Criminality or Sheer Numbers? Attribute Agenda-Setting of Immigration and Asylum in British Newspapers, 2001-15

William L Allen, University of Oxford
Department of Politics and International Relations (DPIR)
Centre on Migration, Policy, and Society (COMPAS)
william.allen@compas.ox.ac.uk

Prepared for the 2017 LSE Applied Quantitative Text Analysis Conference (19 May, London)

ABSTRACT

How does the press cover certain issues, and how does this change over time? Answering these questions is essential for agenda-setting research that assumes amounts of media attention influence levels of public concern. Also, associating particular characteristics or properties (called 'attributes') to an issue can make it more relevant or accessible. This paper combines techniques from corpus and computational linguistics—notably grammatical part-of-speech (POS) tagging and collocation analysis at the levels of words and phrases—to identify and measure 'second level attribute agenda-setting' (McCombs 2014). Examining over 200,000 articles mentioning migration- and asylum-related terms in nine UK newspapers between 2001-2015 reveals how the scale and pace of immigration has recently risen in visibility, especially among tabloids. Also, British tabloids have consistently emphasised criminality more than broadsheets, whether referring to immigration or asylum issues. The paper concludes by suggesting how these techniques can complement current developments in quantitative text analysis.

KEY WORDS: agenda setting, immigration, linguistics, media, text analysis, UK

ACKNOWLEDGEMENTS: This research was funded by the Economic and Social Research Council (UK) and The Migration Observatory (University of Oxford, UK) as part of its ongoing 'Migration in the Media' project. The author would like to thank Paul Baker, Scott Blinder, Tony McEnery, James Tilley, and attendees at both MPSA 2017 and the COMPAS Works-In-Progress Seminar series for their feedback and comments at various stages of this paper's development. Also, staff at the Sketch Engine (www.sketchengine.co.uk) provided assistance in the data collection and advising on corpus linguistic 'best practice': in alphabetical order, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, and Akshay Minocha.

One of the core ideas of agenda setting research is that 'elements prominent in the media pictures not only become prominent in the public's pictures, but also come to be regarded as especially important' (McCombs, 2014: 39). Prominence is often measured in two ways: salience, or sheer visibility ('first level'), and attribution, or the kinds of features and properties linked with an issue (Guo et al., 2012). In order to claim that media have effects on agendas and public attitudes, it is necessary to first identify levels of both kinds of prominence.

This paper focuses on the latter measure of prominence ('second level attribute agenda setting') by asking: how have the ascribed attributes of immigration and asylum changed over time? Looking at specific ways of describing immigration and migrant groups that previous research identifies as especially salient for the British public helps make the case that some of these attributes are more likely to have greater effects than others. But, a challenge facing scholars is how to reliably extract attributes from large amounts of text that do not readily lend themselves to manual or small-sample coding. So, this paper aims to demonstrate how methods from the field of corpus linguistics can provide systematic ways of identifying attributes that are simultaneously grounded in the ways that words actually work in real life—specifically through grammatical relationships.

LARGE-SCALE TEXT ANALYSIS: DIFFERENT APPROACHES

'Text as Data' Approaches

With the advent of digital archives and text mining techniques, political scientists have a wealth of data available to them. Developments in statistical methods have enabled researchers to study these texts along many dimensions: their topics (Rossignol and Sebillot, 2005), positive or negative sentiment (Young and Soroka, 2012), relationship with other texts (Lim, 2010), speakers' policy positions (Laver et al., 2003), and the frames contained within them (Baden, 2010). In their landmark paper summarising the state of text analysis under the banner of 'text as data' approaches, Grimmer and Stewart (2011) argue that automated methods can add value to, and significantly speed up, analyses. This is particularly the case when the objective is either to classify items (into known or unknown categories) or to place them along some kind of ideological scale (such as left-right policy preferences).

One of their main points is that any quantitative approaches to text 'should be evaluated on their ability to perform some useful social scientific task' (Grimmer and Stewart, 2011: 270). Perhaps the goal is to measure the tone of articles to see whether some articles are more positive than others (Young and Soroka, 2012), in which case dictionary-

based methods might be useful: researchers consult (or construct) lists of words that 'belong' to a given category, such as 'positive connotation'.[1] Other approaches, such as supervised learning methods, require researchers to hand-code a 'training set' of documents for whichever feature they are looking for, and then use statistical techniques to 'learn' how the given feature operates in that training set. Then, the model applies this learning to a new group of previously unseen documents (a 'test set') to divide them along the desired categories. Ideally, the model's results match hand-coded results of the same test set.

But perhaps the desired categories or features are not known beforehand. This may be the case in instances where researchers want to identify emergent topics or subjects within texts. By using models that cluster documents together, unsupervised learning methods aim to identify these underlying aspects of texts without necessarily fitting them to a pre-determined coding scheme. They are especially useful in situations when 'the categories of interest in a new project or a new corpus are usually unclear or could benefit from extensive exploration of the data' (Grimmer and Stewart, 2011: 281).

Across all these tasks, Grimmer and Stewart emphasise that the assumptions about language use driving these quantitative analyses are wrong, but useful.[2] For example, a major assumption used in pre-processing texts is that documents are 'bags of words': word order does not matter (Jurafsky and Martin, 2009). Obviously, sentences in real life derive meaning from word order. But if the goal is identifying the topic of an article, or establish its tone, then a simple list of highly frequent words may be sufficient. Their point is that choices about which model or technique to use must be linked to a clear understanding of the intended objective and domain: 'there is no globally best method for automated text analysis' (Grimmer and Stewart, 2011: 270).

In the case of second level agenda-setting research, the task that confronts researchers is how to identify and measure the salience of certain attributes, or larger categories of attributes. Some of these may already be known through previous research and theory. Others may be unknown beforehand, emerging from the corpus itself. Furthermore, agenda-setting theory suggests that attributes express properties or characteristics of a given attitudinal object (McCombs, 2014). This paper argues that these expressions manifest themselves most

[1] In which case the dictionary eventually consulted needs to reflect how words are actually used in that topical domain. Otherwise, there is a risk that 'positive' words in one context may actually have negative connotations in another. See Loughran and McDonald (2011) for an example of this in the case of finance.

[2] They highlight two sentences that are similar in structure, but drastically different in meaning: 'Time flies like an arrow. Fruit flies like a banana'. The phrase 'flies like' moves from a metaphorical use to a literal one.

explicitly at the word-level through specific and predictable patterns of usage.[3] Given these challenges, an ideal approach would have four characteristics: (1) it would draw upon previous topic-specific attribute categories; (2) it would refine and modify them if needed based on empirical observations of language use in similar texts to account for domain-specific connotations; (3) it would create new categories if emergent patterns from the dataset demand them; and (4) it would rely on clear, robust, and theoretically sound measures of object attribution to establish the existence of patterns in the first place.

Corpus Linguistic Approaches

Approaches based in corpus linguistics potentially provide advances on each of these four characteristics. Corpus linguistics does not refer to a discipline. Rather, it is 'an approach that facilitates empirical investigation of language variation and use, resulting in research findings that have much greater generalizability and validity that would otherwise be feasible' (Biber and Reppen, 2015: 1). It is distinguishable by several characteristics: (1) it is empirical, based in actual patterns of use in real-world texts; (2) its objects of study (corpora, or collections of texts) are relatively large and principled in collection; (3) it extensively uses computers and semi- or fully-automated methods for analysis; and (4) it includes room for qualitative as well as quantitative techniques. Although computer-based techniques may have recently popularised and enabled quantitative analyses of texts, scholars have actually used corpora for many years. Biber and Reppen observe that 'the standard practice in linguistics up until the 1950s was to base language descriptions on analyses of collections of natural texts: pre-computer corpora. Dictionaries have long been based on empirical analysis of word use in natural sentences' (2015: 2). In practice, these characteristics enable analyses that can rely on patterns emerging from the data (a 'corpus-driven' approach), test pre-determined hypotheses (a 'corpus-based approach), or a mixture of both (Tognini-Bonelli, 2001).[4]

What advantages does this approach afford to the task of identifying agendas in large amounts of text? First, it is grounded in empirical observations about how language actually works in real life. Instead of relying on hypothetical examples or researchers' own (and necessarily limited) experiences to generate codes or candidate word lists, corpus methods can comprehensively identify all examples of a given pattern—whether it is highly frequent

[3] Though, as observed in subsequent sections, it is possible attributes may also appear in larger units of texts: the approach used in this paper does not claim to capture *all* attributes, but rather the most explicit and immediately-salient ones.

