Applying Computational Linguistic and Text Analysis to Media Content about Migration

William L Allen, University of Oxford

Abstract: As larger and more varied datasets on socially and political relevant issues have become available to researchers, computational methods that aim to make sense of them have also proliferated. In response, social scientists need to critically reflect on and take stock of the opportunities and challenges presented by these shifts. This chapter draws upon on a series of projects that involved collecting, analysing, and communicating results from a large corpus of British newspaper texts about migration, asylum-seekers, and refugees spanning 1985-2015. By reporting the choices I made during these steps, as well as the contexts in which they happened, I aim to draw out both theoretical and practical lessons for researchers wanting to use computational approaches to media data involving political topics like migration—particularly, but not exclusively, from linguistic perspectives.

This is the author accepted manuscript of a chapter that has been published by Oxford University Press. The suggested citation is: Allen, William L. (2022). 'Applying computational linguistic and text analysis to media content about migration', in A. A. Salah, E. E. Korkmaz, and T. Bircan (eds.), *Data Science for Migration and Mobility*. Oxford: Proceedings of the British Academy, Oxford University Press, pp. 352-365.

Within the domain of migration studies, itself a multi- and inter-disciplinary field concerned with a broad array of questions spanning scales of geography and analysis (Gamlen et al. 2013; Pisarevskaya et al. 2020), researchers have regularly addressed issues surrounding how migrants and migration are represented in different settings such as media. Documenting these patterns of representations is important for several reasons. First, media analysis is a form of knowledge production that enables comparison with other epistemological viewpoints, such as migrants' own understandings and journalistic expressions (Shumow 2014). Second, these patterns serve as inputs for other phenomena of interest to social scientists, including immigration attitude formation (see Dinesen and Hjorth 2020 for a review), elections and voting behaviour (e.g., Hobolt, Leeper, and Tilley 2020), integration in host communities (e.g., Bos et al. 2016), policymaking (e.g., Allen and Blinder 2018), and migrant decision-making (e.g., Crawley and Hagen-Zanker 2019).[1] Third, analyses of media texts can serve as (part of) the evidence base which facilitates a range of interventions aiming to either inform audiences—itself a worthy goal (Schudson 2010)—or impact their behaviours such as taking a stand on issues and talking about politics among their networks (King, Schneer, and White 2017).

Thanks to advances in computational and digital methods, social scientists are increasingly able to collect and make sense of larger and more varied sets of texts, whether they involve digitized or digital-native sources. Meanwhile, communicating the results of these analyses in visual and graphical ways is becoming more important as a growing number of visualization tools has placed relatively high-quality outputs within the reach of non-specialist users. On the one hand, these developments have spurred tremendous uptake of these approaches to analyse textual data across a range of domains. On the other hand, the

[1] For a comprehensive review of major themes in recent scholarship on media and migration, see Allen, Blinder, and McNeil (2017).

rise in work claiming to use textual analysis has attracted a degree of scepticism, particularly on grounds of validity and reliability when measured against humans' performance (e.g., Brookes and McEnery 2018; van Atteveldt, van der Velden, and Boukes 2021). Therefore, a key problem confronting social scientists is how to think through the opportunities and limitations of computational text analysis methods as applied to their specific projects.

In this chapter, I aim to provide such a guide for researchers by drawing together several strands of my research into mass media portrayals of migrants and migration. While the examples from my own work mainly draw from the domains of newspapers and journalism, the lessons and observations about computational methods apply to other forms of digital data that researchers increasingly access. My approach has been motivated by the sentiment expressed by Tony McEnery, a prominent linguist who has sought to apply knowledge about how languages operate to social scientific questions. In short, he has urged linguists (as well as humanities and social science researchers using texts) to find what is "distinctive and good" (McEnery 2015, 2) about the interface of expertise and data, rather than abandon this effort in favour of purely algorithmic solutions that are divorced from either theory or practice. This echoes parallel trends in critical data studies examining how quantitative data—as well as the people, methods, and processes which generate them—both shape and are shaped by politics (Amoore and Piotukh 2016; Bigo, Isin, and Ruppert 2019)

I have organized this chapter around four key verbs as both a nod to the linguistic content and to walk through the discrete steps involved in doing textual analysis. The first section, 'orienting', sets out the main approaches to text analysis that I will focus on, including some reflection on their limitations. The second section, 'collecting', focuses on the ways that researchers gather and organise media data. Third, in the section 'analysing', I explain how I identify patterns in texts (e.g., frequencies of words) and how these link to substantive concepts of interest to social scientists (e.g., agendas). Finally, the fourth section

'communicating', explores different ways sharing both the outputs of, and procedures involved in analysis. This is an important step that is often overlooked yet is increasingly consequential for enhancing the impact of research—which is often publicly funded in the first place—and for taking an ethical stance towards Open Science principles that emphasise how researchers should be transparent about how they reached their conclusions (Lewis Jr. 2020).

