How are 'immigrant workers' represented in Korean news reporting?—A text mining approach to critical discourse analysis

Changsoo Lee
Hankuk University of Foreign Studies, Seoul, Republic of Korea

Abstract

The present study explores the usefulness of a text mining approach to investigating the representation of minorities in news reporting. The question is popular for scholars working in the realm of critical discourse analysis (CDA). Their typical approach is qualitative, which involves dissecting a small number of texts at the microlinguistic level. This approach has over the years come under severe criticisms for the lack of objective and reliable empirical evidence it produces for the sweeping claims it makes about the relationship between language and social structures. In response to such criticisms, the discipline has recently been embracing corpus linguistics as a way of making up for its methodological weaknesses. Yet, the corpus techniques applied to CDA have so far been largely confined to collocation and concordance analysis. Despite their proven value, these techniques are not optimal for investigating the overall topic structure and content of media texts because they are primarily designed to probe lexical relationships. Digital humanities have over the years developed a range of effective tools for analyzing the content of a large population of texts. The present study applies two such tools, co-occurrence analysis and topic modeling coupled with network analysis, to analyzing a corpus of Korean straight news reports. The analyses were effective in uncovering five major topics or themes underlying the news corpus, thereby proving the usefulness of the techniques used for the type of CDA questions explored in this article.

Correspondence:
Changsoo Lee, Hankuk, University of Foreign Studies, 107 Imun-ro, Dongdaemun-gu, Seoul 02450, Republic of Korea.
E-mail: dewywag@gmail.com

1 Introduction

The present article adopts a text mining approach to investigating the media representation of immigrant workers in Korea within the framework of critical discourse analysis (CDA), a field of study aimed at revealing the social structures of power, dominance, and inequality as manifested, legitimized, and/or challenged by language or discourse (Van Dijk, 1993: 249–50; Wodak and Meyer, 2009: 3). The article seeks to contribute to CDA and the digital humanities by first showcasing the relevance of state-of-the-art text mining techniques for a field of studies that has traditionally been dominated by qualitative research and second by testing if the findings of existing CDA studies on minorities

Digital Scholarship in the Humanities, Vol. 34, No. 1, 2019.
doi:10.1093/llc/fqy017 Advance Access published on 18 June 2018

hold good for Korea, a burgeoning multiethnic society with a strong background of ethnic homogeneity.

The media representation of ethnic minorities has been a popular topic in CDA since the publication of a series of pioneering studies in the late 1980 and early 1990s by Van Dijk (1987, 1991, 1993). Yet, the majority of these studies have been set in Western countries with long traditions of immigration such as the UK, France, and Australia [see references sorted by nation in KhosraviNik (105, ft.54: KhosraviNik, 2015)]. Whether their findings would be equally applicable to emerging multiethnic societies in Asia with different backgrounds remains largely unexplored. Korea is a case in point. It has traditionally prided itself on being an ethnically homogeneous nation (Shin, 2013: 369), and ethnic diversity is a recent phenomenon that has emerged since the mid-1990s when immigrant workers began flowing in from Southeast Asian nations in response to growing labor shortages in so-called 3D jobs (Dirty, Difficult, Dangerous) (Shin, 2013: 374). The nation now finds itself on the threshold of becoming a true multiethnic society with the share of immigrants in total population having more than doubled from 1.4 in 2005 to 3.2% in 2013 (Table 3, 91: Seol, 2015). Yet, no serious CDA study is reported that explores how immigrant workers are portrayed in Korean news media. The present study aims to fill this gap in the literature.

Another important motivation behind the study is to demonstrate the value of text mining techniques to exploring CDA questions, particularly when they are concerned with uncovering the underlying topical structure of textual data. In recent years, corpus linguistics has been incorporated into CDA as a means of compensating for its weak empirical foundation (Orpin, 2005; Baker, 2006; Baker *et al.* 2008; Mautner, 2009a, b). Text mining may also offer some useful techniques to expand the scope of computer-based analytical tools available for CDA to enrich its methodology. Text mining is generally defined as a knowledge-intensive process which employs a suite of computational tools to identify patterns in unstructured text data (Feldman and Sanger, 2007: 1). The present study introduces two text mining techniques for corpus analysis—co-occurrence network analysis and topic modeling. These techniques originate from text mining and are now popularly adopted in social network analysis and digital humanities (cf. Cho and Kim, 2015; Graham *et al.* 2016; Grayson *et al.* 2016). Of them, the relevance of topic modeling for CDA has been demonstrated recently by Törnberg and Törnberg (2016a, b) who analyzed Internet forum posts to investigate the representation of Muslim and Islam and the discursive connections between Islamophobia and anti-feminism. The present article employs co-occurrence network analysis in addition to topic modeling to demonstrate how they can expedite analysis of a more conventional type of text data familiar to CDA analysts—news texts.

In the following, CDA research on racism and the growing adoption of corpus linguistics for CDA studies will be briefly discussed as the background for the present study before the data and analytical methods are explained and the results of analyses are discussed.

2 CDA on Media Racism

In the literature, racism is generally defined as prejudice or discrimination based on the misguided assumption that individuals' birth and biological features such as skin color determine their identity in terms of races and ethnicities and govern their behavior (Ghurye, 2003: 15). Race and ethnicity, however, are not natural categories, as their boundaries and memberships are hard to define. Yet, these concepts are ideologically and politically manipulated to highlight differences as a basis for power, dominance, and control (Goldberg and Solomos, 2002: 3, 4).

Van Dijk (1991: 24–28; 2002: 145; 2005: 10), one of the pioneers in CDA, defines racism as a system of dominance, which is ideological and structural. Dominance principally stems from unequal access to social resources and brings about social inequality through practices and processes of exclusion and marginalization. CDA views discourse as one of these resources and contends that it plays a pivotal role in re/producing racist prejudices and

legitimizing resultant discriminatory and exclusionary practices (Reisigl and Wodak, 2001: 1; Van Dijk, 2002: 145). At the heart of the connection between racism and discourse is what Van Dijk (2002: 145) refers to as 'elite discourse', the opinions and beliefs disseminated by those in power who enjoy special access to, and control over, public discourse. The media is an important vehicle for the formation and sustainment of elite discourse (Van Dijk, 2002: 152). In most societies, citizens rely on the media to learn about minorities, and the media in turn depend on elite groups as sources of reference in reporting about these groups.

Media-mediated racism, as is the case with racism in other sectors, is based on the dichotomization of group identity between 'Us' versus 'Them' or 'ingroup' versus 'outgroup' (Van Dijk, 2000, 2002), which is also known as 'otherization' (Holliday *et al.* 2004: 3). Media representation of minorities is often geared toward negative representation of others by focusing on their problems such as crime, drugs, and violence. Even, news reports on seemingly neutral topics like immigrant housing and employment are cast in terms of a threat to 'Us' (Van Dijk, 2000: 38). For example, it has been observed that news on ethnic issues in Europe is typically limited to a small set of stereotypical topics such as illegality, demographic threats, cultural differences, crime, violence, and tensions (Van Dijk, 1993, 2012; Santa Ana, 2002; Richardson, 2004; KhosraviNik, 2014; Lirola, 2014; Banda and Mawadza, 2015). This is a part of the overall pattern in which topics of social discourse about immigrants and ethnic groups are skewed toward negative stereotypes (Van Dijk, 2004: 352).

