curiosities without headline editorials, while also being able to see specific figures, too. But this feature required reading across the rows to identify which squares were associated with which collocate. The visualisation produced as part of Project 2 combined these steps by representing collocational strength as words themselves in more saturated colours, as shown in Figure 2. The degree of saturation was determined by Sketch Engine's built-in statistical measure, called 'salience'.[10] This was intended to draw attention to the words most strongly associated with each migrant group without the need for reading multiple rows.

Another difference involved the ordering of collocates. The first output only displayed collocates associated with a target word chosen by the user, then arranged them from strongest to weakest. Meanwhile, also as seen in Figure 2, the second output displayed up to 100 collocates associated with each of the four target words, then highlighted words according to salience depending on which target word the user selected. This created a shifting 'wall of words' effect that aimed to give a more instant sense of the sheer quantity of collocates.

Second, the visualisations handled word frequencies differently. The first output did not consider how mentions of each target word varied over time because its research design treated the entire 2010 to 2012 period as a unit: it was concerned with consistent portrayals, not variation. In contrast, the second output took advantage of its diachronic corpus and showed how mentions of each target group changed on a monthly basis. As shown in Figure 3, it also included annotations indicating when important events in British policy or politics occurred, particularly elections. These annotations were seen as important markers of context that different audiences – perhaps

[10] Details of how this score is calculated can be found in Rychlý (2008).



What words were typically used to describe each migrant group?

Accession Additional Adult Afghan African American Arab Asian Autumn

Average Bangladeshi Benefit Bird Black British Bulgarian Burmese

Caribbean Central Central Child Chinese Clandestine

Contribution Country Desperate Destitute Early East East

Eastern Eastern Economic English Eritrean EU European Extra Fact Family Fellow

Female Few First First-generation Foreign Former Future Generation German Haitian Han

Hard-working Homeless Hopeful Illegal Impoverished Incoming

Indian Individual Internal International Irish Irregular Italian Jewish Job Jobless

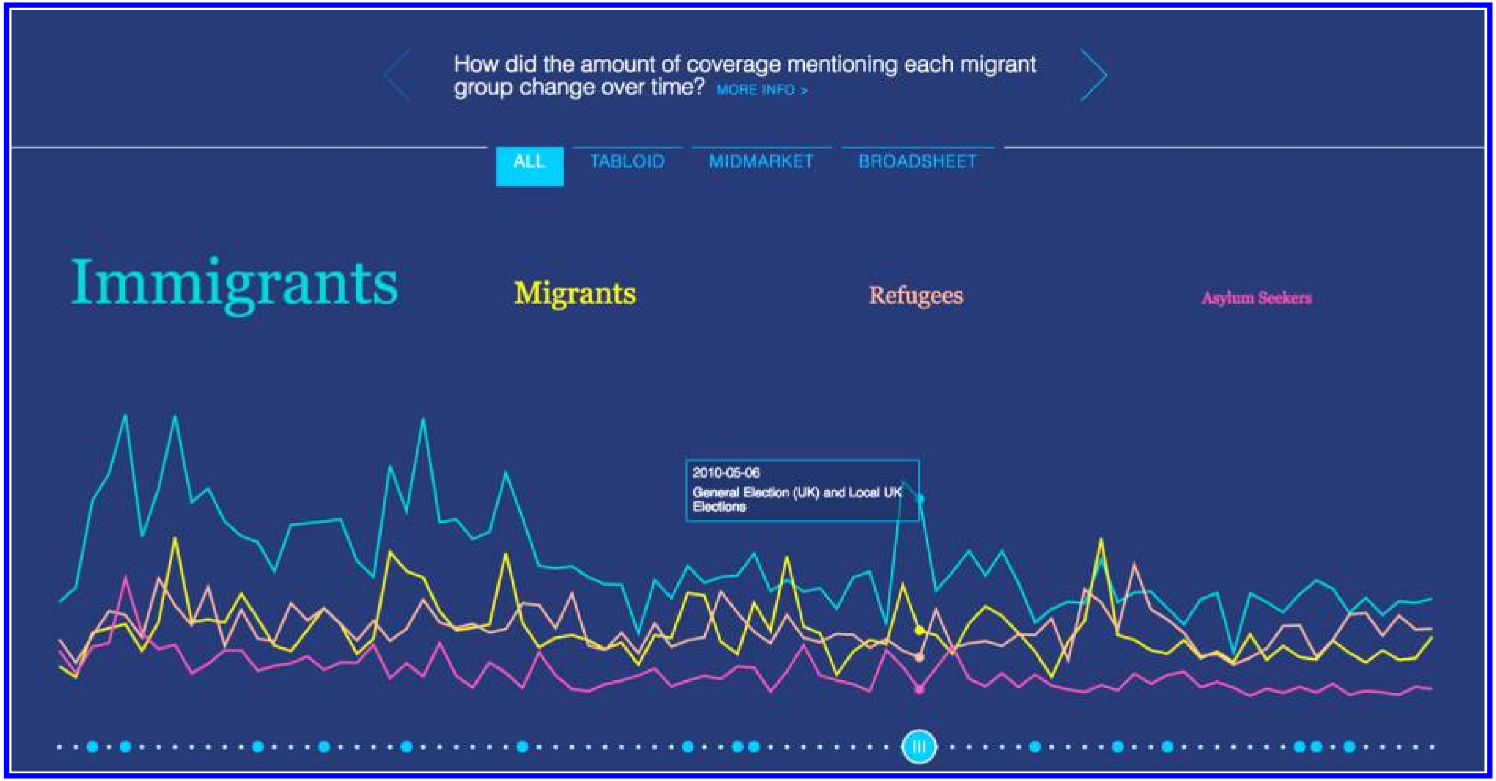
Labour Legal Long-distance Long-term Low-paid