Orienting: Approaches and limitations

In response to the shift towards quantitative methods across the social sciences, as well as broader interest in making sense of the growing volume and varieties of data now available (Mayer-Schönberger and Cukier 2013), researchers increasingly turn to solutions that rely on various computational tools. This is particularly true when it comes to texts and other forms of content which are digitised or digitisable. Given the range of possible methods—as well as the potential complexity of the task—getting to grips with this fast-growing field may seem overwhelming.[2]

My broader body of work draws upon two distinct yet related approaches to analysing large amounts of textual content. The first, typically called 'text as data', comprises statistical techniques that aim to identify patterns within texts while making minimal assumptions about how those texts were produced or how they may relate to one another (Laver, Benoit, and Garry 2003; Grimmer and Stewart 2013). The second, under the labels of corpus and computational linguistics, incorporates models of language use into modes of annotating and analysing texts (McEnery and Hardie 2011; Biber and Reppen 2015). This is related to natural language processing (NLP) which harnesses advances in computing power to

[2] Compare, for example, the methodological development with respect to analysing political texts in just 10 years between the advisory pieces of Quinn et al. (2010) and Barberá et al. (2020)—both of which are excellent guides to the field.

effectively organise and make sense of language appearing in 'the real world' through written or spoken forms (Jurafsky and Martin 2009). These techniques have clear benefits in terms of efficiency and reliability: algorithms, unlike humans, do not tire or become bored when faced with large numbers of repetitive tasks. Moreover, they offer ways by which researchers can test, share, and replicate their analyses—which are important dimensions of Open Science principles.

However, like all methods, they attract criticisms on several grounds. First, do these methods provide valid measures of patterns in texts that are of interest to social scientists? Empirical evaluations of how well computational approaches perform compared to human judgment suggest that there are significant gaps, particularly when categories derived from one set of texts are applied to a different set of texts which has different characteristics (Brookes and McEnery 2018). On this front, there is growing consensus around (and evidence for) the value of subjecting computational analyses to diverse forms of validation (Barberá et al. 2020; van Atteveldt, van der Velden, and Boukes 2021).

Second, since computational approaches to text analysis rely upon and prioritise the quantification of features including sentiment, topics, and populations, do they generate—rather than merely reflect—perspectives in ways that have political implications? This is connected to broader concerns about the assumptions made by researchers who are fixated on harnessing growing amounts of data without considering the impacts of their work (boyd and Crawford 2012). Specifically to modes of computational text analysis, there is a risk that these techniques produce outputs which outwardly give impressions of being objective, coherent, and comprehensive—features which may obscure how they necessarily involve choices and trade-offs that aim to simplify complex realities as manifested through texts (Marciniak 2016). This is especially important to consider when using these approaches to study texts about migration and refugee issues: high levels of abstraction may inhibit public

understanding of, and action towards these potentially vulnerable groups. Here, having a clear self-awareness of one's stance and objectives for doing research is crucial.

Third, can these techniques adequately account for the varied ways in which messengers create and express meaning beyond text? This is relevant for considering portrayals of migration in media that comprise both textual and visual modes (see Smets et al. 2020 for an overview). In various settings—from social media platforms (Hameleers et al. 2020) to websites about salient issues (Engebretsen 2020)—visual elements often accompany text. To be sure, parallel developments in computational image analysis now afford researchers greater flexibility and power in responsibly handling these forms of data (e.g., Byrne, Angus, and Wiles 2017). While this chapter does not focus on visual methods, it is worth highlighting how the realities of digital media present their own challenges to textual analysis.