[4] Many corpus linguists use both approaches in iterative ways: emergent findings inform hypotheses which are tested further. See Baker (2006) for examples.

or not.[5] These observations, in 'corpus-driven' settings, can be as domain-specific as the corpora used.[6] Second, it gives guidance about how to identify 'attributes' of concepts using reliable measures. Linguistics provides conceptual and practical tools for figuring out how words relate to one another in ways that go beyond looking for co-occurrences within whole articles. Instead, by looking at the level of words for patterns of nouns and adjectives—concepts that are well-specified for the purposes of automated searching (Marcus et al., 1993)—researchers can be reasonably sure that the attributes they identify are actually referring to the desired object. Third, it removes some human subjectivity and error by taking advantage of recent developments in computing, but still enables efficient, qualitative validation.[7] For example, disambiguation of similar terms (such as 'asylum' in mental health contexts and 'asylum' in forced migration contexts) is an important step in confirming that the quantitative results reflect the intended object. The remainder of this paper explains and demonstrates the usefulness of corpus methods and techniques for identifying first- and second-level agendas in a large amount of newspaper texts spanning up to 30 years.

DATA AND METHODS

Data Sources and Collection

The main source of data for this paper is a corpus of migration-related articles from nine UK national daily newspapers, spanning 1 January 2001 to 31 December 2015. The corpus data come from Nexus UK and Factiva. These are online archival services that cover many international periodicals and other publications. These were chosen because they can deliver full-text versions of newspaper content in standardised formats that enable large-scale data collection.[8]

Using a search string developed by Gabrielatos (2007), items were retrieved from these sources that contained a selection of migration-related terms: [refugee! OR asylum! OR deport! OR immigr! OR emigr! OR migrant! OR illegal alien! OR illegal entry OR leave to remain) NOT (deportivo OR deportment)]. The '!' symbol is a wildcard, which includes variations of terms such as plural forms ('refugee' and 'refugees') and verb forms

[5] This feature of 'comprehensiveness' can reveal relatively rare or unusual language use: see Baker (2006).
[6] One of the most well-known corpora available for research use is the British National Corpus (of British English, with subcorpora containing different genres of writing). Other corpora can be very specialist, from those that contain the works of Charles Dickens (Mahlberg, 2007) to letters from companies to their shareholders (Pollach, 2011).
[7] More details on all of these points will be explored in the section 'Analytical Procedures'.
[8] However, despite their wide-ranging coverage and point-and-click interfaces, these services still come with health warnings: see 'Limitations' later in this paper.

('immigrating' as well as 'immigration').[9] All sections of each newspaper were included in the search, because it is difficult to presume where migration-related content will appear among each publication. Also, people may encounter information about migration in many forms besides typical current affairs reporting: through mentions of athletes' backgrounds, reviews of films involving refugees, or opinion columns talking about asylum-seekers.[10]

The corpus includes the daily versions of nine national British newspapers, divided into tabloids and broadsheets in Table 1. These publications cover the breadth of the British press, with the important exceptions of the Independent, the News of the World, and the i. In total, the corpus contains 216,384 items. Table 1 shows how these are distributed among the nine publications. Tabloid items comprise about 40% of the corpus, while broadsheet items make up about 60%.

Table 1. Publications Included in the Corpus

Tabloids	Articles	Share of Corpus	**Broadsheets**	Articles	Share of Corpus
Daily Mail	22,119	10.2%	The Guardian	43,747	20.2%
The Sun	21,743	10.0%	The Times	35,219	16.3%
The Express	18,833	8.7%	Daily Telegraph	26,655	12.3%
Daily Mirror	15,319	7.1%	Financial Times	25,272	11.7%
Daily Star	7,477	3.5%			
TOTAL	**85,491**	**39.5%**		**130,893**	**60.5%**
Note: Figures do not add to 100% due to rounding.					

Establishing the size of the monthly 'news hole': constructed week method

A significant limitation of prior studies is that they do not include a measure of the total amount of non-advertising content in each publication—the 'news hole' (Jones and Carter, Jr. 1959). Rather, these studies rely on the raw frequencies of mentions, or the number of items mentioning a given term, as a measure of salience. This is a problem because spikes and lulls in migration coverage may actually be meaningless once the overall number of items published is taken into account. For example, assume Newspaper X published 100 articles that mentioned 'immigration' in both January 2001 and January 2015. Based on these

[9] 'Deportivo' is a Spanish football club, while 'deportment' refers to etiquette. The specific term 'migration' is not included because it might result in retrieving articles not related to human migration, such as the movement of animals ('bird migration') or information ('data migration').

[10] For example, discussions about immigration appeared with coverage mentioning British athletes' performances during the 2012 Olympics—particularly around Mo Farah, born in Somalia, and Jessica Ennis-Hill whose father is of Jamaican/Afro-Caribbean origins (Allen and Blinder, 2012).

equivalent raw figures, one might conclude that Newspaper X gave similar levels of priority to immigration across this period. But, if Newspaper X published 2000 items in January 2001, and 4000 items in January 2015, then the picture is different. Salience, as measured by the proportion of items mentioning 'immigration', would have actually declined from 5% of items in January 2001 to 2.5% in January 2015.

This fictional example illustrates the potential problems that the lack of a baseline—even if roughly estimated—introduces to the task of making claims about how the visibility of some aspect of language has changed over time. To address this problem, this paper uses a constructed week method (Jones and Carter, Jr. 1959) to estimate the number of items that each publication produced every month. The underlying principle of this method is that randomly selected days, when summed and scaled up to match the actual distribution of days observed in a given month, create a baseline that approximates the actual news hole.[11]

First, six days per year were randomly chosen. These days corresponded with each day of the week, excluding Sundays: one Monday, one Tuesday, etc. Accounting for the fact that each year and month has slightly different frequencies of each day matters because not all news days are equal: Wednesdays tend to have more content, for example (Lacy et al. 2001). Then, all the content held in Nexis or Factiva was downloaded for each selected day to generate a number of total articles for that day as well as the number of words they contained.[12] Next, consulting available calendars provided the actual number of Mondays, Tuesdays, etc., in each month between 2001 and 2015. This accounted for variation among months, particularly February that can have up to four fewer weekdays and Saturdays. Then, the actual numbers of each day per month were multiplied by the corresponding value of items or words for the sampled day. Summing all of these estimated values for all the days in each month produced an estimated baseline number of items and words in the news hole.

Comparing estimated counts with external corpora

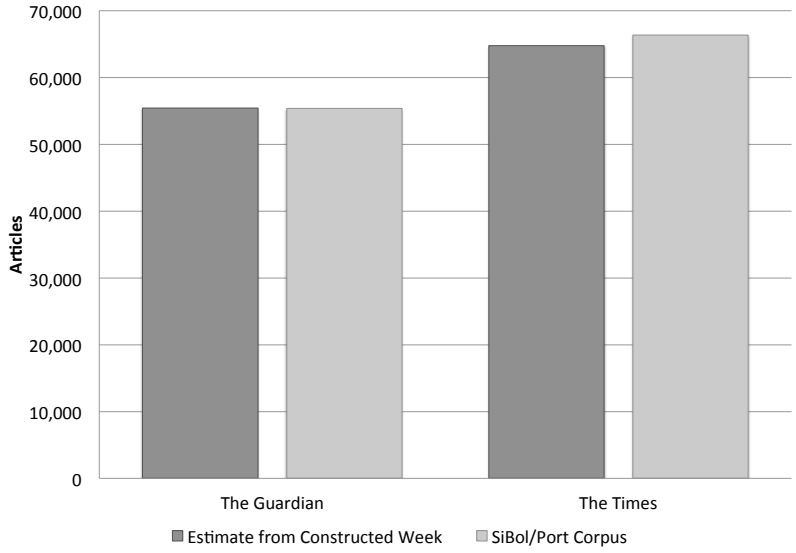
Validating the accuracy of these estimates, particularly over multiple decades and for several publications, poses some challenges. Notably, there are few systematically collected and

[11] In their 1959 research note, Jones and Carter, Jr. report on a study done by the Association for Education in Journalism (AEJ) that analysed three weeks' worth of news in 90 US newspapers comprising about 70,000 pages. The sizes of many of these newspapers' news holes were already available, but for four publications that did not have news hole data available, the AEJ study used a similar 'constructed week' approach used in this paper. The results, in their words, were 'gratifying': after building 30 constructed weeks for each paper, and comparing these results with manually collected 'true' measurements, the study found that 85% of the constructed weeks fell within +/- 2% of the actual figures (Jones and Carter, Jr. 1959: 403).

[12] This excluded online content, as the original dataset also did.

documented datasets of all newspaper content available to researchers. One exception is a corpus of selected UK broadsheet newspaper content, collected by a team of linguists for linguistic research. Called the 'SiBol/Port Corpus',[13] this dataset contains all content from five British broadsheets in 1993, 2005, and 2010 (Sketch Engine, 2011). It was built in successive waves: the first wave, for example, contained texts that the publishers were legally able to release with no duplicates. However, later waves were manually downloaded from archive services, and involved removing duplicates, multiple editions, online content, or other extraneous articles.