Low-skilled M Male Many Many Mexican Missing More Most Muslim Net New Non-European

North Old Only Other Overseas Pakistani Passage Permanent Polish

Poor Portuguese Potential Prospective Rare Recent Roma Romanian

Figure 2: Screenshot of Project 2 collocational visualisation.



How did the amount of coverage mentioning each migrant group change over time?

2010-05-06
General Election (UK) and Local UK Elections

Figure 3: Screenshot of frequency analysis with annotation.

students, civil society organisations or members of the press – would be able to use as they navigated the dataset.

Third, the visualisations aimed to enable users to access different levels of analysis, but without necessarily prioritising one over the others. This constraint related to MigObs' organisational values of independence and engagement. In the first example, buttons along the right-hand side allowed users to toggle quickly among two kinds of collocational analysis, three publication types and four target words. Hovering over each square also revealed specific frequencies in both raw and normalised forms. These features were included for two reasons: an aim of being transparent with results for the benefit of users, and for consistency with other charts MigObs had already created with the same Tableau software. This was handled differently in the second visualisation. Although users could customise their view based on publication type or target word, the output did not include the ability to read off individual values. Rather, the intention was to give viewers a more immediate sense of how collocates related to one another through colour and size differences.

These comparisons suggest that choosing which linguistic features to visualise – and how – is connected to organisational values (Section 4.2), intended purposes (Section 5.1), as well as available design options presented by the tools at hand. The example from Project 1 that used Tableau Public had fewer annotations about the political context of the corpus, but allowed users to select, access and read greater details about the analysis. These decisions stemmed partly from the research questions: they aimed to show central tendencies rather than changes over time – and partly from the context in which the visualisation eventually existed, alongside a static research report that only showed particular aspects of the analysis. Also, the mix of available skills and capacities held by MigObs meant that Tableau Public presented an achievable solution.

In contrast, the bespoke visualisation produced by Clever Franke as part of the Seeing Data project was intended to show changes in linguistic features over time. It also aimed to take advantage of the greater range of skills available through the collaboration with design professionals by using more advanced software and libraries to express differences through colour and layout, and not just size and reading values. These decisions were made through iterative discussion involving the designers, researchers and MigObs staff. Therefore, the value of building time for communication (Section 5.2) remains important here, too.