Collecting: Media data and the politics of their origins

Having outlined some of the main contours of contemporary computational text analysis, I now turn attention to the next step of collecting textual data. This stage is often overlooked, yet it has important implications for the analysis which follows (though see chapters X and X in this volume for further discussion involving specific datasets). Where can researchers access media content? Within the context of digital research, two broad avenues are available: use established repositories—possibly third-party archival services—or collect it using webscraping techniques. Both approaches present opportunities and limitations.

A growing number of platforms and services make media data available to researchers who possess varying levels of computing proficiency.[3] These intermediaries

[3] Here, I use 'media' in its widest sense to include textual content as conveyed through a variety of communication modes—not just mass media.

provide valuable services for different audiences. Some of the most well-known, Factiva and LexisNexis, aim their services at businesses for market research purposes. One advantage of these kinds of services is that they generally have good coverage of international, national, and regional media outlets. Another relates to the ease with which users can search, filter, and download media texts (within some limits). Yet gains in accessibility and convenience come with losses in understanding precisely how and to what extent these texts are stored, labelled, and eventually delivered. For instance, ready-made categories and classification systems built into these platforms may reflect the broader interests in more generalized audiences rather than the needs of social scientists (e.g., Salah et al. 2012). What is more, the terms that generated these typologies may not be transparent or even published at all, preventing replication by other researchers or by the same researcher at a later point in time. Even the total number of documents in these archives—an important feature for making conclusions about the proportion of coverage dedicated to a topic such as migration—may not be available, as this could be considered proprietary information. Therefore, current guidance advises developing and using one's own search terms to identify relevant texts in these databases (Gabrielatos 2007; Barberá et al. 2020).

Given these limitations, collecting media data at their sources (known as 'scraping' data when it involves computational approaches) is appealing. Not only does this expand the universe of potential media beyond those sources which have already been collected, but also it promises researchers more control over the scope, organization, and reporting of the content. There are many off-the-shelf tools available for webscraping, developed by academics and for-profit companies alike: for example, the Digital Methods Initiative (n.d.) of the University of Amsterdam offers several scrapers for Google Images and other online sources. For those who possess more coding skills, the Python programming language remains the preferred way for quickly and effectively collecting textual data from online

sources. Using packages such as *BeautifulSoup* and *pandas*, which offer powerful ways of parsing and organizing webscraped data, researchers can assemble their own datasets and publish their methods for doing so.[4] This can be made even more effective by using Application Programming Interfaces (APIs) provided by major media sources, allowing access to large and growing portions of their databases for research purposes (e.g., Twitter 2021).

Yet as in the case of using existing repositories, these solutions also involve engaging with intermediaries, whether they are companies providing APIs or the tools themselves which are doing the webscraping. Moreover, it is not always clear which data are being scraped, and whether this will remain consistent in the future. For example, chapter X considers the specific affordances of Twitter's modes of API access, and how this has changed over time—with consequences for academic research. Some critical scholarship observes how this creates a situation where researchers are actually studying a "black box" (Driscoll and Walker 2014). Moreover, all of these platforms and code are part of digital infrastructures that generate their own sets of politics and privileges around knowledge creation and sharing (Bigo, Isin, and Ruppert 2019; Allen 2020). For example, comparative research shows how Google search queries in developing countries tend to show results from US or Europe-based sites rather than local sources, contributing to further digital inequalities (Ballatore, Graham, and Sen 2017). Finally, the growth of webscraping as a technique raises questions about researchers' fixation on 'freshness' and how this may impact what kinds of research questions receive priority (Marres and Weltevrede 2013). These aspects of digital data collection bear some similarities to curation: factors both internal and external to the research process, including algorithms, potentially impact the eventual contents of media datasets.

[4] For more practical guidance on these topics, see McKinney (2018) and Mitchell (2018).

My own research on media portrayals of migrants and migration has engaged with both avenues of data collection. One study into British press coverage relied on national newspaper data collected from NexisUK (Blinder and Allen 2016; Allen and Blinder 2018). In the process of assembling the dataset, it became apparent that NexisUK did not hold several years' worth of data from a particular publication. Yet this period *was* available using the Factiva service. Qualitative examination of the output from Factiva did not reveal any obvious systematic differences or gaps compared to what was otherwise available in the NexisUK database, so I merged the results. What is more, between stages of the study, an entire publication (*The Independent*) was removed from the NexisUK database when it ceased publishing printed copies in 2016. These examples demonstrate how commercial archival services present challenges for data collection.