These discussions inform us that modern racism in the media is covert, subtle, and implicit. It works mostly by denying racism on the surface and highlighting facts suggestive of problems, deficiencies, and differences in minority groups (Van Dijk, 2000: 33, 34; Simmons and Lecouteur, 2008: 667). If this is the case, then discovering major topics in the media coverage of ethnic minorities should be a critical part of CDA research on media racism. CDA research on racism, however, has been dominated by qualitative linguistic analysis focused on the interpretation of specific linguistic features rather than on the discovery of topic distribution in media coverage (cf. Teo, 2000; Flowerdew *et al.* 2002; Pietikainen, 2003; Liu and Mills, 2006; Simmons and Lecouteur, 2008; Belmonte *et al.* 2010; Van Dijk, 2012; Costelloe, 2014; Don and Lee, 2014; KhosraviNik, 2014; Banda and Mawadza, 2015; Burroughs, 2015). In the small number of studies that have attempted to identify underlying themes or topics, analyses have been done by manual coding and/or categorization (cf. Gale, 2004; Farquharson and Marjoribanks, 2006; Kim, 2012) in keeping with the traditional qualitative approach to media content analysis (cf. Riff *et al.* 2014; Drisko and Maschi, 2016; Neuendorf, 2017), rather than analyzing language itself.

3 CDA, Corpus Linguistics, and Text Mining

As noted above, CDA analysts have traditionally relied on qualitative analysis of texts. Typically, the analyst would choose a small number of texts such as newspaper articles and political speeches and subject them to microscopic analysis of various linguistic features, drawing extensively on popular concepts from diverse fields of linguistic inquiry such as systemic functional linguistics, cognitive linguistics, pragmatics, rhetoric, and so on (Van Dijk, 2002: 147; Fairclough, 2003: 191–194; Talbot, 2007: 45–7).

CDA's traditional approach to text analysis has over the years come under severe criticisms for a lack of methodological rigor and interpretation problems (Widdowson, 1995, 2004; Stubbs, 1997; O'Halloran, 2003; Bartlett, 2004; O'Halloran and Coffin, 2004). Limiting our discussion to the methodological side, the validity of CDA studies has been called into question with regard to generalizability, reliability, and representativeness for being based on a small number of texts arbitrarily selected or possibly cherry-picked to justify the researcher's preconceived notions. In this context, corpus linguistics has been suggested as a way of complementing for CDA's methodological weaknesses by providing an empirical basis for validating and reinforcing its findings (Orpin, 2005; Baker, 2006; Baker

et al. 2008; Mautner, 2009a, 2009b). Baker (2006: 10–17) mentions four advantages of using corpora for discourse analysis: (1) reduction of researcher bias, (2) 'cumulative effect' of repeated patterns in a corpora as evidence of a particular hegemonic discourse, (3) ease of finding counter examples or 'resistant discourse' to reveal a fuller range of discourse positions, and (4) easy combination with other methods of analysis to achieve 'triangulation'.

The standard procedure of investigating corpora for discourse analysis as illustrated by Baker (2006, 2014a, b) starts with determining a list of search terms relevant to a research question. For example, Baker (2014b) selected, on the basis of introspection and various reference sources, twelve terms including 'transgender', 'transsexual', and 'transvestite', to investigate the representation of trans people in British newspapers. As a second step, collocation analysis is carried out by extracting lists of collocates of the search terms from the corpus with the aid of a text analysis program. A collocate is a word which occurs within a certain range of the search term, for example, within five words to its left and right. The collocates are then examined to get an idea of how the search term is used in the corpus by consulting their textual environments through concordance searches. Finally, the results of collocation analysis are used as a basis for concluding how the subject in question is represented in the corpus.[1] Concordance and collocation analysis can be quite useful and fruitful as have been illustrated by many CDA studies that have adopted them. They can reveal, for example, that the word 'feminism' connotes 'homo/sexuality' in the British media, while it occurs more in the context of academic inquiry in the German press (Jaworska and Krishnamurthy, 2012) or that refugees are usually described in negative terms in British newspaper texts and UN reports (Baker and McEnery, 2005).

Despite the proven usefulness of these corpus methods, there may be situations where other text analytic techniques could be more useful and offer fresh insights in investigating CDA questions, particularly when the researcher is more interested in the overall topic structure of his/her corpus than a set of keywords and their lexical associations within limited spans of text. There are cases where concordance or frequency analysis was used to discover major themes or 'corpus topics' in the media representation of ethnic minorities (cf. Baker and McEnery, 2005; Baker *et al*. 2013), but categorization was done manually, and there is no guarantee that the analysis covered the whole topics. The same task could be handled more efficiently with improved reliability by using state-of-the-art text mining techniques that automate the process. In this study, two such techniques are introduced to summarize topics in our data—co-occurrence network analysis and topic modeling.

The two techniques share some commonalities with collocation-based corpus analysis in that they all deal with word co-occurrences. They differ, however, in two notable ways as explained by Murakami *et al* (2017). First, collocation looks at co-occurrences within limited spans of text as noted earlier (five words to the right or left of the search term or node word), while text mining techniques cover co-occurrences in larger units of text such as sentences, paragraphs, or entire texts. Second, in collocation analysis, co-occurrences are confined to the node words selected by the analyst, whereas the text mining methods cover co-occurrences among all words within the entire corpus. As a result, text mining is better suited for uncovering the thematic structure of a corpus, which enables a 'global monitoring of the entire system' (Fortuna et al. 2009: 27), while collocation analysis is more adept at capturing prosodic meaning at phraseological level.

4 Data and Analytic Techniques Used

The data used in the current research consist of 404 straight factual news reports collected from a public news archive website (https:// www. kinds. or. kr/) for the entire year of 2013. The website maintains archives of articles (reports, editorials, commentaries, etc.) from major nationwide and provincial news outlets. The corpus, as shown in Table 1, represents forty-eight news outlets, including broadcast and print, online and offline, and provincial and nationwide news media. The number of articles

Table 1 Composition of the news corpus

Type/number	Name (number of articles in the corpus)
Broadcast (nationwide)/3	KBS (15), MBC (9), SBS (7)
Broadcast (provincial)/1	KNN (2)
Newspaper (nationwide)/14	*Asia Today* (2), *Donga Daily* (2), *Financial News* (18), *Hankook Economy* (6), *Hankook Daily* (7), *Herald Economy* (16), *Kookmin Daily* (12), *Kyonghyang Newspaper* (3), *Hankyoreh* (4), *Maeil Economy* (12), *Moonwha Daily* (9), *Sekye Daily* (9) *Seoul Economy* (7), *Seoul Newspaper* (6)
Newspaper (provincial)/27	*Boosan Daily* (13), *Choogbook Daily* (10), *Choongchung Today* (21), *Daejun Daily* (16), *Hanla Daily* (4), *Hongsung Newspaper* (2), *Inchun Daily* (13), *Jeymin Daily* (6), *Joongdo Daily* (13), *Joongboo Daily* (9), *Junbook Daily* (4), *Junbook Domin Daily* (6), *Junnam Daily* (4), *Kangwon Daily* (10), *Kangwon Domin Daily* (3), *Kimpo News* (3), *Kookje Newspaper* (11), *Kyonggi Daily* (10), *Kyongin Daily* (17), *Kyongnam Newspaper* (12), *Kyongnam Domin Daily* (17), *Kyongsang Daily* (3), *Kwangjoo Daily* (11), *Moodung Daily* (3), *New Junbook Newspaper* (4), *Okchun Newspaper* (1), *Youngnam Daily* (14)
Online/3	*E-Daily* (4), *E-Today* (11), *Prime Economy* (3)

they individually contribute to the corpus ranges one to twenty one.