Figure 1. Number of Articles in The Guardian and The Times, 1993, Constructed Week Method v. SiBol/Port



To test how the systematic 'constructed week' approach compared to the 'official' SiBol/Port corpus, Figure 1 compares SiBol/Port data about the number of articles from 1993 for The Guardian and The Times with the estimated baseline totals for the same year.[14] The results indicate that the constructed week method produced article estimates that differed 0.07% from the SiBol/Port corpus for The Guardian (55,415 estimated articles in 1993, compared to 55,377 in SiBol/Port). Meanwhile, the difference for The Times was 2.47% (64,757 estimated articles in 1993, compared to 66,357 in SiBol/Port). These differences

[13] For the institutions where its original contributors came from: University of Siena, University of Bologna, and Portsmouth University.

[14] In a fuller version of this paper, which also looks at first level agenda-setting, the analysis includes findings from 1985 onwards, where available in the archive services. 1993 was chosen because, upon correspondence with one of the original SiBol/Port authors, this year in the dataset is the closest reflection of the content in the actual paper version. Also, subsequent years' results were collected under largely different conditions to those of the constructed weeks.

closely follow expectations found in Jones and Carter, Jr. (1959). Although similar benchmarks for tabloids are not available in SiBol/Port, these results do lend support for the systematic approach afforded by the constructed week method.[15]

Analytical Procedures

The corpus was organised, stored, and analysed using the software Sketch Engine (Kilgarriff et al., 2014). This is a comprehensive tool, typically used by lexicographers, that enables researchers to generate snapshots of how a chosen term (called a 'headword') functions in a corpus, with links to the headword as it appears in the actual texts. These snapshots are called 'word sketches' because they potentially give the researcher an initial picture of how a given headword actually operates in real-world language, with the option to go into more detail.

Word sketches rely upon part-of-speech (POS) tagging. This is a technique that attaches information about how each word is used—its grammar—that allows Sketch Engine to look for patterns in usage. For example, if a word modifies a noun (such as 'immigrant'), its part of speech would typically be an adjective. So, if a corpus is POS tagged, researchers could search for all adjectives associated with the word 'immigrant'.[16] The specific set of tags used by Sketch Engine come from Marcus et al. (1993). Two techniques from corpus linguistics feature in this paper. The first is frequency analysis, where specified terms and the articles in which they appear are totalled up to show how prevalent they are in a given corpus or subcorpus.[17] The second is collocation analysis, a method that determines how strongly one word is linked with a target word, as opposed to them appearing together by random chance. Conventionally, collocation is defined as 'a co-occurrence relationship between two words' (McEnery and Hardie, 2011), typically a 'node' and its 'collocate'.[18]

Figure 2 shows what this analysis looks like by displaying an example word sketch for the headword 'immigrant', with the results sorted by frequency. The third column displays a statistical test for collocation.[19] Drawing upon the POS tagging, word sketches

[15] Also, as explained in the next section detailing analytical procedures, both SiBol/Port and the corpus collected for this paper are stored and organised within the same software. This also lends support for the claim that the counts are similar: using the same software removes any differences that might have been introduced by different programming schemes.

[16] This would include comparative and superlative adjectives, such as 'larger' and 'largest'.

[17] A subcorpus is a research- or researcher-defined subset of the larger body of text.

[18] Determining what counts as 'co-occurrence' is also debatable: the answer depends on several choices taken by the researcher (Brezina et al., 2015; McEnery and Hardie, 2011). For fuller discussion of collocation in the linguistics literature, see Lehecka (2015).

[19] The statistic is logDice, developed by Rychlý (2008). This research uses collocations sorted by frequency instead of the logDice measure. This is because agenda-setting relies on the visibility of an object or attribute. Therefore, prioritising collocates' frequencies in determining which ones to include is more appropriate.

display collocates that are grammatically linked to the headword. In this example, it is clear that the adjective most frequently associated with 'immigrant' or 'immigrants' across the entire corpus is 'illegal'.[20] Also, nouns that are frequently modified by 'immigrant' include 'worker', 'population', and 'labour'. Meanwhile, in line with corpus linguistic practice, Sketch Engine enables researchers to look more closely at how these collocations behave in real texts. Called 'concordance analysis', this technique displays collocation results with the surrounding text (Baker, 2006). Figure 3 shows what this concordance view looks like. This technique, when linked with quantitative approaches, provides an important validity check and an efficient way to see how collocates appear. These two relationships—adjective and noun collocations—form the basis of the second-level attribute agenda setting analysis.

Figure 2. Word Sketch for the Noun 'Immigrant'

immigrant *(noun)*
COMPAS 2016 NM_All freq = 110,807 (487.96 per million)

modifiers of "immigrant"			nouns and verbs modified by "immigrant"		
	63,065	0.57		19,840	0.18
illegal +	23,509	12.68	community +	1,864	9.46
many +	1,727	7.84	worker +	1,546	8.79
new +	1,456	7.03	population +	1,168	9.41
jewish +	1,356	8.75	family +	1,106	8.71
irish +	932	8.17	parent +	737	9.36
chinese +	848	8.22	group +	615	6.97
more +	848	7.05	labour +	425	8.20
european +	829	7.48	child +	395	7.41
would-be +	760	8.49	council +	330	7.07
russian +	739	7.91	experience +	291	7.67

Figure 3. Concordance View of 'Illegal Immigrant(s)'

Word sketch item **23,509** > GDEX **23,509** (103.53 per million)

Page 1 of 1,176 Go Next | Last

doc#213808... Mr Davis, aged 34, who comes from Ghana, had been in Britain as an ***illegal*** immigrant for 16 years.

doc#211293... Greek police yesterday arrested 75 ***illegal*** Kurdish immigrants on Crete after a tanker captain had convinced them that they had arrived in Italy. - AP.

doc#206164... It came amid dire warnings there are not enough detention places to lock up failed asylum seekers and ***illegal*** immigrants .

doc#205948... Normally, asylum seekers would be arrested on suspicion of being an ***illegal*** immigrant , taken to a police station and interviewed by an immigration officer.

doc#205246... The Greek people don't want ***illegal*** immigrants ,' he says.

[20] This confirms previous results from a pilot study that looked at the 2010-12 period (Allen and Blinder, 2013). That study used more mechanistic rules to identify adjective collocates—such as focusing on collocates appearing immediately before a target word. Sketch Engine and POS-tagging provides a more systematic and comprehensive way of identifying collocates in texts that, by their very nature as qualitative data, do not always follow mechanical rules.

Limitations

One of the most obvious limitations relates to the data sources. Nexis and Factiva rely on publishers uploading their content to the services. For major, national publications like those used in this study, most content is readily available. However, it does leave the possibility of duplicates—such as multiple editions—entering the corpus and artificially inflating the results if publishers decide to provide them. So, to counteract this problem, any duplicate articles appearing on the same day and in the same publication were removed.[21]

Another limitation relates to the choice of analysing national print media at the expense of other forms of media. There are open questions about how, where, and when people access news or other information (de Zuniga et al., 2012; McCombs, 2014). However, as Vliegenthart and Walgrave (2008) demonstrate, newspapers play especially strong roles in setting agendas among media types. Furthermore, many articles that circulate online have their basis in 'traditional' publications. Finally, since this paper is concerned with tracking changes in visibility over a longer period of time, limiting the dataset to newspaper texts that were stored digitally was a practical decision.

Also, collocation at the word-level presents some limits in the analysis. Considering only noun and adjective collocates is just one way of identifying the properties or attributes required for second-level agenda setting. It is possible that larger units of text, at the sentence or paragraph level for example, may express relevant attributes, too. However, given the theoretical and conceptual strength of these kinds of collocations for explicitly marking how certain words ascribe meaning to objects, they lend a more systematic and efficient way of handling large amounts of text.

EMPIRICAL MATERIAL

Second-Level Agenda Setting: Identifying Attributes Using Collocation Analysis

In an agenda-setting context, attributes are 'those characteristics and properties that fill out the picture of each object' (McCombs, 2014: 40). The question is how to efficiently, yet accurately, identify attributes in large amounts of text. The approach this paper takes is collocation analysis combined with POS tagging. Attributes can link with their objects in several ways. One way is through explicit modification, as in the phrase 'illegal

[21] Journalists often rely on widely-circulated press releases. These usually contain pre-approved quotes or other background material. Therefore, removing all articles that contained similar text on the same day could remove individual articles from two different publications—a problem for trying to measure overall visibility. This is the reason for including the extra criterion of duplicates within the same publication: two very-similar articles in the same newspaper on the same day are likely to be copies. De-duplication was done within the Sketch Engine interface using techniques developed in Pomikálek (2011).

immigration'. In this example, the word 'illegal' modifies the object 'immigration'. This relationship is an adjective collocation. Another way is when objects are used as modifiers themselves for another object: 'immigration policy' is a case where the term 'immigration' is now used as a modifier for the object 'policy'. This is a noun collocation. When a corpus contains POS tags, it is possible to distinguish among collocations that refer to a specific term (such as 'immigration' or 'refugee') and those which do not. Those that do can be thought of as 'attributes' or 'properties' in the second-level agenda setting sense.