In other projects, I have turned to variations of webscraping to collect media data from online sources. For example, to study the dominant features of migration data visualizations (Allen 2021), I used the Google Image Scraper (Digital Methods Initiative, n.d.) to identify and download sets of highly-ranked results for subsequent content analysis. Although the tool allows a high degree of control over different search parameters, it nevertheless reflects what Google's algorithm reports. To check the sensitivity of the tool to different search environments, I used a virtual private network (VPN) to simulate what users in several countries would see, while also deleting browser histories between searches. Fortunately, most of the results were consistent across searches. The broader point of this example is to highlight how, despite the potential convenience and comprehensiveness that computational tools offer for social science, they do not absolve researchers from asking critical questions about the scope, provenance, and veracity of the data at hand.[5]

[5] This point is equally applicable to those who use more conventional quantitative datasets such as observational surveys (Gray et al. 2015).

Analysing: Linking textual patterns and mental schemas

Having collected media data, how can social scientists analyse that data to generate meaningful insights? The key word here, of course, is 'meaningful'. One the one hand, meaning might arise or emerge from patterns identified in textual data. For example, topics comprising clusters of related words might be a form of meaning that could tell researchers something about a set of documents as demonstrated through unsupervised learning methods (see Barberá et al. 2020). On the other hand, meaning might come from a theory-driven view of which characteristics or aspects of texts are more relevant or important: in this mode, a researcher might be looking for evidence of a mechanism or property. Both approaches display their own susceptibilities. Results from 'data-driven' modes can be sensitive to parameters and settings which are set by researchers, opening concerns about the risk of generating and interpreting spurious relationships. Meanwhile, theory-driven modes of enquiry may overlook newer concepts that are not well-captured by existing models or categories but nevertheless are significant.

In my own practice, I have mainly used text analysis as a means of generating results that serve as evidence of theoretical mechanisms or outcomes. For example, a long tradition of political communication scholarship has developed the concepts of 'agendas' and 'media agenda-setting' (McCombs and Shaw 1972; Boydstun 2013) as ways of linking changes in media content with subsequent changes in public attitudes. Broadly, media can make some issues seem more important by mentioning them more frequently ('first-level agenda-setting') or by linking them with other issues that are also perceived to be important ('second-level agenda-setting'). These second-level agendas are sometimes seen as attributes of issues (Soroka 2002): for example, media might link the issue of asylum-seekers with attributes relating to security and terrorism rather than economic development. Therefore, security

might be considered a second-level agenda of asylum that makes people think the issue is more important. If found, this result would be evidence of an agenda-setting effect.

But how does this precisely happen via media content? Here is where I have used the theory of 'lexical priming' proposed by the linguist Michael Hoey (2005). Essentially, lexical priming proposes a way by which some words are more strongly associated with a target word such as 'immigrant' or 'refugee'. When considering the word 'pitch', for example, you might think of related words like 'football' (if you are inclined toward sports) or possibly 'music' if you have an arts background. These words, called 'primes', become related to target words through repeated use or familiarity: if certain primes are more readily available and top-of-mind than others, they will become more strongly related to the target word. Of course, the strength of these relations may change through use.

Linking this theorised mechanism with the concept of a second-level agenda led me to a particular kind of textual feature: collocation with key migration terms. The best candidates for strong primes of 'immigration' or 'refugees' would likely be those words which were frequently linked with those target words in the dataset. In this case, those links took the form of a grammatical association: adjectives referring to nouns. By noting the most highly-frequent adjectives linked with nouns of interest (e.g., *asylum-seeker*, *refugee*, *immigrant*), I began generating second-level agendas associated with migration. Clustering these terms using prior qualitative work on similar topics in the British press (Baker, Gabrielatos, and McEnery 2013) revealed six sets of issue attributes that I later interpreted as representing second-level agendas of migration in British newspapers: the economy; legal status; policymaking; sociocultural dimensions; geographic origins; and the scale and pace of immigration.

To be clear, there are multiple ways of identifying and measuring agendas in media texts, some of which can include computationally inducing topics or patterns from content.