6. Conclusion: visualisation and corpus linguistic expertise

In a recent Editorial in *Corpora*, McEnery (2015) reflected on the emergence of large-scale data mining and what it meant for the broader field of humanities research. He encouraged corpus linguists to 'find a voice and show what is distinctive – and good – about the interaction between data and linguistic expertise as opposed to simply data and the algorithm' (McEnery, 2015: 2). As the use of statistical techniques and tools becomes more common in linguistic research, they should aid rather than replace linguists' own expertise in making sense of real world language.

In a parallel way, I began from an observation that visualisation tools are also becoming more widely available to humanities and social science researchers. But availability is not enough on its own. Researchers also need to appreciate how visualisations, and the processes involved in making them, are situated in particular contexts of values, skills, objectives and issue areas (Kennedy *et al*., 2016b). As shown in Sections 2.1 and 2.2, parts of the university-based research world have changed dramatically – and are likely to continue to do so. Increasing demands for impact and knowledge exchange, and the presence of intermediaries to facilitate those processes, create new challenges. Visualisations offer some solutions to these challenges, but the field of corpus linguistics is only beginning to fully engage with them and understand their complexities.

My analysis reveals three key lessons: (*1*) consider values and aims before design begins; (*2*) develop clear communication among all team members who have different areas of expertise; and (*3*) link choices of linguistic features with intended purposes and design options. These lessons direct attention away from the technical 'how-to' of making visualisations, and towards the contexts in which these decisions are made. Such reorientation makes these lessons applicable in any field concerned with displaying analyses of textual data.

At a time when more data and tools for corpus analysis are available to researchers, and these same researchers are increasingly asked to engage with groups outside universities, it is appropriate to take time for reflection. If the process of making visualisations is located within sets of objectives, values, participants and skillsets that are contingent and shifting, then this suggests there is no single 'right' way to visualise corpora. Instead, researchers would be well-served by developing an awareness of how a range of factors feature in decisions about visualisation. Linguists are especially attuned to this kind of contextual thinking as they examine how people make their worlds through language. In the future, critical thinking about how researchers make their analyses visible presents another such opportunity.

Acknowledgments

Funding for visualising the first case study, 'Migration in the News, 2010–2012', was provided by the John Fell Fund (Oxford University Press). Funding for visualising the second case study that covered the 2006–2013 period was provided by the UK Arts and Humanities Research Council (AHRC) through the project 'Seeing Data: are good Big Data visualisations possible?' (grant number AH/L009986/1). Helen Kennedy, Rosemary Lucy

Hill and Andy Kirk collectively contributed to the visualisation research contained in the second case. Also, many thanks to Migration Observatory staff members who provided feedback on the visualisations during the research (Scott Blinder, Mikal Mast, Robert McNeil and Carlos Vargas-Silva).

References

Allen, W. 2012. UK Migration, Political Elites and Public Perceptions: Possibilities for Large-Corpus Textual Analysis of British Print Media Coverage. University of Oxford: COMPAS.

Anderson, C. 2008. ‘The end of theory’, Wired Magazine 16 (1). Available online at: https:// www.wired.com/2008/06/pb-theory/.

Andrejevic, M. 2014. ‘The big data divide’, International Journal of Communication 8, pp. 1673–89.

Baker, P., T. McEnery and C. Gabrielatos. 2007. ‘Using collocation analysis to reveal the construction of minority groups: the case of refugees, asylum seekers and immigrants in the UK Press’, Paper presented at the Corpus Linguistics 2007 Conference. 27–30 July 2007. University of Birmingham, UK.

Blinder, S. and W. Allen. 2016. ‘Constructing immigrants: portrayals of migrant groups in British national newspapers, 2010–2012’, International Migration Review 50 (1), pp. 3–40.

Bostock, M., V. Ogievetsky and J. Heer. 2011. ‘D3: data-driven documents’, IEEE Transactions on Visualization and Computer Graphics 17 (12), pp. 2301–9. Available online at: https://d3js.org/.

Boswell, C. 2009. The Political Uses of Expert Knowledge: Immigration Policy and Social Research. Cambridge: Cambridge University Press.

boyd, D. and K. Crawford. 2012. ‘Critical questions for big data’, Information, Communication and Society 15 (5), pp. 662–79.