Studying adjective and noun collocations of key terms gives a window into the ways that key terms regularly appear in a corpus or subcorpus. Lists of individual collocations can give some early, exploratory indication of 'typical' usages (McEnery and Hardie, 2011). However, collocations often can fit within broader categories of attributes—whether inferred from the corpus data, or informed by prior research and theory.[22] Tracing these categories' shifts over time can reveal important trends in second-level agenda setting.

Building attribute categories: a corpus linguistic approach

Constructing salient and theoretically appropriate attribute categories required several steps. First, word sketches provided collocation candidates for two sets of target headwords: 'immigration', 'migration', 'immigrant(s)', and 'migrant(s)' on the one hand; and 'asylum', 'asylum seeker(s)', and refugee(s)' on the other.[23] These lists contained up to 200 of the most-frequent noun and adjective collocations explicitly associated with each headword.[24] Then, prior studies and theory—combined with close inspection of the collocation lists themselves—informed the attribute category construction process. Table 2 provides a summary of the six attribute categories generated by the collocation analysis, a brief description of the categories, and example collocates within them. The descriptions reflect the refinements described above, as well as prior theory and empirical studies of British press coverage relating to migration or minority groups. The chosen collocates are illustrative, not exhaustive, and are not necessarily the most frequently observed in the corpus.

Two studies were particularly informative in building this scheme of attribute categories, especially since they focus on different aspects of migration in the UK press

[22] Vollmer (2017) also uses this inductive method in the Sketch Engine interface to study how language around the key term 'border' differs between British newspapers and policy spheres.

[23] Concordance analysis identified instances where 'asylum' and its collocates appeared in the context of mental health. These were subsequently excluded from the analysis.

[24] Collocations also included alternative forms: for example, 'high immigration' and 'higher immigration'. Some headwords did not have 200 noun or adjective collocations. In these cases, all collocations identified by Sketch Engine were examined.

context. McLaren et al. (2017) also aimed to identify second-level agenda setting about migration in four UK newspapers.[25] After identifying the most frequent words in their corpus, omitting stopwords such as 'and' or 'the', they used four human coders to determine whether these words were relevant to the topic of immigration. Reserving those words that at least three out of four coders thought were relevant to immigration left a list of 350 terms. After using statistical clustering methods,[26] they conclude that there are five coherent issues ('factors') present in the corpus: the economy, crime/security, government policymaking, foreign wars/rising numbers of refugees, and education. Crucially, for them the unit of analysis lies at the level of articles: 'an issue is considered to be present in a story if at least three of the words loading on a given factor occur in the story' (McLaren et al., 2017: 10).

Table 2. Attribute Categories Scheme and Examples

Attribute Category	Description	Example Collocates
Economic and Occupational	Relating to real or perceived financial background or situations, either past or present; labour, working, or (un)employment; specific jobs	*cleaner, doctor, economic, labour, low-skilled, poor, scrounger, skilled, unemployed, worker*
Criminality and Legal Status	Relating to explicit crimes or anti-social behaviour; evaluations of individuals' legal right to be in the country, whether actual or perceived; procedures or activities involved in establishing legal status	*abuse, amnesty, criminal, dangerous, failed, fraud, illegal, overstayer, smuggler, would-be*
Legislative, Policy, and Governmental	Relating to actions, individuals, or procedures occurring in policy, legislative bodies, or government; references to organisations, bodies, or associations operating in these areas	*advisor, application, bill, department, minister, official, procedure, regulation, scheme, system*
Demographic and Sociocultural	Relating to characteristics used to differentiate along dimensions of age, sex, family structure, education, relationships, ethnicity, health status, religion, sexual orientation, or ability; references to collective views, attitudes, activities, shared cultural artefacts and ideas, or emotions	*ancestor, black, child, culture, experience, gay, history, husband, mother, Muslim*
Geographical	Relating to national origins, countries, or regions	*Afghan, British, Chinese, EU, European, Indian, non-EU, Pakistani, Polish, Romanian*
Scale and Pace	Relating to the speed or amounts of mobility, whether real or perceived; references to the ways in which mobility happens	*boom, chaotic, excessive, flood, mass, number, stock, substantial, unlimited, unrestricted*

[25] Their study included two broadsheets (The Times and The Guardian) and two tabloids (Daily Mirror and Daily Mail).
[26] Their paper does not provide full details of the clustering methods, but does refer to Hellsten et al. (2010) as representative of their techniques.

McLaren et al.'s (2017) study was informative in that it provided some guidance on sets of attributes which also were likely to appear in the corpus used in this paper, given the similar newspaper sources. They found that the issue of 'economy', for example, was signaled by terms such as 'employment', 'job/jobs', and 'work/workers/working'. Also, the issue of 'government policymaking' comprised words including 'application', 'convention', 'department', and 'order'.[27] These kinds of words were also salient in this paper's collocation lists. So, these categories were retained in this paper's final scheme.

The second study from which this paper draws insight comes from Baker et al. (2013b).[28] This was a linguistic study of the term 'Muslim' in over 200,000 articles that appeared in British national newspapers from 1998 to 2009. Also using collocation analysis in combination with POS tagging in the Sketch Engine interface, this study generated a comprehensive list of 1,256 noun collocates associated with 'Muslim'. Manual concordance analysis of all of these collocates produced a scheme of categories and subcategories. These included 'culture', comprising social practices, education, and attitudes; and 'characterising/differentiating attributes', which combined demographic features such as age and sex with other references to kinship, occupation, or nationality.

To a certain extent, this scheme is useful because it is grounded in a close reading of portions of newspaper articles that specifically refer to the term 'Muslim' in some way: 'the categorization does not rely so much on the dictionary meaning of the noun collocates, as on the topics they index in the corpus articles' (Baker et al., 2013b: 262). The categories emerging from the concordance analysis also have some relevance for, and consonance with, general press coverage about immigrants. Specifically, some subcategories referring to 'characterizing/differentiating attributes' were combined into a larger attribute set comprising demographic, religious, social, and cultural features.[29] Also, their subcategory of 'ethnicity/race/nationality' was narrowed to a set of 'geographic' terms referring to specific countries or regions, such as 'Afghan', 'EU', or 'Pakistani'. This differs from McLaren et al. (2017), where national terms including 'Polish' were subsumed into other categories.[30]

[27] However, their issue of 'government policymaking' also included the terms 'asylum' and 'seeker/seekers'. This illustrates a problem with lumping together all types of migration together: the argument in this paper is that differentiating among 'migrants' and 'asylum-seekers', even at a basic level, reveals important differences in attribution.

[28] Expanded findings also appeared in book form (Baker et al., 2013a).

[29] Initially, 'cultural/social' and 'demographic' terms were kept separate. However, subsequent analysis revealed that these categories were too infrequent on their own to warrant this approach.

[30] In the case of 'Polish', McLaren et al. (2017) place articles mentioning this term in the 'economic' attribute category.

Using noun collocates—and collocation analysis more generally—also informed other key differences and refinements in the attribute scheme developed by McLaren et al. (2017). For example, 'crime' was a prominent category in their scheme, as indicated by terms such as 'officer/officers' and 'association'. However, examining the noun collocates reveals that these terms are often modified by 'immigration', 'border', or 'refugee', and do not connote a criminal sense. So, while the scheme used in this paper does include a category of attributes related to criminality based on McLaren et al. (2017), the terms in that category differ. At the article level, it is likely that a range of words would appear alongside other crime-related terms. But, collocation analysis at the word level enables researchers to draw a more specific, grammatically precise relationship between key terms such as 'immigration' and other words appearing nearby.[31]

A final category of terms that does not appear in either McLaren et al. (2017) or Baker et al. (2013b) relates to the scale or pace of migration. These terms, illustrated by words such as 'excessive', 'flood', and 'mass', emerged from the collocation lists. Prior studies into the drivers of immigration attitudes debate the extent to which the sizes of outgroups—whether real or perceived—matters for public preferences (Blinder, 2015; Herda, 2010; Pottie-Sherman and Wilkes, 2017). Accounting for collocates referencing the rate at which migration happens, as well as its gross levels, is an important addition to the attributes already considered, especially when linking these second-level agendas to public concern.[32]

Descriptive features of the attribute categories: comparing migration types

How are these categories distributed across the corpus, and how well does this scheme cover the observed variations in collocations? Table 3 displays the number of noun and adjective collocations contained within each attribute category, broken down by each set of headwords. Each cell displays the total number of collocations observed within a given attribute category, as well as the percentage of all collocations that category comprises. Adding up all the percentages within each set of headwords shows that this scheme captures 77% of all noun and adjective collocations for immigration-related terms, and 53.67% of all similar collocations for asylum-related terms.

[31] There are also potential ways of combining the statistical approaches used in McLaren et al. (2017) with those used by corpus linguists: see McEnery (2015) for some provocations on this point of balancing expertise and theory with algorithms.