Moreover, I could have also used these techniques at later stage of my analysis to sort and make sense of the lists of collocates for each target word. My point, however, is to demonstrate how the analytical choices that researchers must make about handling text can be driven by theoretical goals as well as technical concerns about efficiency. In this case, my goal was to measure a theoretically useful concept (i.e., second-level agendas) using a computational technique (i.e., collocational analysis deriving from grammatical rules). Whichever approaches social scientists use, they should be mindful of the needs to clearly justify their choices and make explicit links to their research questions or objectives.

Communicating: Sharing results and decision-making processes for public impact and replicable science

Analysing texts, while obviously important to do correctly, is not the end point. Rather, communicating these results to different audiences—whether they are other scholars, members of the interested public, or decision-makers in policy and civil society domains—is a crucial step that is closely linked to analysis. This is becoming even more true as researchers are increasingly expected to demonstrate how their work has public relevance or impact, especially when that work is funded by public sources.

Scholars using computational text analysis need to be attuned to this question of how to effectively communicate their research for at least two reasons. First, specific to the domain of media analysis, large-scale text analysis offers valuable avenues for making sense of these communication forms in ways that can inform decisions or change public perceptions. Those shifts, in turn, might have profound impacts on the lives of migrants. For example, in a report that analysed British media content on migration, Heaven Crawley and her colleagues (2016) showed how migrant voices tend to be missing from mainstream reporting. This presents implications for journalism practice: it is possible that greater

migrant representation in news, especially through the means of direct quotation, may change public perceptions. Second, speaking to other researchers, textual datasets and methodologies are important artifacts of the scientific process that potentially enable future replication and reuse, in the tradition of 'Open Science' principles. Therefore, clear documentation that explains the steps taken to produce a given output is vital.

In my own work, I have experimented with visualizing textual analyses for and among public users. Visualization, or the visual representation of data to enhance understanding (Kirk 2019), is a potentially powerful way of communicating quantitative information. The kinds of quantities and patterns that computational analyses of text produce—frequencies, clusters, changes over time or among subsets of documents—lend themselves to visual representation. Yet there are multiple choices involved in creating effective visualizations, which in turn involve several sets of 'hands' through which the analysis runs (Allen 2018). For example, there are potentially several ways of showing the strength of a relationship between a target word and its primes: intensity of colour, widths of connecting lines, or varying sizes of symbols (Allen 2017). Each approach presents advantages and limitations depending on the intended purpose and audience. The point here is not to recommend any particular approach, but rather to draw attention to the need for social scientists using computational approaches to also consider how they communicate their results—and whether these are the most appropriate for their objectives and key stakeholders.

Conclusion: Towards an agenda for social scientific—and socially responsible—text analysis

In this chapter, I have outlined what I think are four key steps in doing computational text analysis: orienting oneself in the field, collecting textual materials, analysing those texts with respect to given research goals (such as testing theories), and communicating the results in

effective ways. Here, it is important to note that tools and techniques, while important in their own rights, are parts of the wider process involved in using computational methods. Other chapters in this volume provide more detailed examples of specific datasets and methods for identifying, assembling, and analysing data at scale. Instead, my goal has been to drawn attention to the ways that these aspects are linked to broader questions of research design.

Indeed, computational approaches to text analysis like the ones I have illustrated through my own work on media representations of migrants and migration are neither value neutral nor divorced from deeply political questions involving human mobility, data, or media (Hovy and Spruit 2016). Rather, they raise important ethical and normative questions about *how* and *for what purposes* social scientists use these methods to say things about the world. For example, the practice of media monitoring—particularly for the purposes of gathering data on and about flows of asylum-seekers—has been restricted by the European Union on several grounds including potential misuse, privacy concerns, and the questionable veracity of collected data (European Data Protection Supervisor 2019). Meanwhile, there is a growing body of work showing how algorithms and computational methods can reproduce their creators' own prejudices (Leese 2014; Amoore and Piotukh 2016; Koenecke et al. 2020). This is not merely a theoretical problem: as more aspects of everyday life become quantified and algorithimically informed, the risk of reinforcing latent or unnoticed biases looms large (see Escalante et al. 2020 for an example involving job hiring procedures). More broadly, the large-scale processing of many types of data, including text, enable and disenable different forms of data politics (boyd and Crawford 2012; Bigo, Isin, and Ruppert 2019). In the context of mobility, answers to these questions have real-world consequences for the lives of migrants.