Two Korean noun phrases, *oykwukin kunloca* ('foreign worker') and *icwu notongca* ('immigrant laborer'), were used as search words. Each news report searched was scanned to ensure that it dealt with foreign/immigrant workers (shortened to 'immigrant workers' hereafter) as the main or a part of the main topic. The news reports vary in length, ranging from a single to several paragraphs. The corpus in total amounts to 21,652 ejels. Ejels are a lexical unit unique to the Korean language. They are separated by spaces like words in English, but ejels in Korean are complex morphological units made up of two words conjoined or a word with particles or endings attached to it, which would be denoted by separate words in English. For this reason, Korean and English corpora cannot be compared simply in token counts. By a rule of thumb, a Korean ejel would be equal to two to three words in English, which can be used as a rough basis for assessing the size of our corpus.

The corpus compiled was loaded into R, an open-source statistical computing environment, and was preprocessed to extract only nouns from the documents after eliminating numbers, punctuation marks, and other unnecessary characters or symbols. Various packages designed for natural language and text processing were used in this process, including KoNLP (for part-of-speech tagging and extracting nouns from Korean texts) and tm (for preprocessing and creating a term–document matrix). Keeping only nouns and verbs in documents is a common practice for analyzing the content of a collection of documents. In English, stop words are used to remove grammatical/function words. This approach, however, is not feasible for Korean because Korean equivalents of function words take the form of particles attached to nouns and verbs. Part-of-speech analysis, therefore, is essential to identify and remove them. Incidentally, the nouns extracted in this process include what are actually verbs, like *chamka* ('participation') and *woonyeng* ('operation'). In the original unprocessed texts, these words mostly occur conjoined with the particle *-hata* ('do') which turns them into verbs. This makes noun extraction in Korean a highly effective method for distilling documents to words with high information load.

Three analyses were carried out on the noun-only corpus—collocation analysis, co-occurrence-based network analysis, and topic modeling. Collocation analysis is not part of the text mining toolset, but it is included in the current research to illustrate how effective it could be in discovering underlying topics compared with the other two techniques. Co-occurrence analysis and topic modeling were carried out with R, and Gephi was used to visualize the results as networks. Each of these techniques and the specifics of their implementation in the current research will be explained in full in Sections 5.1, 5.2, and 5.3, respectively.

5 Results

5.1 Collocation analysis

The collocation analysis reported here was carried out, using Wordsmith Tools, Version 7 (Scott, 2017). The noun-only corpus generated in R was exported to a text file, which was then loaded into the text analysis software. Using the Concordance Tool, collocates of the keyword *oykwukin* ('foreigner') were extracted for the span of five words to the left and right of the search word. Top fifty collocates are provided in Table 2 in the descending order of frequency. Since the news reports were merged into a single document, and it was already stripped of function words, frequency was considered as the basis for computing significant collocates instead of using statistical measures available in the software such as MI or T statistics.

To begin with, we exclude from consideration some obvious collocates such as 'worker', 'resident', and 'laborer', which top the list. They come right after the search word 'foreigner' to form reference terms like 'foreigner worker', 'foreigner resident', and 'foreign laborer'. The term 'multicultural' also occurs mostly paired with words like 'center', 'family', and 'children' to form fixed phrases. With these terms out of the way, we scan the list to see if the remaining words can be grouped into some meaningful themes. The words, 'crime', 'police', 'anticrime unit', '(criminal) charge', and 'crime prevention' seem to form a thematic cluster. They suggest foreign workers are frequently mentioned in the context of crime. Let us label this group CRIME. There appears to be an EVENTS theme as well, as suggested by the terms, 'event', 'attend', and 'hold (event)'. This indicates that foreign workers are regularly described in connection with events, presumably as participants rather than as organizers, a point verified by separate searches of collocates of the two verbal terms 'attend' and 'hold (event)'. The verbal term 'conduct' may be linked to 'survey' or 'education'. A separate search of 'conduct' indeed proved 'education' to be its top collocate, but 'survey' was not on its collocate list. Some terms in Table 1 are too general to be pinned to a particular theme. Take 'increase' for example. It may go with 'crime', 'support', 'population', 'employ (ment)', and 'marriage', anything that can be described as going up in quantity. This suggests danger in relying on our intuition to draw links among the terms in Table 1. To avoid this danger, further investigation of the terms will be necessary, by examining their own collocates and also checking their contexts in concordances. This is not a very effective way of finding underlying topics in our corpus. More importantly, the fact that all examination and classification is done by the human analyst makes the process vulnerable to subjectivity and prone to errors in case she/he misses important collocation links. Also, there is no guarantee that the process will yield a complete list of underlying topics. We will go no further with the collocation analysis because it is not the focus of our analysis. Our exercise to this point is sufficient to give us an idea of how the process will go as it is applied to finding underlying topics in our corpus.

5.2 Co-occurrence analysis

Co-occurrence is similar to collocation in the sense that it looks into relationships between words. The

Table 2 Top fifty collocates of *oykwukin* ('foreigner')[3]

근로자 **Worker**	다문화 **Multicultural**	국적 **Nationality**	도내 **Provincial**	운영 **Operate**
주민 Resident	센터 Center	유학생 Foreign student	결과 Outcome	제주 JEJU
노동자 Laborer	지원 Support	전체 Total	현황 Status	인구 Population
범죄 Crime	체류 Stay	여성 Women	개최 Hold (event)	사업장 Workplace
거주 Reside	가족 Family	경기 Kyunggi	불법체류 Illegal Stay	하기 Summer
대상 Target	이민 Immigration	등록 Register	교육 Education	한국 Korean
고용 Employ	행사 Event	조사 Survey	전국 Nationwide	생활 Life
경찰 Police	방범대 Anticrime unit	참석 Attend	혐의 (Criminal) Charge	개국 Number of nations
증가 Increase	자율 Autonomous	부산 Pusan	기업 Enterprise	실시 Conduct (program)
결혼 Marriage	자녀 Children	활동 Activity	올해 This-year	범죄예방 Crime prevention

difference is that it computes relationships for all terms in a document instead of limiting them to those linked to a specific keyword. That way, co-occurrence tells us not just what terms a certain keyword hangs together with but how all major terms in a document are associated with one another, which makes it an ideal data format for exploring the overall semantic structure of any given text collection. As such, co-occurrence is the 'core functionality' of any text mining system (Feldman and Sanger, 2007: 9).

Co-occurrence can be formally defined as 'the relationship between pairs of terms occurring within a constant-sized context window' (Grayson *et al*., 2016: 67). Here, the window could be any length of text—a certain number of words, a sentence, a paragraph, or a whole document. In the current analysis, documents were used as the unit for computing co-occurrence pairs. The decision depends on how many topics a document is expected to have. For novels, it is customary to segment the text into smaller chunks like 1,000 words to capture transient topics that phrase in and out as the story unfolds (Jockers, 2013: 134). In contrast, news reports are invariably constructed around a single topic. Therefore, analyzing co-occurrence for the entire document is a more sensible approach.

Since we are dealing with a large number of words cross-related across multiple documents, the relationships are best expressed as a matrix. In R, there are several ways to construct a co-occurrence matrix from a collection of documents. In the current study, the text mining package 'tm' was used to extract a term–document matrix from our noun-only corpus, and a simple script was used to transform it into a term-to-term co-occurrence matrix, where the score in each cell represents the frequency with which the column and row terms co-occur in the same document across the entire corpus. Table 3 shows a matrix of the top five co-occurring terms in our news corpus.