[32] A previous study of the period January 2006-May 2015 found that the proportion of collocates referencing the scale or pace of migration was increasing in the British press (Allen, 2016). This finding is replicated and expanded upon later in the paper.

Table 3. Distribution of Attribute Categories by Migration Type

Reference Terms (Headwords)	Economic, Occupational	Criminality, Legal Status	Legislative, Policy, Governmental	Demographic, Sociocultural	Geographic	Scale, Pace
'Immigration', 'Migration', 'Immigrant(s)', 'Migrant(s)'	19,828 instances (9.03% of all collocations)	39,306 (17.91%)	53,160 (24.22%)	12,066 (5.5%)	17,915 (8.16%)	26,738 (12.18%)
'Asylum',[33] 'Asylum-Seeker(s)', 'Refugee(s)'	1,566 (1.21%)	11,003 (9.5%)	22,565 (17.44%)	9,059 (7%)	16,867 (13.03%)	7,102 (5.49%)

At a broad level, mentions of terms related to both immigration and asylum were most often attached to 'legislative, policy, and governmental' attributes. This makes some intuitive sense: mainstream newspapers tend to report and rely on 'official' government sources or activities for much of their content (Entman, 2003).[34] This finding matches McLaren et al. (2017), who found the issue of 'legal processes' was the most visible from 1995-2012 with few exceptions.

However, distinguishing between types of migration reveals differences in levels of the remaining attribute categories. Regarding immigration/migration/immigrants/migrants, the second most visible category was 'criminality and legal status': when there was a noun or adjective collocation of one of these four terms in the corpus, about 18% of the time it related to crime or legality. This also echoes McLaren et al. (2017) whose 'crime' issue remains in second place until the mid-2000s before declining to third place. But this level was not observed in mentions of asylum/asylum-seekers/refugees, where only about 10% of collocates related to this category, placing it in third. This difference suggests that the press agenda links immigration and immigrants with criminality or breaking established procedures more than it does asylum seekers or refugees.

In fact, the second-most frequent category associated with asylum seekers and refugees was 'geographic'.[35] These attributes, as seen in Table 2, relate to identifying places of origin or association. In one way, this could be interpreted as simply statements of fact: when reporting on flows of people out of a country due to potentially visible and international

[33] See note 23.
[34] Manual content analysis of British newspaper coverage of 'illegal immigration' and 'EU/European migration' between 2006-15 also showed that politicians and other government officials were most visible—and blamed for perceived problems—in articles (Allen, 2016).
[35] Previous studies also find that geographic terms tend to appear with mentions of asylum-seekers or refugees: see Blinder and Allen (2016).

conflicts, for example, it makes sense to refer to these people in relation to the country or countries they are from—especially when reporting from the perspective of an external country like the UK. However, in another way, the fact that these attributes are not as strongly associated with immigration/migration/immigrants/migrants—geographic attributes only make up about 8% of collocations with these terms, placing it in fifth—suggests this could be a subtle way of establishing difference or distance between British audiences on the one hand and asylum-seekers or refugees on the other.[36]

'Scale and pace', as an emergent set of attributes not necessarily predicted by prior studies or theory, was the third-most frequent category applied to immigration. About 12% of the time when a noun or adjective collocate appeared with one of the four immigration reference terms, it was within this set of words. This was not the case in collocations of asylum terms, where they appeared only about 5.5% of time, behind four other categories.

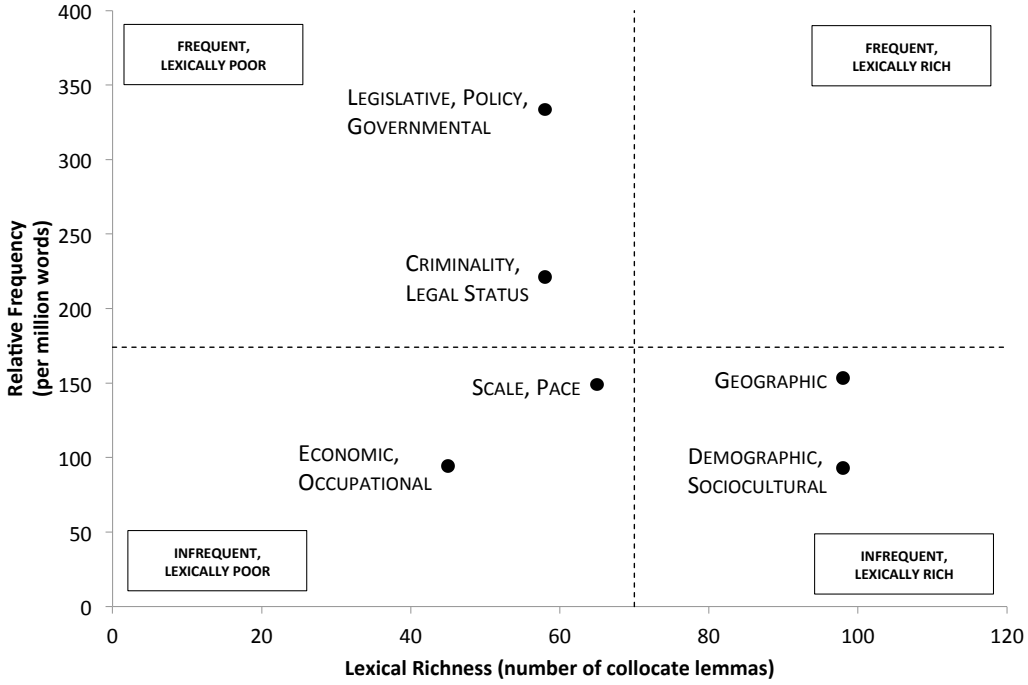
It is also revealing to turn attention towards those categories that were not as visible. For example, 'economic and occupational' attributes were the fourth-most frequent in relation to immigration, only comprising about 9% of observed collocations. It was even more infrequent in reference to asylum, appearing in only about 1% of collocations and making it the least visible category for that reference group. Similarly, 'demographic and sociocultural' attributes were relatively infrequent in relation to both groups of terms. On the one hand, this result is somewhat surprising given the enormous debate in public opinion research about whether economic or sociocultural concerns drive attitudes (Ceobanu and Escandell, 2010; Hainmueller and Hopkins, 2014). It would be reasonable to expect economic considerations about immigration, such as competition over jobs or lowered wages, to feature more prominently in press coverage.

On the other hand, it may be that the relatively specific tests of noun or adjective collocation as indicators of attributes may be missing ways of talking about immigration or asylum in economic and sociocultural terms that do not necessarily occur via explicit attribution in a single word. Instead, they may arise through more complex phrases, sentences, or larger sections of text. Even the simple sentence 'immigrants are taking British jobs' would not necessarily count as a collocation in the sense used in this paper. Including all such constructions, while possible to do in relatively simple cases, opens the door for other quantitative and qualitative methods. This paper argues that looking for nouns and

[36] Scholars of discourse analysis call this ability to establish and maintain (often unequal) relations among groups 'social power' (van Dijk, 2008). Other media and communication scholars have documented how this approach has been a staple feature of British coverage of asylum issues for decades, particularly among tabloids: see Greenslade (2005).

adjectives appearing with an object is a simpler, yet effective, way of identifying its properties.[37] Nevertheless, despite this methodological point, it is striking that McLaren et al. (2017) *also* find that the issue of 'economy' is third (out of four attribute categories) with few exceptions until mid-2006, when it moves into second-place behind 'legal processes'.

Figure 4. Comparison of Relative Frequency and Lexical Richness Among Attribute Categories



Comparing these categories in terms of their relative visibility and diversity of terms also shows some broad aspects of British press coverage about migration issues.

Figure 4 plots the relative frequency of all the terms in each attribute category in the corpus against the number of terms in that category.[38] Relative frequency is calculated by dividing the total raw frequency of all collocations in a category. The number of distinct

[37] Of course, it is possible to imply economic attributes via the phrase 'asylum-seekers who take benefits', the attribute being 'someone who in receipt of financial payments from the state'. The point of using noun and adjective collocations is that they are theoretically grounded and practically efficient ways of accurately identifying explicit characterisations or properties ascribed to an object or issue—a key part of the second-level attribute agenda setting concept. This paper does not claim to capture *all* variations of attributes.

[38] This includes all references to either 'immigration/migration/immigrants/migrants' or 'asylum/asyum-seekers/refugees'.

collocate lemmas[39] within each category can be thought of as a measure of 'lexical richness' indicating the level of variety used in a given context (Baker et al., 2013b). The dotted lines indicate average relative frequency and lexical richness among the six categories.

The 'legislative, policy, and governmental' category is the most visibly collocated with migration and asylum, as mentioned above. Its relative frequency is the highest. 'Criminality and legal status' is the second-most visible, having an above average relative frequency. Meanwhile, the categories of 'economic and occupational' and 'demographic and sociocultural' are the least visible, having roughly the same relative frequencies in the overall corpus. However, both 'demographic and sociocultural' and 'geographic' attribute categories are the richest in terms of lexical variety: they are comprised of the most number of distinct collocates compared to the other four categories. Meanwhile, 'scale and pace' is very near average in both relative frequency and lexical richness.