Given this reality, what would a socially responsible computational approach to media and text analysis look like? Recognizing how notions of 'responsibility' are

themselves context- and time-specific, I cautiously advance a few suggestions. First, it would be attuned to issues of power and inequality by acknowledging how researchers' choices throughout the lifecycle of designing, implementing, and communicating textual analyses are the products of particular contexts and characteristics. Second, it would consider how the research would be used—or could be used—for different objectives, and whether these objectives correspond with researchers' ethical stances. Third, it would adhere to Open Science principles of transparency and replicability, as far as the data and materials allow (see Lewis Jr. 2020 in the field of communication studies). Yet, building on the first two points, it would also be mindful of how the implementation of these principles relate to broader societal values and objectives—such as justice and public benefit—that are particularly salient when working with marginalised populations (Fox et al. 2021). Together, these aspects represent what I think are an emerging agenda for socially responsible computational text analysis. Developing and revisiting such an agenda will undoubtedly remain important as greater volumes and forms of data continue to come within the grasp of researchers, while issues such as migration and the movement of people remain visible in the collective consciousness of the public, policymakers, and media.

Works Cited

Allen, William. 2017. ‘Making Corpus Data Visible: Visualising Text With Research Intermediaries’. *Corpora* 12 (3): 459–82. https://doi.org/10.3366/cor.2017.0128.

———. 2018. ‘Visual Brokerage: Communicating Data and Research through Visualisation’. *Public Understanding of Science* 27 (8): 906–22. https://doi.org/10.1177/0963662518756853.

———. 2020. ‘Mobility, Media, and Data Politics’. In *The SAGE Handbook of Media and Migration*, edited by Kevin Smets, Koen Leurs, Myria Georgiou, Saskia Witteborn, and Radhika Gajjala, 180–91. London: SAGE Publications Ltd.

———. 2021. ‘The Conventions and Politics of Migration Data Visualizations’. *New Media & Society*, June, 1–22. https://doi.org/10.1177/14614448211019300.

Allen, William, and Scott Blinder. 2018. ‘Media Independence through Routine Press-State Relations: Immigration and Government Statistics in the British Press’. *The International Journal of Press/Politics* 23 (2): 202–26. https://doi.org/10.1177/1940161218771897.

Allen, William, Scott Blinder, and Robert McNeil. 2017. ‘Media Reporting of Migrants and Migration’. World Migration Report 2018. Geneva: International Organization for Migration. https://publications.iom.int/system/files/pdf/wmr_2018_en_chapter8.pdf.

Amoore, Louise, and Volha Piotukh. 2016. *Algorithmic Life: Calculative Devices in the Age of Big Data*. Abingdon, Oxon: Routledge.

Atteveldt, Wouter van, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. ‘The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms’. *Communication Methods and Measures*, January, 1–20. https://doi.org/10.1080/19312458.2020.1869198.

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. 'Sketching Muslims: A Corpus Driven Analysis of Representations around the Word 'Muslim'in the British Press 1998–2009'. *Applied Linguistics* 34 (3): 255–78.

Ballatore, Andrea, Mark Graham, and Shilad Sen. 2017. 'Digital Hegemonies: The Localness of Search Engine Results'. *Annals of the American Association of Geographers* 107 (5): 1194–1215. https://doi.org/10.1080/24694452.2017.1308240.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2020. 'Automated Text Classification of News Articles: A Practical Guide'. *Political Analysis*, 1–24. https://doi.org/10.1017/pan.2020.8.

Biber, Douglas, and Randi Reppen, eds. 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Bigo, Didier, Engin Isin, and Evelyn Ruppert, eds. 2019. *Data Politics: Worlds, Subjects, Rights*. London and New York: Routledge.

Blinder, Scott, and William Allen. 2016. 'Constructing Immigrants: Portrayals of Migrant Groups in British National Newspapers, 2010-2012'. *International Migration Review* 50 (1): 3–40. https://doi.org/10.1111/imre.12206.

Bos, Linda, Sophie Lecheler, Moniek Mewafi, and Rens Vliegenthart. 2016. 'It's the Frame That Matters: Immigrant Integration and Media Framing Effects in the Netherlands'. *International Journal of Intercultural Relations* 55 (November): 97–108. https://doi.org/10.1016/j.ijintrel.2016.10.002.

boyd, danah, and Kate Crawford. 2012. 'Critical Questions for Big Data'. *Information, Communication & Society* 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Boydstun, Amber. 2013. *Making the News: Politics, the Media and Agenda Setting*. Chicago: University of Chicago Press.