The actual co-occurrence table is huge with 3,442 rows and columns respectively, matching the number of terms in our corpus. For practical reasons, the table is pruned to the top 100 terms. The best way to represent the crisscrossing relationships among these terms is by showing them as a network. Co-occurrence-based network analysis has recently flourished in conjunction with social network analysis, for example, for analyzing a network of people mentioned in news articles (cf. Danowski and Cepela, 2010; Liu *et al*, 2012), and it is being adopted by linguistic studies, for example, to find key sentences (Cho and Kim, 2015) or word associations (Tanev, 2014). In a network analysis, each term in our co-occurrence matrix will become a node, and the co-occurring relationships among them will be represented by edges connecting them. Additionally, the nodes can be grouped into clusters or 'communities', which can be used a basis for uncovering topics. Figure 1 shows a network representation of the 100-term matrix in our news collection.

The network was drawn with Gephi, an open-source network visualization program.[2] Edge thickness matches the frequency of the pair. Node size shows the popularity of the term, that is, how many other nodes are linked to it. The network has one center node, which is our search term, 'foreign(er)'. Numerous links extend out from the center node in radial fashion, linking it to other terms directly or indirectly. The different colors for the nodes indicate the five communities identified by the software's modularity statistics (at the resolution of 0.8), a measure of the strength of division of a network into modules or clusters. The original network had its nodes labeled in Korean. The English glosses were added later using a photo editor.

Table 3 Top five co-occurring terms

	외국 **Foreign**	경찰 **Police**	결혼 **Marriage**	국내 **Domestic**	여성 **Female**
외국 Foreign	277	36	28	35	25
경찰 Police	36	56	3	6	9
결혼 Marriage	28	3	45	10	17
국내 Domestic	35	6	10	42	7
여성 Female	25	9	17	7	42

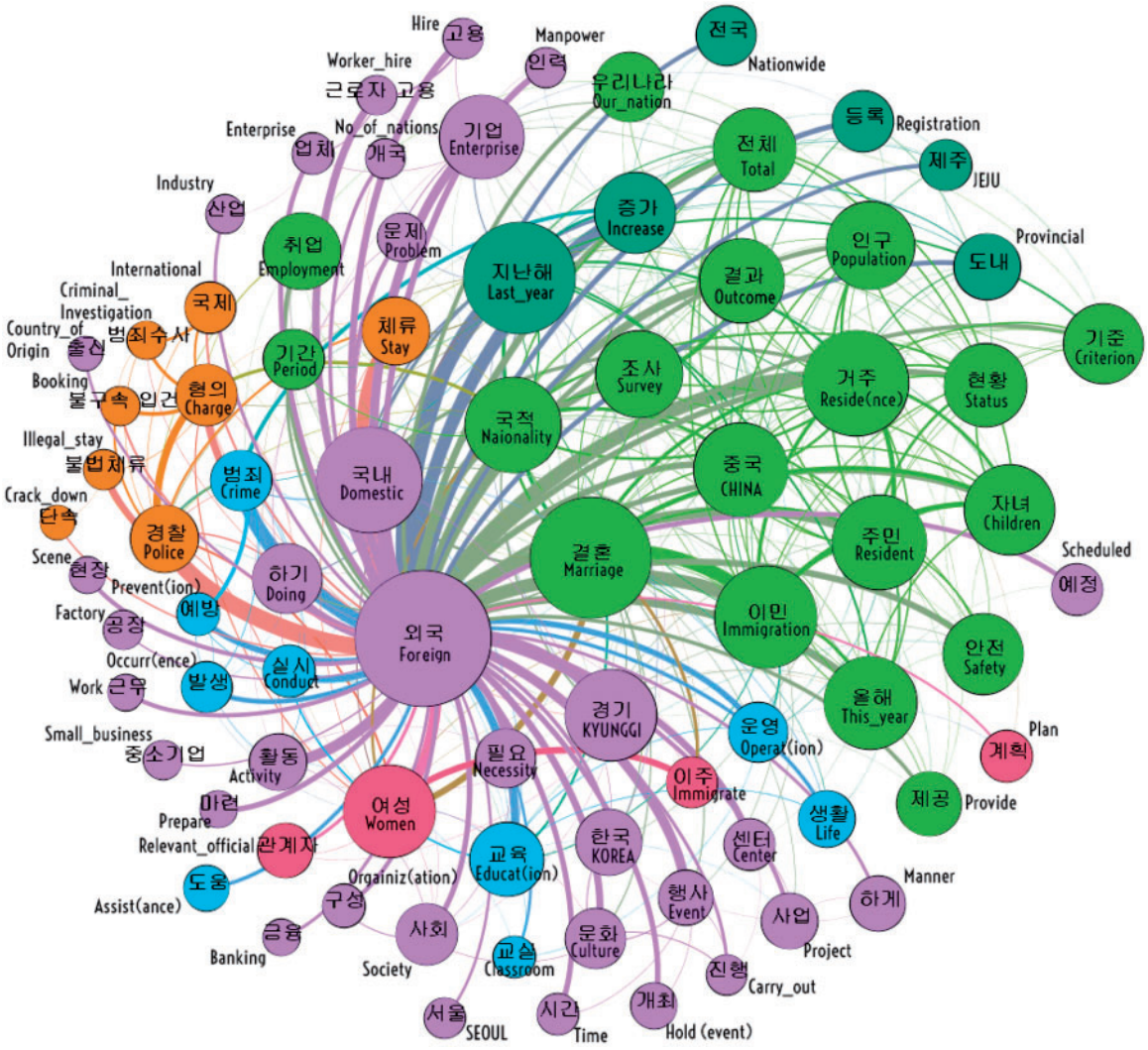

Fig. 1 A network representation of top 100 co-occurring terms

Let us examine each community to see if the membership constitutes a coherent topic. If we divide the network into four planes, the upper right part is occupied by the light green cluster. The membership consists of terms like 'population', 'nationality', 'immigration', 'resident', 'children', 'status', 'survey', and 'outcome'. These are terms used for reporting about the outcome of a demographic survey. The temporal terms 'period' and 'this year', the numerical term, 'total', and the comparative term, 'criterion' (which is actually used to mean 'as of a certain point of time') also belong to the language of reporting year-to-year statistics. This community suggests that in Korean news media foreigners are frequently mentioned in terms of their numbers. The dark green community adjacent to the light green one is made up of six terms, 'last year', 'increase', 'nationwide', 'register (ation)', 'provincial', and 'JEJU' (the biggest island off the southern tip of the peninsular). The relatively thick edges linking these terms to those in the light green community indicate the two communities are

closely interrelated. They all relate to survey statistics. Let us call these two communities together STATISTICS.

Moving to the upper left side of the network, the brown community along the edge of the ball-shaped network is made up of terms like 'illegal stay', 'police', 'investigation', '(criminal) charge', and 'booking'. These terms indicate anticrime police work. The term 'charge' is directly linked to the term 'crime' in the adjacent sky-blue community. Although the two terms belong to different communities, the thick edge linking the two suggests strong association, supporting the interpretation that police work here is concerned with crime. The other terms in the sky-blue community, 'occurrence', 'prevention', 'education', 'class', 'operate', and 'conduct', suggest the planning and implementation of preventive activities against crime. Based on this analysis, we can conclude that the brown and sky-blue communities represent two sides of the same coin. Both relate to crime, one focusing on fighting it and the other on preventing it. They indicate that Korean news media frequently focus on immigrant workers' association with crime. Let us name these communities together CRIME.

This leaves us with two remaining communities to be analyzed. The red one, composed of just four terms, does not appear to represent any important topic, other than that 'immigrant' and 'women' co-occur frequently, referring to the large number of foreign women who have married Korean nationals. Moving on to the final purple community that spans top to bottom on the left side of the network, the upper part of it is made up of terms like 'industry', 'business', 'factory', 'small-and-medium enterprises', 'work', 'manpower', 'problem', and 'hiring'. They point to manpower problems faced by Korean manufacturing firms and reflect the situation where foreign workers are imported to ease these problems. The lower part of the community is populated with terms like 'society', 'culture', 'activities', 'event', 'project', 'preparation', 'hold (event), and 'stage (event)'. They are terms used for reporting on events and activities, organized by Korean hosts for the immigrant community as was discussed in the foregoing collocation analysis. Consequently, the purple community can be divided into two sub-groups. They supply us with two more themes that figure importantly in Korean news about foreigners, MANPOWER, and EVENTS.