The findings in **Table 3** and

Figure 4 focus attention on the features of each attribute categories at the broadest level of the entire corpus. They suggest that both immigration and asylum as issues in the press tend to be portrayed through governmental and policy aspects, followed by criminality (in the case of immigration) and geographic origins (in the case of asylum). However, when considering all types of mobility together, a relatively smaller set of collocates related to government and policymaking accounts for the large visibility the category gets in the press. This is also observed, to a lesser degree, with terms related to criminality: both have above-average frequency, but below-average lexical richness. What's more, the lack of any categories in the upper right-hand quadrant of

Figure 4—an area of both high visibility and lexical richness—is telling. This suggests that the British press as a whole has not tended to portray mobility in larger, sustained ways featuring diverse, varied vocabularies.[40]

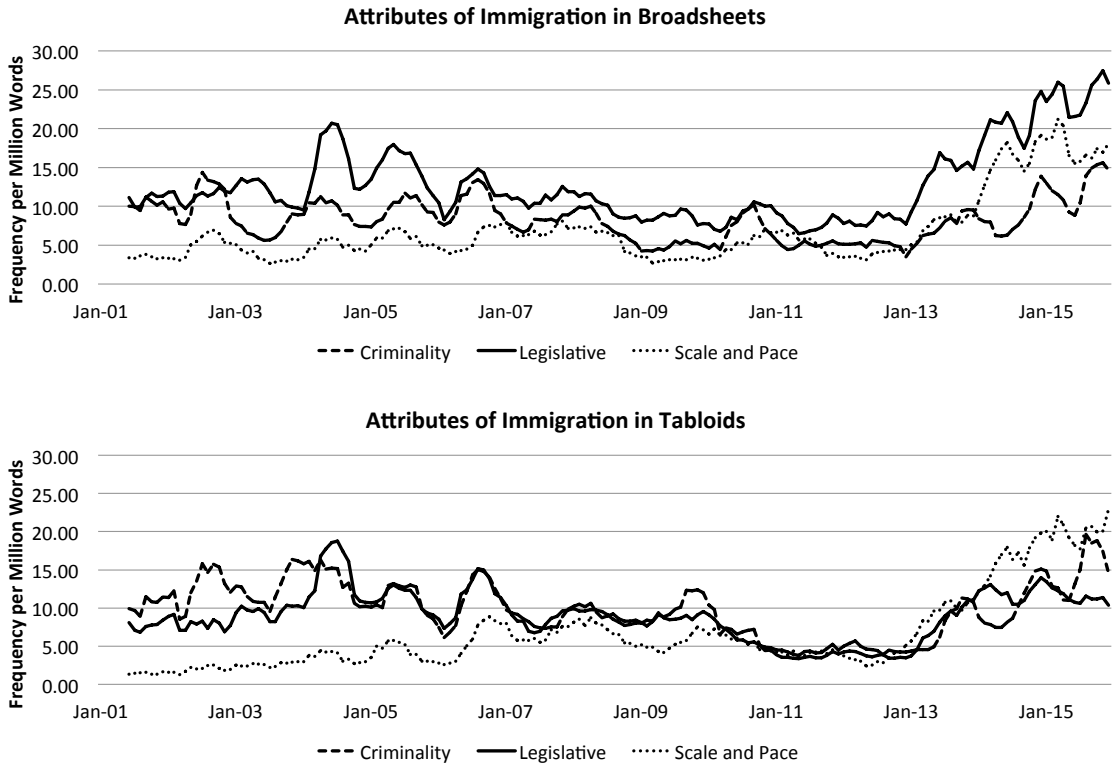
[39] A 'lemma' is the basic form of a word, not accounting for pluralisation or possessives (for nouns), intensifiers (for adjectives), or different verb forms. 'Distinct collocate lemmas' refers to the number of unique collocates in their basic form, omitting variations within each form. So, for the purposes of measuring lexical richness, the collocate set 'girl, girls, boy, boys, pineapples' would have three distinct lemmas: 'girl', 'boy', and 'pineapple'. See Thomas (2015: 19).

[40] Baker et al. (2013b) do a similar analysis of the term 'Muslim' and its collocates. They found that their category of 'characterizing/differentiating attributes' was frequent and lexically rich. However, their scheme combines demographic, kinship, occupational, and nationality terms into this category, whereas the scheme used in this paper separates these into different categories. Also, their category 'ethnic/national identity', which includes terms related to 'governance' that echo those within the category of 'legislative, policy, and governmental' used in this paper, is highly frequent and slightly more lexically poor than the average category in their scheme. These two categories between the two studies, which include overlapping sets of collocates, appear in nearly identical positions on the grid, suggesting some degree of convergent validity.

Changes in Attribute Agenda Setting: Analysing Trends and Variation

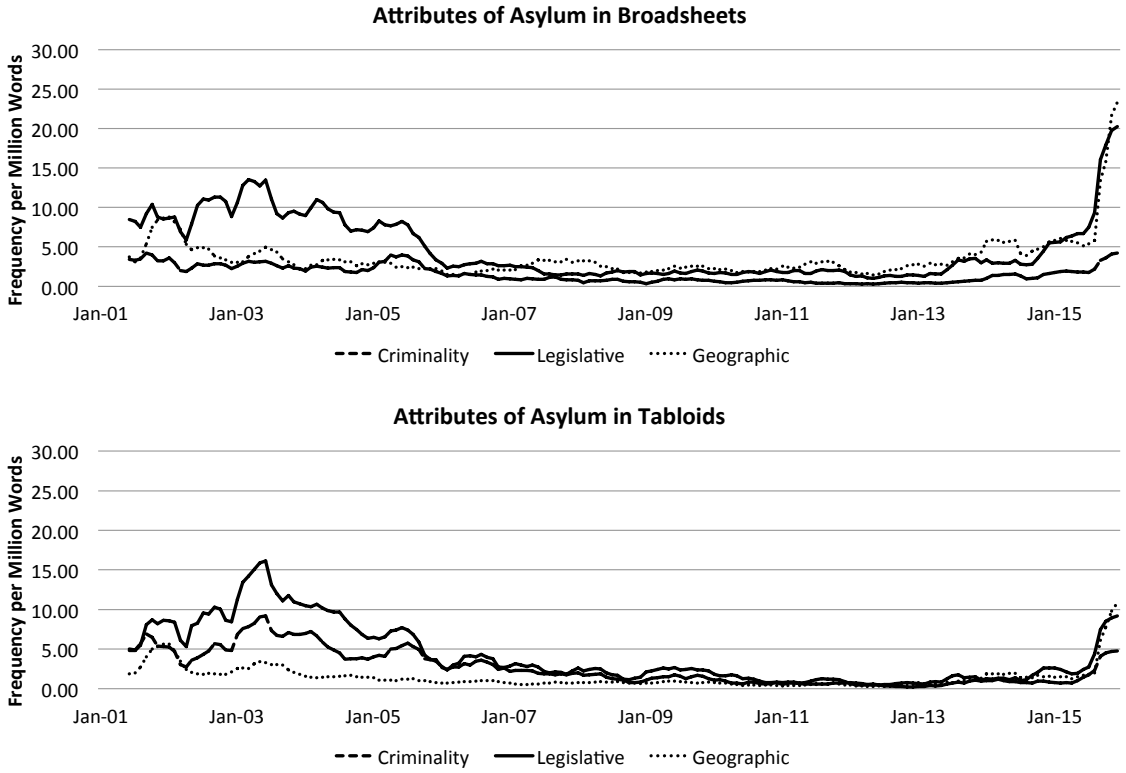
In addition to showing snapshots of attributes at the level of the whole corpus, it is important to examine how these attributes' have changed—or remained constant—over time. This is a step that McLaren et al. (2017) take in their study, but do not differentiate between different types of migration. Figure 5 and Figure 6 display how the visibility of the three most frequent categories for each issue changed over time from 2001-15, divided by both migration type ('immigration' which includes collocations of 'immigration, migration, immgrants, migrants' versus 'asylum' which includes collocations of 'asylum, asylum-seekers, refugees') and publication type (tabloids versus broadsheets). The frequencies indicate how often attributes related to each category appear as a proportion of the estimated total number of words in tabloids or broadsheets.[41] Also, the series are six-month rolling averages to make them clearer. Finally, the charts have the same axes to enable comparison of levels among them.

Figure 5. Selected Attributes of Immigration by Publication Type, 2001-2015



[41] For clarity, the numerator is the number of times collocates in a given category appear in either tabloids or broadsheets. The denominator is the estimated number of words (based on the constructed week method) that appeared in the same publication type. Displaying results per million words is standard, corpus linguistic practice (McEnery and Hardie, 2011).