Brezina, V., T. McEnery and S. Wattam. 2015. ‘Collocations in context’, International Journal of Corpus Linguistics 20 (2), pp. 139–73.

Couldry, N. and J. Turow. 2014. ‘Advertising, big data and the clearance of the public realm: marketers’ new approaches to the content subsidy’, International Journal of Communication 8, pp. 1710–26.

Culy, C. and V. Lyding. 2010. ‘Double tree: an advanced KWIC visualization for expert users’ in Information Visualisation (IV), 2010 14th International Conference, pp. 98–103. 26–29 July 2010. London: IEEE.

Diakopoulos, N., D. Elgesem, A. Salway, A. Zhang and K. Hofland. 2015. ‘Compare clouds: visualizing text corpora to compare media frames’ in proceedings of the IUI Workshop on Visual Text Analytics. Presented at the 4th Workshop on Visual Text Analytics. 29 March 2015. Atlanta, Georgia, USA.

Di Cristofaro, M. 2013. ‘Visualizing chunking and collocational networks: a graphical visualization of words’ networks’, paper presented at the 2013 Corpus Linguistics Conference, Lancaster University.

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

Dörk, M., P. Feng, C. Collins and S. Carpendale. 2013. ‘Critical InfoVis: exploring the politics of visualization’, Extended Abstracts on Human Factors in Computing Systems: Association for Computing Machinery, pp. 2189–98. New York: ACM.

Driscoll, K. and S. Walker. 2014. ‘Working within a black box: transparency in the collection and production of big Twitter data’, International Journal of Communication 8, pp. 1745–64.

Gabrielatos, C. and P. Baker. 2008. ‘Fleeing, sneaking, flooding a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005’, Journal of English Linguistics 36 (1), pp. 5–38.

Gatto, M.A.C. 2015. Making Research Useful: Current Challenges and Good Practices in Data Visualisation. University of Oxford: Reuters Institute for the Study of Journalism.

Gough, P., X. Ho, K. Dunn and T. Bednarz. 2014. ‘Art and chartjunk: a guide for NEUVis’ in Proceedings of the 7th International Symposium on Visual Information Communication and Interaction, pp. 171–7. 5–8 August 2014. Sydney, Australia: ACM.

Graham, M. 2011. ‘Time machines and virtual portals: the spatialities of the digital divide’, Progress in Development Studies 11 (3), pp. 211–27.

Gregory, I.N. and A. Hardie. 2011. ‘Visual GISting: bringing together corpus linguistics and geographical information systems’, Literary and Linguistic Computing 26 (3), pp. 297–314.

Hardaker, C. 2010. ‘Trolling in asynchronous computer-mediated communication: from user discussions to academic definitions’, Journal of Politeness Research 6 (2), pp. 215–42.

Hilpert, M. 2011. ‘Dynamic visualizations of language change: motion charts on the basis of bivariate and multivariate data from diachronic corpora’, International Journal of Corpus Linguistics 16 (4), pp. 435–61.

Hullman, J. and N. Diakopoulos. 2011. ‘Visualization rhetoric: framing effects in narrative visualization’, IEEE Transactions on Visualization and Computer Graphics 17 (12), pp. 2231–40.

Kennedy, H. and W. Allen. 2017. ‘Data visualisation as an emerging tool for online research’ in N. Fielding, R. Lee and G. Black (eds) The SAGE

Handbook of Online Research Methods, pp. 307–26. (Second edition.) London: SAGE Publications Ltd.

Kennedy, H. and R.L. Hill. 2017. 'The feeling of numbers: emotions in everyday engagements with data and their visualisation', Sociology. DOI: 0.1177/0038038516674675.

Kennedy, H., R.L. Hill, G. Aiello and W. Allen. 2016a. 'The work that visualisation conventions do', Information, Communication and Society 19 (6), pp. 715–35.

Kennedy, H., R.L. Hill, W. Allen and A. Kirk. 2016b. 'Engaging with (big) data visualizations: factors that affect engagement and resulting new definitions of effectiveness', First Monday 21 (11). DOI: 10.5210/fm.v21i11.6389.