In summary, our co-occurrence-based network analysis has revealed four major angles in news coverage. Foreign workers in Korean news media are represented predominantly as statistics with the focus on their increasing numbers, as crime-prone people, as a solution to manpower shortages in manufacturing and as the target of various social activities and events. The analysis confirms the two topics, CRIME and EVENTS, that we identified in the earlier collocation analysis but adds two new significant ones, suggesting co-occurrence-based analysis is more thorough for discovering the thematic structure of a large corpus, in addition to being more efficient and less at risk of subjectivity and errors.

5.3 Topic modeling (latent Dirichlet allocation) analysis

Topic models are probabilistic models for automatically extracting latent topics from a large collection of documents based on a hierarchical Bayesian analysis. Latent Dirichlet allocation (LDA) is the simplest, yet most widely used form of topic modeling. The model assumes that there are a fixed number of underlying topics in a corpus, and each document is composed of these topics to varying degree. Each topic is represented by a different set of terms with high probability of co-occurring with that topic. LDA computes and presents these sets of terms for us (Blei *et al.* 2003; Blei, 2012). In actual practice, we ask the algorithm for a certain number of topics, five for example, and it will provide us with five lists of terms. It is up to us, the human analyst, to interpret these lists and identify topics.

Topic modeling in the current analysis was carried out by using the 'lda' package in R. The package implements LDA by using a Bayesian Gibbs sampler. The trickiest part of topic modeling is deciding on the number of topics we ask for. There are some statistical methods of determining this, like perplexity, which is a measure of the model's fit for data, but having the right number of topics in a mathematical sense does not guarantee the topics are more interpretable (Jacobi *et al*, 2015: 7). Therefore, it is

customary to produce models for a range of topic numbers and choose the one that offers the most coherent topic structure. In the current analysis, models were generated for topic numbers ranging from ten to three. A careful examination of the topic lists from each run led to the conclusion that five topics best summarize the thematic content of our corpus. Table 4 shows top twenty terms our topic model assigned to the five topics we asked for. Let us find out how interpretable the lists are.

Looking at Topic 1, we can first identify terms used as referents. 'Multicultural' typically pairs with 'family', and 'foreign' combines with 'worker' to form 'multicultural families' and 'foreign workers'. 'Children', 'immigrant', and 'woman' also refer to the membership of these minority groups. Aside

Table 4 Top twenty terms for five topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	다문화	외국	외국	외국	경찰
	Multicultural	Foreign	Foreign	Foreign	Police
2	외국	주민	근로자	근로자	여성
	Foreign	Resident	Worker1	Worker1	Woman
3	근로자	외국인	고용	범죄	혐의
	Worker1	Foreigner	Hire1	Crime	Charge
4	지원	지난해	국내	경찰	외국
	Support	Last_Year	Domestic	Police	Foreign
5	가족	제주	인력	노동자	불법체류
	Family1	Jeju	Manpower	Worker2	Illegal_Stay
6	센터	거주	노동자	대상	공장
	Center	Reside	Worker2	Target	Factory
7	행사	근로자	중소기업	자율	노동자
	Event1	Worker1	Small business	Autonomous	Worker2
8	결혼	등록	기업	활동	근로자
	Marriage1	Registration	Enterprise1	Activity	Worker1
9	가정	전체	이주노동자	외국인	차량
	Family2	Total	Immigrant_Worker	Foreigner	Vehicle
10	이주민	국적	양산	음성	단속
	Immigrant	Nationality	Yangsan	Umsung	Crack_Down
11	사회	증가	업체	지역	부산
	Society	Increase	Enterprise2	Region	Pusan
12	이민	마약	채용	범죄예방	출신
	Immigration1	Drugs	Hire2	Crime prevention	Country_of_orgin
13	여성	중국	필요	교육	국적
	Woman	China	Necessity	Education	Nationality
14	교육	결혼	지역	방범대	이주
	Education	Marriage1	Region	Anticrime_Unit	Immigration2
15	사업	올해	언어	예방	결혼
	Project	This_Year	Language	Prevention	Marriage1
16	대상	포함	정책	체류	출입국관리
	Target	Include	Policy	Stay	Immigration_Control
17	문화	국내	한국어	증가	사진
	Culture	Domestic	Korean_Language	Increase	Photograph
18	거주	이민	개선	실시	사무
	Reside	Immigration1	Improvement	Conduct (Program)	Clerical_Work
19	자녀	전국	조사	치안	피해자
	Children	Nationwide	Survey	Security	Victim
20	개최	인구	결과	지난해	외국인
	Hold (event)	Population	Outcome	Last_Year	Foreigner

Note: In descending order of term–topic co-occurrence probability.

from them, we have activity terms like 'education', 'project', and 'event', which relate to the verb 'hold', indicating certain events and activities being organized and offered. The key term linking the two groups of referent and activity terms is '대상 (target)'. Functionally, it is equivalent to 'for' as in 'hold a concert for children'. It specifies for whom something is planned and organized. Finally, we have the term, 'support'. Apparently, the terms under Topic 1 point to events and activities offered for the benefit of, and to support, foreign workers and their families. The EVENTS theme we identified in the foregoing co-occurrence analysis can be subsumed under a new topic, which we will label SUPPORT.

Moving to Topic 2, aside from the usual referential terms, the topic is built around terms used for describing demographic surveys. 'Population', 'survey', 'reside', and '(registered) resident' directly point to this. The temporal referents, 'this year' and 'last year', the quantity-related terms, 'total' and 'increase', and the terms of geographical scope, 'nationwide' and 'domestic', are also part of the typical language for describing statistical trends in a nation. These terms are brought together under 'nationality', which relates to 'foreign', 'immigration (immigrant)', and 'China'. Conclusively, Topic 2 is composed of terms for discussing increasing trends in the population of immigrant workers. This relates to the STATISTICS community we identified in the earlier co-occurrence-based network analysis. Let us keep the same label for this topic. Incidentally, the term 'drugs' appears to be irrelevant to this topical theme. Terms like this are called 'invaders', and it is not rare that topic lists include a few invaders.

Topic 3 is distinguished from other topics with its focus on manpower and employment. In addition to what appears to be the key term for this topic, 'manpower', we have two variant forms of 'hire' and three terms referring to businesses, 'enterprise1', 'enterprise2', and 'small business'. The last term suggests immigrant workers are mostly hired by small businesses in Korea. The other terms like 'policy', 'necessity', 'improvement', and 'language' appear to point to certain problems connected with the use of immigrant manpower, for which policy measures are needed in order for them to be addressed. In a nutshell, Topic 3 can be labeled MANPOWER, which is also the name we gave to one of the communities found in the earlier co-occurrence analysis.

Now, let us skip Topic 4 momentarily and go to Topic 5. A quick glance through the terms here tells us that it concerns criminal activities and law enforcement by police, as indicated by terms like 'police', '(criminal) charges', 'crackdown', and 'victim'. The term 'immigration control' refers to immigration authority, and it relates to 'illegal stay' in the context of efforts to control 'illegal' foreign workers. The fact that 'nationality' and 'country of origin' are included on the list shows that news reports are interested in articulating the nationality of immigrant workers implicated in criminal and illegal activities. One interesting inclusion is 'photograph'. A concordance search of this term showed that it mostly occurs in situations where foreigners are caught taking sneak photos of women in swimsuits at beaches, and 'Pusan' is where this happens a lot. Based on this interpretation, let us label Topic 5 CRIME.