Figure 6. Selected Attributes of Asylum by Publication Type, 2001-2015



Immigration in broadsheets and tabloids

Focusing on Figure 5, there are a few key trends and differences to note. First is the prominence of the 'legislative, policy, and governmental' category over time. For the press, this is one of the main sets of attributes associated with immigration and immigrants since 2001. Also, the visibility of this category has increased since 2013 in the broadsheets while remaining relatively level during the same period in the tabloids. However, 'criminality and legal status' in tabloids closely matches levels of 'legislative', even exceeding them in the early 2000s.

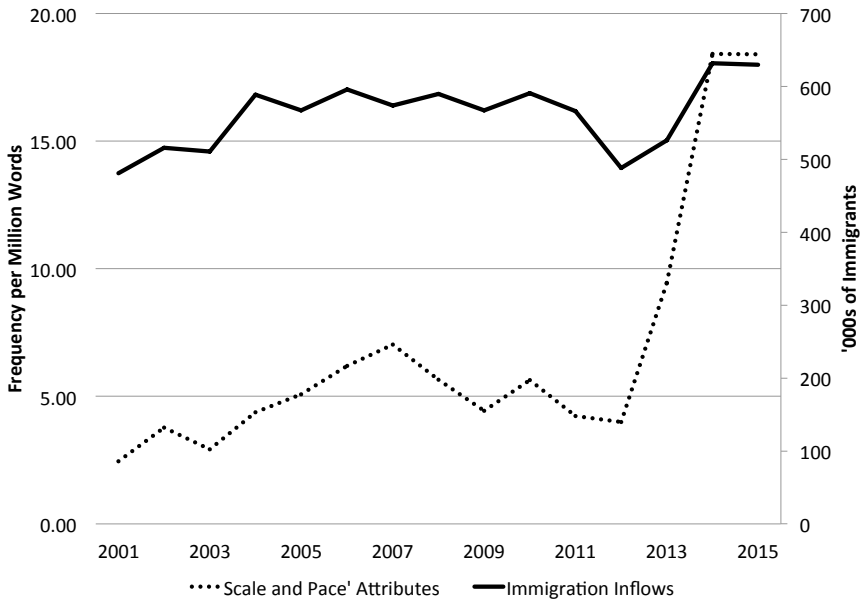
Second, the 'scale and pace' category has risen dramatically over the last 15 years in both tabloids and broadsheets. The relative frequency of these terms increased over five times in the broadsheets between 2001 and 2015 (from 3.36 instances per million words in June 2001 to 18.08 in December 2015), and near 17 times in tabloids (from 1.35 instances per million words to 22.92).[42] Interestingly, much of this increase actually occurred from 2013 onwards, moving this category into second-place among broadsheets and first-place in

[42] Breaking down these figures by publication reveals that The Express, Daily Mail, and The Sun contributed most to this increase.

tabloids—although the actual levels are similar between the publication types. This builds upon previous findings that also documented the rise of 'scale and place' attributes in the British press (Allen, 2016): visibility of these kinds of words remained low even before 2006, where that study began.

What might explain this rise in 'scale and pace' attributes? One possibility is a similar rise in the actual numbers of immigrants entering the UK. Figure 7 shows the annual frequencies of 'scale and pace' attributes for all publications, displayed as a proportion of all estimated newspaper content for that year as calculated for previous figures. It plots these frequencies against the annual number of immigrants entering the UK as reported in Home Office data (Markaki and Vargas-Silva, 2016).[43] Although the results are correlational, they do suggest that the most recent increase in 'scale and pace' attributes from 2012-15 was also accompanied by increases in the numbers of immigrants. This lends further weight for the case of including, but not blindly relying upon, 'real-world' factors in media effects research.[44]

Figure 7. Visibility of 'Scale and Pace' Attributes Compared to Immigration Inflows, 2001-15



[43] Note that these are different from 'net migration' figures that report the number of immigrants minus the number of emigrants.
[44] Van Klingeren et al. (2015) test this idea using the issue of immigration in the Netherlands and Denmark.

Asylum in broadsheets and tabloids

Meanwhile, Figure 6 shows how the three most-frequent attribute categories actually tend to remain fairly invisible for much of the 2001-15 period, with fewer than 5 instances per million words. When attributes are present, they tend to appear in the early 2000s and in late 2015, periods when asylum and refugee policies—as well as external conflicts—were particularly high in public awareness (Gabrielatos and Baker, 2008; Greenslade, 2005). Indeed, during the 2001-06 period, 'legislative' attributes were the most visible in both broadsheets and tabloids, referring to government policies and agencies involved in dealing with asylum-seekers. However, tabloids linked 'criminality' attributes with asylum issues more than broadsheets.[45]

The large spike in 2015 demands attention, too.[46] Both tabloids and broadsheets attributed more 'legislative' and 'geographic' properties to mentions of asylum and refugee groups that year. This was particularly pronounced in broadsheets: 'legislative' attributes nearly quadrupled from 5.58 instances per million words in January 2015 to 20.24 instances by the end of the year. Similarly, 'geographic' attributes rose in broadsheets from 5.84 instances per million words to 23.34 instances—by far their highest levels even compared to the 2001-06 period. Although broadsheets used both kinds of attributes relatively more often than tabloids, the 'geographic' category moved into first-place for both publication types.

DISCUSSION AND CONCLUSION

Scholars, practitioners, and policymakers often identify the British press as a major source of (mostly negative) public attitudes towards immigration and asylum (Duffy and Frere-Smith, 2014; Threadgold, 2009). Recent agenda-setting research observes that the kinds of properties associated with objects, as well as their sheer visibility, can impact how much importance people attach to those objects (McCombs, 2014; McLaren et al., 2017). Given this backdrop, this paper sought to answer two questions. First, how have the frequencies of articles containing migration-related terms changed over time? Second, how have the ascribed attributes of immigration and asylum changed over time? Identifying these trends and variations is crucial for linking media agendas to public agendas. Also, differentiating

[45] Again, as observed in note 42, this increase was largely due to The Express, Daily Mail, and The Sun.
[46] This increase was driven by the 2015 refugee 'crisis' and the release of Alan Kurdi's photo, a finding led by intuition and confirmed by concordance analysis: most 'geographic' mentions refer to Syria. See Fotopoulos and Kaimaklioti (2016) for comparative analysis across Greek, German, and British press.

between different migration types is an important step that prior studies tend not to take for either methodological or conceptual reasons.[47]

With respect to each migration type, the results paint two different pictures. Immigration has generally risen in press visibility, particularly in the last three years. While the press, particularly broadsheets, have tended to link the issue most strongly with governmental or policy attributes, this has been challenged—if not eclipsed—by rising concerns about the speed and rate of immigration. When the press brings attributes related to criminality or legal status into discussion about immigrants or immigration, it tends to be more prevalent among tabloids, a trend that has somewhat grown in recent periods. Meanwhile, issues of asylum and refugees tend to be most visible in specific periods: the early- to mid-2000s, and most recently in 2014-15. While they continue to be linked with governmental attributes, during the recent 'crisis' the press also attributed them with references to both their geographic origins as well as their numbers.

This paper's findings, while consistent with prior analyses of the British press and the ways it presents immigration and asylum issues (Allen, 2016; Crawley et al., 2016; Gabrielatos and Baker, 2008; McLaren et al., 2017), also go further in showing how coverage varies among attribute categories, and how these have changed in visibility over time. For example, it reveals how the speed and size of immigration and asylum flows alike is also increasingly salient in the press. Also, it uncovers some important differences among publication types: tabloids tend to highlight criminal attributes when referencing immigration, whereas broadsheets have consistently linked the issue with governmental or policy aspects. These differences may have implications for the ways that agenda-setting studies handle and subdivide their corpora.

The paper also makes some methodological contributions to the study of texts and corpora in political science. First, it demonstrates the need for, and viability of, systematically establishing a baseline of newspaper content using existing digital archives and relatively straightforward sampling. This stands to benefit future agenda-setting research that aims to examine longer periods of time where accessing complete versions of newspapers is not feasible. Second, it shows how established tools and techniques from linguistics can identify key relationships that serve as reliable indicators of attributes at the word-level. Where there is a good body of theory and empirical work that can inform category building, as in the case

[47] Although McLaren et al. do acknowledge that future agenda-setting research could 'focus more specifically on asylum and/or refugees to understand in greater detail how this topic is framed' (2017: 18). This paper aimed to respond to their work, and continue in a similar path.

of migration in the media (Baker et al., 2013b; Balabanova and Balch, 2010), this is an advantage over current trends in text analysis that often use articles as the unit of analysis to induce categories.

WORKS CITED

Allen W (2016) *A Decade of Migration in the British Press*. Migration Observatory Report, Oxford: COMPAS, University of Oxford. Available from: http://www.migrationobservatory.ox.ac.uk/wp-content/uploads/2016/11/Report-Decade_Immigration_British_Press-1.pdf.