Brookes, Gavin, and Tony McEnery. 2018. 'The Utility of Topic Modelling for Discourse Studies: A Critical Evaluation'. *Discourse Studies* 21 (1): 3–21. https://doi.org/10.1177/1461445618814032.

Byrne, Lydia, Daniel Angus, and Janet Wiles. 2017. 'Figurative Frames: A Critical Vocabulary for Images in Information Visualization'. *Information Visualization* 18 (1): 45–67. https://doi.org/10.1177/1473871617724212.

Crawley, Heaven, and Jessica Hagen-Zanker. 2019. 'Deciding Where to Go: Policies, People and Perceptions Shaping Destination Preferences'. *International Migration* 57 (1): 20–35. https://doi.org/10.1111/imig.12537.

Crawley, Heaven, Simon McMahon, and Katharine Jones. 2016. 'Victims and Villains: Migrant Voices in the British Media'. Coventry University: Centre for Trust, Peace and Social Relations. http://www.migrantsrights.org.uk/files/news/Victims_and_Villains_Digital.pdf.

Digital Methods Initiative. n.d. *Google Image Scraper*. University of Amsterdam. https://tools.digitalmethods.net/beta/googleImages/.

Dinesen, Peter Thisted, and Frederik Hjorth. 2020. 'Attitudes towards Immigration: Theories, Settings, and Approaches'. In *The Oxford Handbook of Behavioral Political Science*, edited by Alex Mintz and Lesley Terris, 1–30. Oxford: Oxford University Press. 10.1093/oxfordhb/9780190634131.013.26.

Driscoll, Kevin, and Shawn Walker. 2014. 'Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data'. *International Journal of Communication* 8 (June): 1745–64.

Engebretsen, Martin. 2020. 'From Decoding a Graph to Processing a Multimodal Message: Interacting with Data Visualisation in the News Media'. *Nordicom Review* 41 (1): 33–50.

Escalante, Hugo Jair, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, et al. 2020. 'Modeling, Recognizing, and Explaining Apparent Personality from Videos'. *IEEE Transactions on Affective Computing*, 1–18. https://doi.org/10.1109/TAFFC.2020.2973984.

European Data Protection Supervisor. 2019. 'Social Media Monitoring Reports'. 2019. https://edps.europa.eu/sites/default/files/publication/19-11-12_reply_easo_ssm_final_reply_en.pdf.

Fox, Jesse, Katy E Pearce, Adrienne L Massanari, Julius Matthew Riles, Łukasz Szulc, Yerina S Ranjit, Filippo Trevisan, et al. 2021. 'Open Science, Closed Doors? Countering Marginalization through an Agenda for Ethical, Inclusive Research in Communication'. *Journal of Communication*, no. jqab029 (August). https://doi.org/10.1093/joc/jqab029.

Gabrielatos, Costas. 2007. 'Selecting Query Terms to Build a Specialised Corpus from a Restricted-Access Database'. *International Computer Archive of Modern and Medieval English (ICAME) Journal* 31: 5–44.

Gamlen, Alan, Alexander Betts, Alexandra Délano, Thomas Lacroix, Emanuela Paoletti, Nando Sigona, and Carlos Vargas-Silva. 2013. 'Faultlines and Contact Zones: A New Forum for Migration Studies'. *Migration Studies* 1 (1): 1–3.

Gray, Emily, Will Jennings, Stephen Farrall, and Colin Hay. 2015. 'Small Big Data: Using Multiple Data-Sets to Explore Unfolding Social and Economic Change'. *Big Data & Society* 2 (1). https://doi.org/10.1177/2053951715589418.

Grimmer, Justin, and Brandon M Stewart. 2013. 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21 (3): 267–97.

Hameleers, Michael, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. 2020. 'A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media'. *Political Communication*, February, 1–21. https://doi.org/10.1080/10584609.2019.1674979.

Hobolt, Sara B., Thomas J. Leeper, and James Tilley. 2020. 'Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum'. *British Journal of Political Science*, 1–18. https://doi.org/10.1017/S0007123420000125.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. Psychology Press.