Finally, Topic 4 also appears to be concerned about crime, but terms like 'conduct', 'education', 'prevention', and 'crime prevention' indicate the focus is on preventing crime. The two terms, 'autonomous' and 'anticrime unit' actually go together as 'autonomous anticrime unit', and it refers to a patrol unit composed of voluntary foreign residents. Topic 4 also includes terms related to hazards and safety (presumably at the workplaces employing foreign workers), though they are not among the top twenty terms. In this sense, the term, 'security', appears to best sum up Topic 4. Hence, let us label Topic 4 SECURITY.

In general, Table 3 is highly interpretable with clear distinctions among the topics. So far, we reviewed only top twenty terms for each topic, but it may be necessary to examine a larger number of terms to get a more concrete understanding of the topic structure of our corpus. But examining sixty or seventy terms in the way we did with Table 3 may be impractical. Instead, we can visualize them in a network for a bird's eye view of the entire topic structure. Figure 2 plots top seventy terms for each topic in our model as a network, which was

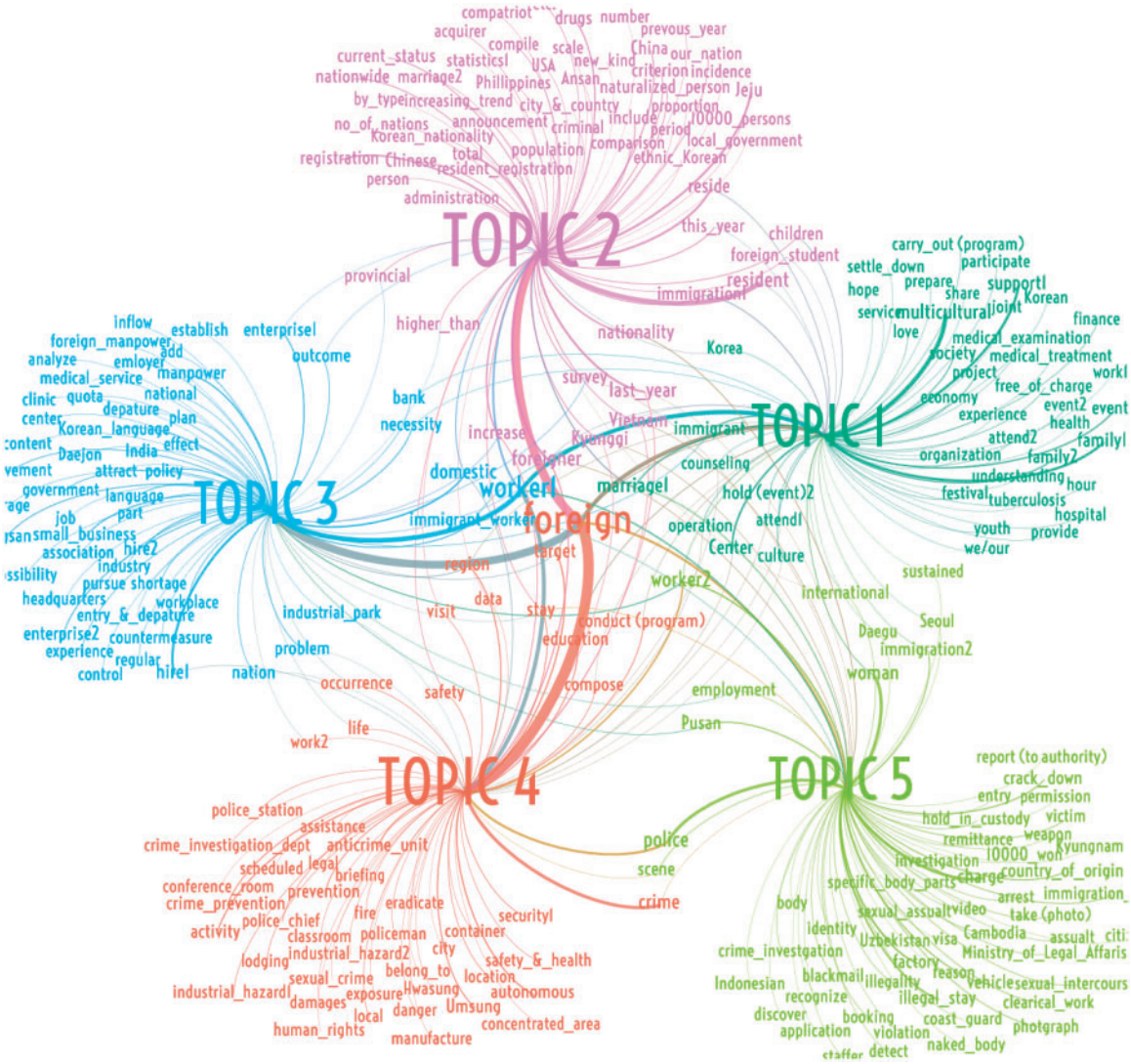

Fig. 2 A network representation of five topics

created with Gephi. The original Korean node labels are replaced by English glosses.

The network has five center nodes corresponding to our five topics, which are anchored by our keywords, 'foreign' and 'worker' in the center. The thickness of edges indicates the co-occurrence probabilities of individual terms with individual topics. The size of the topic nodes is indexical of their relative weights in the overall topical structure. It shows that the five topics occupy almost equal shares in our corpus, though Topic 2 (STATISTICS) enjoys a slightly higher level of exposure than the others.

Without rehashing the observations made about Table 3, let us see how considering a larger number of terms can add clarity to our interpretation. Starting with Topic 1 (SUPPORT), the term cluster now includes words like medical 'examination' and 'treatment', 'service', 'festival', and 'experience (activities)', which specify what kind of services and activities are offered, and they are 'free of

charge'. These terms reinforce our interpretation of this topic as structured around support for immigrant workers. Additionally, terms like 'hope', 'love', and 'settle down' relate to the motivations of support, namely, to help immigrant workers settle down, to give them hope, and show them love, as were verified by searching their concordance lines. Topic 2 (STATISTICS) now shows two more terms of quantity increase, 'increasing trend' and 'higher than', on top of 'increasing' from Table 3. This shows that growing numbers are indeed a major concern in reporting on the population of immigrant workers. Among the new terms showing under Topic 3 (MANPOWER) are 'shortage', 'problem', 'countermeasure', and 'attract'. Together with the terms, 'manpower', 'enterprise', 'small business', and 'policy' from Table 3, they clarify that the topic is indeed about addressing the problem of manpower shortages in industry by attracting foreign workers. Topic 4 (SECURITY) newly includes words like 'concentrated area' and 'location', which point to places where crime prevention efforts are concentrated. The two places appearing under Topic 4, 'Umsung (County)' and 'Hwasung (City)', are home to the nation's large concentrations of immigrant workers. The fact that these two places are associated with SECURITY among other possible topics reflects negative biases toward the immigrant community as a whole. Besides this, the terms, 'police station' and 'police chief', indicate the leading role of police in crime prevention efforts, and 'fire', 'industrial hazard', and 'safety and health' suggest that safety is also another important dimension to police-led preventive measures. Finally, Topic 5 (CRIME) now includes new crime-control terms such as 'investigation', 'arrest', 'discover', 'detect', 'book(ing)', and 'hold in custody'. Terms like 'blackmail', 'assault', 'sexual assault', and 'weapon' specify crime types. The multiple body references, 'body', 'specific body parts', and 'naked body' relate to 'photograph' and 'take (photo)' and 'video' in the context where immigrant workers are often caught secretly photographing and videotaping women in exposed clothing. 'Cambodian', 'Uzbekistan', and 'Indonesian' are nationalities frequently associated with this and other crimes in the news. Finally, the term 'illegality', along with 'illegal stay' and 'immigration control' from Table 3, shows that control of illegal foreign workers is a major element of this topic. Overall, considering larger number of terms not only improves the interpretability of each topic but it also allows a more richly textured reading of each topic.