Allen W and Blinder S (2012) *Jessica Ennis, Mo Farah and Identity Language in the British Press: A Case Study in Monitoring and Analysing Print Media*. Migration Observatory Report, University of Oxford: COMPAS.

Allen W and Blinder S (2013) *Migration in the News: Portrayals of Immigrants, Migrants, Asylum Seekers and Refugees in National British Newspapers, 2010 to 2012*. Migration Observatory Report, University of Oxford: COMPAS.

Anderson B and Blinder S (2017) *Who Counts as a Migrant? Definitions and Their Consequences*. MIgration Observatory Briefing, Oxford: COMPAS, University of Oxford. Available from: http://www.migrationobservatory.ox.ac.uk/wp-content/uploads/2016/04/Briefing-Who_Counts_Migrant.pdf.

Baden C (2010) Contextualizing frames in political discourse: using semantic network analysis to investigate political parties' framing strategies in the Dutch EU referendum campaign. In: .

Baker P (2006) *Using Corpora in Discourse Analysis*. London: Bloomsbury Academic.

Baker P, Gabrielatos C and McEnery T (2013a) *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge University Press.

Baker P, Gabrielatos C and McEnery T (2013b) Sketching Muslims: a corpus driven analysis of representations around the word 'Muslim'in the British press 1998–2009. *Applied linguistics* 34(3): 255–278.

Balabanova E and Balch A (2010) Sending and receiving: The ethical framing of intra-EU migration in the European press. *European Journal of Communication* 25(4): 382–397.

Biber D and Reppen R (2015) Introduction. In: Biber D and Reppen R (eds), *The Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, pp. 1–8.

Blinder S (2015) Imagined Immigration: The Impact of Different Meanings of `Immigrants' in Public Opinion and Policy Debates in Britain. *Political Studies* 63(1): 80–100.

Blinder S (2016) *Migration to the UK: Asylum*. MIgration Observatory Briefing, Oxford: COMPAS, University of Oxford. Available from: http://www.migrationobservatory.ox.ac.uk/wp-content/uploads/2016/04/Briefing-Asylum.pdf.

Blinder S and Allen WL (2016) Constructing Immigrants: Portrayals of Migrant Groups in British National Newspapers, 2010–2012. *International Migration Review* 50(1): 3–40.

Brezina V, McEnery T and Wattam S (2015) Collocations in context. *International Journal of Corpus Linguistics* 20(2): 139–173.

Ceobanu AM and Escandell X (2010) Comparative analyses of public attitudes toward immigrants and immigration using multinational survey data: A review of theories and research. *Annual Review of Sociology* 36: 309–328.

Crawley H, McMahon S and Jones K (2016) *Victims and Villains: migrant voices in the British media*. Coventry University: Centre for Trust, Peace and Social Relations. Available from: http://www.migrantsrights.org.uk/files/news/Victims_and_Villains_Digital.pdf (accessed 13 March 2016).

de Zuniga HG, Correa T and Valenzuela S (2012) Selective Exposure to Cable News and Immigration in the U.S.: The Relationship Between FOX News, CNN, and Attitudes Toward Mexican Immigrants. *Journal of Broadcasting & Electronic Media* 56(4): 597–615.

Duffy B and Frere-Smith T (2014) *Perceptions and Reality: Public Attitudes to Immigration*. London: Ipsos MORI. Available from: http://www.ipsosmori.com/researchpublications/publications/1634/Perceptions-and-Reality-Publicattitudes-to-immigration.aspx.

Entman RM (2003) Cascading Activation: Contesting the White House's Frame After 9/11. *Political Communication* 20(4): 415–432.

Fotopoulos S and Kaimaklioti M (2016) Media discourse on the refugee crisis: on what have the Greek, German and British press focused? *European View*: 1–15.

Gabrielatos C (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.

Gabrielatos C and Baker P (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): 5–38.

Greenslade R (2005) *Seeking Scapegoats: The Coverage of Asylum in the UK Press*. London: Institute for Public Policy Reearch.

Grimmer J and Stewart BM (2011) Text as data: the promise and pitfalls of automatic content analysis methods for political texts.

Guo L, Vu HT and McCombs M (2012) An Expanded Perspective on Agenda-Setting Effects. Exploring the third level of agenda setting Una extensión de la perspectiva de los efectos de la Agenda Setting. Explorando. *Revista de Comunicación* 11: 51–68.

Hainmueller J and Hopkins DJ (2014) Public Attitudes Toward Immigration. *Annual Review of Political Science* 17(1): 225–249.

Hellsten I, Dawson J and Leydesdorff L (2010) Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science* 19(5): 590–608.

Herda D (2010) How Many Immigrants?: Foreign-Born Population Innumeracy in Europe. *Public Opinion Quarterly* 74(4): 674–695.

Jones RL and Carter RE (1959) Some Procedures for Estimating 'News Hole' in Content Analysis. *The Public Opinion Quarterly* 23(3): 399–403.

Jurafsky D and Martin J (2009) *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech*. Upper Saddle River, NJ: Prentice Hall.

Kilgarriff A, Baisa V, Bušta J, et al. (2014) The Sketch Engine: ten years on. *Lexicography* 1(1): 7–36.

Laver M, Benoit K and Garry J (2003) Extracting Policy Positions from Political Texts Using Words as Data. *The American Political Science Review* 97(2): 311–331.

Lehecka T (2015) Collocation and colligation. In: *Handbook of Pragmatics Online*, Benjamins.

Lim YS (2010) Semantic Web and Contextual Information: Semantic Network Analysis of Online Journalistic Texts. In: Breslin JG, Burg TN, Kim HG, et al. (eds), *Recent Trends and Developments in Social Software*, Lecture Notes in Computer Science, pp. 52–62. Available from: ://WOS:000285700200006.

Loughran T and McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66(1): 35–65.

Mahlberg M (2007) Clusters, key clusters and local textual functions in Dickens. *Corpora* 2(1): 1–31.

Marcus MP, Marcinkiewicz MA and Santorini B (1993) Building a large annotated corpus of English: the penn treebank. *Computational Linguistics* 19(2): 313–330.

Markaki Y and Vargas-Silva C (2016) *Long-Term International Migration Flows to and from the UK*. MIgration Observatory Briefing, Oxford: COMPAS, University of Oxford. Available from: http://www.migrationobservatory.ox.ac.uk/wp-content/uploads/2016/04/Briefing-LTIM_FLows_UK-1.pdf.

McCombs ME (2014) *Setting the Agenda: The Mass Media and Public Opinion*. 2nd Edition. Cambridge, England: Polity Press.

McEnery T (2015) Editorial. *Corpora* 10(1): 1–3.

McEnery T and Hardie A (2011) *Corpus linguistics: method, theory and practice*. Cambridge University Press.

McLaren L, Boomgaarden H and Vliegenthart R (2017) News Coverage and Public Concern about Immigration in Britain. *International Journal of Public Opinion Research*.

Pollach I (2011) Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis. *Organizational Research Methods* 15(2): 263–287.

Pomikálek J (2011) Removing Boilerplace and Duplicate Content from Web Corpora. Brno: Masaryk University.

Pottie-Sherman Y and Wilkes R (2017) Does Size Really Matter? On the Relationship between Immigrant Group Size and Anti-Immigrant Prejudice. *International Migration Review* 51(1): 218–250.

Rossignol M and Sebillot P (2005) Combining statistical data analysis techniques to extract topical keyword classes from corpora. *Intelligent Data Analysis* 9(1): 105–127.

Rychlý P (2008) A Lexicographer-Friendly Association Score. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing*, Brno: Masaryk University. Available from: http://nlp.fi.muni.cz/raslan/2008/raslan08.pdf#page=14 (accessed 3 August 2015).

Sketch Engine (2011) SiBol/Port corpus. Available from: https://www.sketchengine.co.uk/sibolport-corpus/ (accessed 13 March 2017).

Thomas J (2015) *Discovering English with Sketch Engine: A Corpus-Based Approach*. Versatile.

Threadgold T (2009) The Media and Migration in the United Kingdom, 1999 to 2009. In: Migration Policy Institute.

Tognini-Bonelli E (2001) *Corpus Linguistics at Work*. Netherlands: John Benjamins B.V.

van Dijk T (2008) *Discourse and Power*. New York: Palgrave MacMillan.

van Klingeren M, Boomgaarden HG, Vliegenthart R, et al. (2015) Real World is Not Enough: The Media as an Additional Source of Negative Attitudes Toward Immigration, Comparing Denmark and the Netherlands. *European Sociological Review* 31(3): 268–283.

Vliegenthart R and Walgrave S (2008) The Contingency of Intermedia Agenda Setting: A Longitudinal Study in Belgium. *Journalism & Mass Communication Quarterly* 85(4): 860–877.

Vollmer BA (2017) Security or insecurity? Representations of the UK border in public and policy discourses. *Mobilities*: 1–16.

Young L and Soroka S (2012) Affective News: the automated coding of sentiment in political texts. *Political Communication* 29(2): 205–231.