Kilgarriff, A. and G. Grefenstette. 2003. 'Introduction to the special issue on the web as corpus', Computational Linguistics 29 (3), pp. 333–47.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý and V. Suchomel. 2014. 'The Sketch Engine: ten years on', Lexicography 1 (1), pp. 7–36.

Kirk A. 2016. Data Visualisation: A Handbook for Data Driven Design. London: SAGE.

Kitchin, R. 2014. 'Big data, new epistemologies and paradigm shifts', Big Data and Society 1 (1), pp. 1–12.

Knight, C. and C. Lyall. 2013. 'Knowledge brokers: the role of intermediaries in producing research impact', Evidence and Policy: A Journal of Research, Debate and Practice 9 (3), pp. 309–16.

Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. Available online at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Leveson, The Right Honourable Lord Justice. 2012. Culture, Practice and the Ethics of The Press. The Stationery Office, London.

McEnery, T. 2015. 'Editorial', Corpora 10 (1), pp. 1–3.

Oancea, A. 2013. 'Interpretations of research impact in seven disciplines', European Educational Research Journal 12 (2), pp. 242–50.

Oliver, K., S. Innvar, T. Lorenc, J. Woodman and J. Thomas. 2014. 'A systematic review of barriers to and facilitators of the use of evidence by policymakers', BMC Health Services Research 14 (2), pp. 1–12.

Pandey, A.V., A. Manivannan, O. Nov, M. Satterthwaite and E. Bertini. 2014. 'The persuasive power of data visualization', IEEE Transactions on Visualization and Computer Graphics 20 (12), pp. 2211–20.

REF. 2011. Assessment Framework and Guidance on Submissions. Available online at: http://www.ref.ac.uk/media/ref/content/pub/assess mentframeworkandguidanceonsubmissions/GOS20including 20addendum.pdf.

Rohrdantz, C., A. Niekler, A. Hautli, M. Butt and D.A. Keim. 2012. 'Lexical semantics and distribution of suffixes: a visual analysis' in Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH, pp. 7–15. Avignon, France: Association for Computational Linguistics.

Rychlý, P. 2008. 'A lexicographer-friendly association score' in Proceedings of Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University.

Scott, J. and H. Jabbar. 2014. 'The hub and the spokes: foundations, intermediary organizations, incentivist reforms, and the politics of research evidence', Educational Policy 28 (2), pp. 233–57.

Sebba, J. 2013. 'An exploratory review of the role of research mediators in social science', Evidence and Policy: A Journal of Research, Debate and Practice 9 (3), pp. 391–408.

Siirtola, H., T. Säily, T. Nevalainen and K.-J. Räihä. 2014. 'Text variation explorer: towards interactive visualization tools for corpus linguistics', International Journal of Corpus Linguistics 19 (3), pp. 417–29.

Smith, K.E., L. Kay and J. Torres. 2013. 'Think tanks as research mediators? Case studies from public health', Evidence and Policy: A Journal of Research, Debate and Practice 9 (3), pp. 371–90.

Tableau Public. 2016. Software Package. Available online at: https://public.tableau.com/s/.

Threadgold, T. 2009. 'The media and migration in the United Kingdom, 1999 to 2009', paper presented at the Public Opinion, Media Coverage, and Migration Conference. 6–8 May 2009. Bellagio, Italy.

Tseng, V. 2012. 'The uses of research in policy and practice', Social Policy Report 26 (2), pp. 3–16.

Viegas, F.B., M. Wattenberg and J. Feinberg. 2009. 'Participatory visualization with Wordle', IEEE Transactions on Visualization and Computer Graphics 15 (6), pp. 1137–44.

Ward, V., S. Smith, A. House and S. Hamer. 2012. 'Exploring knowledge exchange: a useful framework for practice and policy', Social Science and Medicine 74 (3), pp. 297–304.

Wise, J.A., J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur and V. Crow. 1995. 'Visualizing the non-visual: spatial analysis and interaction with information from text documents' in Proceedings of Information Visualization, 1995, pp. 51–8.