Hovy, Dirk, and Shannon L. Spruit. 2016. 'The Social Impact of Natural Language Processing'. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 591–98. Berlin, Germany. https://www.aclweb.org/anthology/P16-2096.pdf.

Jurafsky, Daniel, and James Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech*. Upper Saddle River, NJ: Prentice Hall.

King, Gary, Benjamin Schneer, and Ariel White. 2017. 'How the News Media Activate Public Expression and Influence National Agendas'. *Science* 358 (6364): 776. https://doi.org/10.1126/science.aao1100.

Kirk, Andy. 2019. *Data Visualisation: A Handbook for Data Driven Design*. 2nd ed. London: SAGE.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. 'Racial Disparities in Automated Speech Recognition'. *Proceedings of the National Academy of Sciences* 117 (14): 7684. https://doi.org/10.1073/pnas.1915768117.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. 'Extracting Policy Positions from Political Texts Using Words as Data'. *American Political Science Review* 97 (2): 311–31. https://doi.org/10.2307/3118211.

Leese, Matthias. 2014. 'The New Profiling: Algorithms, Black Boxes, and the Failure of Anti-Discriminatory Safeguards in the European Union'. *Security Dialogue* 45 (5): 494–511. https://doi.org/10.1177/0967010614544204.

Lewis Jr., Neil A. 2020. 'Open Communication Science: A Primer on Why and Some Recommendations for How'. *Communication Methods and Measures* 14 (2): 71–82. https://doi.org/10.1080/19312458.2019.1685660.

Marciniak, Daniel. 2016. 'Computational Text Analysis: Thoughts on the Contingencies of an Evolving Method'. *Big Data & Society* 3 (2). https://doi.org/10.1177/2053951716670190.

Marres, Noortje, and Esther Weltevrede. 2013. 'Scraping the Social?' *Journal of Cultural Economy* 6 (3): 313–35. https://doi.org/10.1080/17530350.2013.772070.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

McCombs, Maxwell E., and Donald L. Shaw. 1972. 'The Agenda-Setting Function of Mass Media'. *The Public Opinion Quarterly* 36 (2): 176–87.

McEnery, Tony. 2015. 'Editorial'. *Corpora* 10 (1): 1–3.

McEnery, Tony, and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McKinney, Wes. 2018. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd ed. Boston: O'Reilly Media.

Mitchell, Ryan. 2018. *Web Scraping with Python: Collecting More Data from the Modern Web*. 2nd ed. Boston: O'Reilly Media.

Pisarevskaya, Asya, Nathan Levy, Peter Scholten, and Joost Jansen. 2020. ‘Mapping Migration Studies: An Empirical Analysis of the Coming of Age of a Research Field’. *Migration Studies* 8 (3): 455–81. https://doi.org/10.1093/migration/mnz031.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. ‘How to Analyze Political Attention with Minimal Assumptions and Costs’. *American Journal of Political Science* 54 (1): 209–28. https://doi.org/10.1111/j.1540-5907.2009.00427.x.

Salah, Almila Akdag, Cheng Gao, Krzysztof Suchecki, and Andrea Scharnhorst. 2012. ‘Need to Categorize: A Comparative Look at the Categories of Universal Decimal Classification System and Wikipedia’. *Leonardo* 45 (1): 84–85. https://doi.org/10.1162/LEON_a_00344.

Schudson, Michael. 2010. ‘Political Observatories, Databases & News in the Emerging Ecology of Public Information’. *Daedalus* 139 (2): 100–109. https://doi.org/10.1162/daed.2010.139.2.100.

Shumow, Moses. 2014. ‘Media Production in a Transnational Setting: Three Models of Immigrant Journalism’. *Journalism* 15 (8): 1076–93.

Smets, Kevin, Koen Leurs, Myria Georgiou, Saskia Witteborn, and Radhika Gajjala, eds. 2020. *The SAGE Handbook of Media and Migration*. London: SAGE Publications Ltd.

Soroka, Stuart. 2002. ‘Issue Attributes and Agenda-Setting by Media, the Public, and Policymakers in Canada’. *International Journal of Public Opinion Research* 14 (3): 264–85. https://doi.org/10.1093/ijpor/14.3.264.

Twitter. 2021. ‘Introducing the New Academic Research Product Track’. 26 January 2021. https://twittercommunity.com/t/introducing-the-new-academic-research-product-track/148632.