5.4 Discussion of analysis results in CDA terms

As the final step of our analysis, let us try to relate the findings of our analyses to racism in CDA terms. We will focus on the results of topic modeling because they are inclusive of the results of the other two analyses. First of all, it is important to note that the news reports in our corpus reflect ethnic Koreans' perspectives. The evidence for this comes from two terms included in our term clusters in Figure 2. They are *wuri* ('we/our') and *wurinala* ('our nation') from Topic 1 (SUPPORT) and 2 (STATISTICS), respectively. A concordance search of the Korean first-person plural pronoun *wuri* ('we/our') showed that it mainly co-occurs with 'society', a term also found in the Topic 1 cluster, to form the phrase 'our society'. It serves to locate the recipients of support and assistance described under Topic 1 as immigrants in 'our society'. The term *wurinala* ('our nation') similarly occurs in the context of reporting about the population of immigrants in 'our nation'. The use of this first-person plural pronoun in news reporting indicates that the news writer treats immigrants as outside members of the in-group of 'we' Koreans as if they were guests. What this means is that the topic structure of our news corpus identified above is a synopsis of the aspects of immigrants' lives that 'we' Korean news writers are interested in looking into.

With this point in mind, we first find that some of the typical topics known to dominate news on ethnic groups in Europe are also found in Korean news. In Section 2, it was noted that news on ethnic minorities in Europe tended to focus on negative topics such as illegality, demographic threats, cultural differences, crime, violence, and tensions. Additionally, KhosraviNik *et al.* (2012) found numbers, economic burdens, threat, danger, and law to be among the major argumentative strategies used

by British newspapers in describing RASIM (refugees, asylum seekers, and immigrants). Similarly, Baker and McEnery (2005) examined concordances of 'refugee' in British newspapers and found that they were often described by using descriptive words related to quantification, movement, tragedy, help, and crime. STATISTICS, SECURITY, and CRIME in our topic model show that many of these topics are also prominent with Korean media. The topic STATISTICS indicates immigrants in Korea are often described as numbers, and SECURITY and CRIME prove that they are strongly associated with illegality and crime.

Our analysis also reveals what appears to be positive angles on immigrants. The MANPOWER topic reflects appreciation of immigrants as contributors to the national economy, and the SUPPORT topic manifests humanitarian care for immigrants. Yet, these, too, can be a potential source of racist discourse. The MANPOWER topic reveals an exploitive attitude of assessing the value of immigrants in economic terms, which relates to the angle of usefulness mentioned by Reisigl and Wodak (2001: 74–81). As with numbers or quantification, this attitude is potentially racist as it dehumanizes and objectifies immigrants as a kind of commodity (KhosraviNik *et al*. 2012: 289). The SUPPORT topic is particularly interesting because it appears unique to the Korean news media. Above, Baker and McEnery (2005: 207) found 'official help' as an important theme of reporting on refugees, but the level of interest shown by Korean media in keeping track of the events, activities, and services being offered to immigrants by a variety of social groups has not been reported in the literature. Nevertheless, this topic, too, has a strong racist slant to it. Analyzed in terms of 'we' versus 'they, the news reports about support and assistance are, in essence, accounts of what 'we' Koreans are doing for 'them' immigrants. Here, 'we' are positively represented as 'philanthropists', while 'they' are portrayed as people in need and incapable of surviving without 'our' care and support. This interpretation is supported by the strict division of roles in reported events and activities. Events, activities, and services require two major actors, those who organize and offer them and those who attend and benefit from them. Concordance searches show that the agents of such verbal terms as 'hold (event), 'carry out (program)', 'provide', and 'prepare' are invariably ethnic Korean groups, and that immigrants are always confined to passive acts like 'attend', 'participate', and 'experience'.

6 Conclusion

The analyses in the previous section led us to the following findings. First, among the three text analytic techniques used, topic modeling, aided by network visualization, was most effective and thorough in uncovering the underlying topic structure of our news corpus. It not only discovered more topics automatically but the terms grouped under each topic allowed detailed analysis of the content of each topic. Second, the topic structure discovered both at global and sub-topical levels were immediately relevant to CDA research on media racism. It showed many of the racist topics used by the press in Europe such as demographic figures and crime were equally prominent in Korean media. Third, the analyses also enabled us to discover a topic unique to Korean media—support and assistance to the immigrant community. The topic, however, was analyzed as equally racist in its implications, as it serves the positive self-representation of 'us' Koreans while promoting the negative image of immigrants as people at the disposal of 'our' kindness.

The findings above support the relevance of text mining techniques for CDA studies. Admittedly, CDA's concerns go far beyond discovering the latent topic structure of corpora. Yet, content analysis has long been employed for CDA studies as a way of investigating the textual representation of various social groups like immigrants, homosexuals, women, feminists, and so on (Ashley and Olson, 1998; Lilleker and Jackson, 2011; Mendes, 2011; Kim, 2012; Jones, 2015). Computational techniques promise to automate a large portion of manual work involved in such projects, thereby freeing the analyst from the limitation of having to work with small samples of text, and produce reliable and reproducible results, which is always a challenge with human coders (Leetaru, 2012: 2–3). Additionally, text mining allows the researcher to explore and

discover new insights, rather than being limited by a pre-defined list of analytic categories (Waldherr *et al*., 2016: 213). The relevance of text mining techniques for CDA will only increase in the future as the scale of text collections to be processed for analysis keeps growing with the flourishing of online media such as social networks, blogs, and online news websites, in addition to the vast amounts of print texts being digitalized including books, journals, notes, and so on (Wiedemann, 2016: 8).

Notes

1. There are other corpus techniques being exploited for CDA beyond collocation analysis explained here, such as keyword analysis and semantic annotation for extracting semantic categories from a corpus with Wmatrix (cf. Rayson, 2008; Al-Hejin, 2014).
2. https:// gephi. org/
3. The words in the list are grammatically nouns in Korean, but some are mostly used to qualify other nouns following them, which makes them function like adjectives. Hence, they are glossed in English as such.

Funding

This work was supported by Hankuk University of Foreign Studies Research Fund of 2018.

References

Al-Hejin, B. (2014). Covering Muslim women: semantic macrostructures in BBC news. *Discourse and Communication*, **9**(1): 19–46.

Ashley, L. and Olson, B. (1998) Constructing reality: print media's framing of the women's movement, 1966–1986. *Journal of Mass Communication Quarterly*, **75**(2): 263–77.

Banda, F. and Mawadza, A. (2015). 'Foreigners are stealing our birth right': moral panics and the discursive construction of Zimbabwean immigrants in South African media, *Discourse and Communication*, **9**(1): 47–64.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. (2014a). *Using Corpora to Analyze Gender*. London: Bloomsbury Academic.

Baker, P. (2014b). 'Bad wigs and screaming mimis': using corpus-assisted techniques to carry out critical discourse analysis of the representation of trans people in the British press. In Hart C. and Cap P. (eds), *Contemporary Critical Discourse Studies*. London: Bloomsbury, pp. 211–35.

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

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse and Society*, **19**(3): 273–306.

Baker, P. and McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics*, **4**(2): 197–226.

Bartlett, T. (2004) Mapping distinction: towards a systemic representation of power in language. In Young L. and Cl Harrison. (eds), *Functional Linguistics and Critical Discourse Analysis*. London: Continuum, pp. 68–84.

Belmonte, I. A., McCabe, A., and Chornet-Roses, D. (2010). In their own words: the construction of the image of the immigrant in Peninsular Spanish broadsheets and freesheets. *Discourse and Communication*, **4**(3): 227–42.

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, **2**(1). http:// journalofdigitalhumanities. org/ 2- 1/ topic- modeling- and- digital- humanities- by- david- m- blei/ (accessed 21 January 2017)

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**: 993–1022.

Burroughs, E. (2015). Discursive representations of 'illegal immigration' in the Irish newsprint media: the domination and multiple facets of the 'control' argumentation. *Discourse and Society*, **26**(2): 165–83.

Cho, S. G. and Kim, S. B. (2015). Summarization of documents by finding key sentences based on social network analysis. In Ali, M., Kwon, Y. S., Lee, C-H., Kim, J., and Kim, Y. (eds), *Current Approaches in Applied Artificial Intelligence (Proceedings of 28th*

International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems). London: Springer, pp. 285–292.

Costelloe, L. (2014). Discourses of sameness: expressions of nationalism in newspaper discourse on French urban violence in 2005. *Discourse and Society*, **25**(3): 315–340.

Danowski, J. A. and Cepela, N. (2010). Automatic mapping of social networks of actors from text corpora: time series analysis. In Memon, N., Xu, J. J., Hicks, D. L, and Chen, H. (eds), *Data Mining for Social Network Data*. London: Springer, pp. 31–46.

Drisko, J. W. and Maschi, T. (2016). *Content Analysis*. Oxford: Oxford University Press.

Don, Z. M. and Lee, C. (2014). Representing immigrants as illegals, threats and victims in Malaysia: Elite voices in the media. *Discourse and Society*, **25**(6): 687–705.

Fairclough, N. (2003). *Analyzing Discourse, Textual Analysis for Social Research*. London: Routledge.

Farquharson, K. and Marjoribanks, T. (2006). Representing Australia: race, the media and cricket. *Journal of Sociology*, **42**(1): 25–41.

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Flowerdew, John., Li, D. C. S., and Tran, S. (2002), Discriminatory news discourse: some Hong Kong data. *Discourse and Society*, **13**(3): 319–45.

Fortuna, B., Galleguillos, C., and Cristianini, N. (2009). Detection of bias in media outlets with statistical learning methods. In Srivastava, A. N. and Sahami, M. (eds), *Text Mining: Classification, Clustering, and Applications*. Boca Raton: CRC Press, pp. 27–50.

Gale, P. (2004). The refugee crisis and fear: Populist politics and media discourse. *Journal of Sociology*, **40**(4): 321–340.

Ghurye, G. S. (2003) Caste and race in India. In Reilly, K., Kaufman, S., and Bodino, A. (eds), *Racism: A Global Reader*. New York, NY: M. E. Sharpe, pp, 16–19.

Goldberg, D. T. and Solomos, J. (2002). General introduction. In Goldberg, D. T. and Solomos, J. (eds), *A Companion to Racial and Ethnic Studies*. Oxford: Blackwell, pp. 1–12.

Graham, S., Milligan, I. and Weingart, S. (2016). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press.

Grayson, S., Wade, K., Meaney, G., and Greene, D. (2016). The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature. In Bozic, B., Mendel-Gleason, G., Debruyne, C., and O'Sullivan, D. (eds), *Computational History and Data-Driven Humanities (Second IFIP WG 12.7 International Workshop)*. Cham, Switzerland: Springer, pp. 65–77.

Holliday, A., Hyde, M., and Kullman, J. (2004) *Intercultural Communication: An Advanced Resource Book*. London: Routledge.

Jacobi, C., van Atteveldt, W., and Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, **4**: 89–106. DOI: 10.1080/21670811.2015.1093271.

Jaworska, S. and Krishnamurthy, R. (2012). On the F word: a corpus-based analysis of the media representation of feminism in British and German press discourse, 1990-2009. *Discourse and Society*, **23**(4): 401–31.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Jones, T. (2015) *Policy and Gay, Lesbian, Bisexual, Transgender and Intersex Students*. London: Springer.

KhosraviNik, M., K. (2014). Immigration discourses and critical discourse analysis: dynamics of world events and immigration representations in the British press. In Hart, C. and Cap, P. (eds), *Contemporary Critical Discourse Studies*. London: Bloomsbury, pp, 501–519.

KhosraviNik, M., K. (2015). *Discourse, Identity and Legitimacy: Self and Other in Representations of Iran's Nuclear Programme*. Amsterdam: John Benjamins

KhosraviNik, M. K., Krzyżanowski, M., and Wodak, R. (2012). dynamics of representation in discourse: immigrants in the British press. In Messer, M., Schroeder, R., and Wodak, R. (eds), *Migrations: Interdisciplinary Perspectives*. London: Springer, pp. 283–295

Kim, S. (2012). Racism in the global era: Analysis of Korean media discourse around migrants, 1990–2009. *Discourse and Society*, **23**(6): 657–678.

Leetaru, K. H. (2012). *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content*. New York, NY: Routledge.

Lilleker, D. and Jackson, N. (2011). *Political Campaigning, Elections and the Internet: Comparing the US, UK, France and Germany*. London: Routledge.

Lirola, M. M. (2014). Legitimating the return of immigrants in Spanish media discourse. *Brno Studies in English*, **40**(1): 129–47.

Liu, J. H. and Mills, D. (2006) Modern racism and neoliberal globalization: the discourses of plausible

deniability and their multiple functions. *Journal of Community and Applied Social Psychology*, **16**(2): 83–99.

Liu, J., Li, Z., and Lu, H. (2012). Correlation mining for web news information retrieval. In Abraham, A. (ed.), *Computational Social Networks: Mining and Visualization*. London: Springer, pp. 103–28.

Mautner, G. (2009a) Checks and balances: how corpus linguistics can contribute to CDA. In Wodak, R. and Meyer, M. (eds), *Methods of Critical Discourse Analysis*. London: SAGE, pp. 122–43.

Mautner, G. (2009b) Corpora and critical discourse analysis. In Baker, P. (ed.), *Contemporary Corpus Linguistics.* London: Continuum. pp. 32–46

Mendes, K. (2011). *Feminism in the News: Representations of the Women's Movement since the 1960s*. Hampshire: Palgrave Macmillan.

Murakami A., Thompson, P., Hunston, S., and Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus, *Corpora*, **12**(2): 243–77.

Neuendorf, K. A. (2017). *The Content Analysis Guidebook*. Lost Angeles: SAGE.

Orpin, D. (2005). Corpus linguistics and critical discourse analysis: examining the ideology of sleaze. *International Journal of Corpus Linguistics*, **10**(1): 37–61.

O'Halloran, K. (2003) *Critical Discourse Analysis and Language Cognition*. Edinburgh: Edinburgh University Press.

O'Halloran, K. and Coffin, C. (2004) Checking Overinterpretation and underinterpretation: help from corpora in critical linguistics. In Coffin, C., Hewings, A., and O'Halloran, K. A. (eds) *Applying English Grammar: Functional and Corpus Approaches*. London: Hodder Arnold, pp. 275–97.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, **13**(4): 519–49.

Pietikainen, S. (2003), Indigenous identity in print: representations of the Sami discourse, *Discourse and Society*, **14**(5): 581–609.

Reisigl, M. and Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of Racism and Antisemitism*. London: Routledge.

Richardson, J. E. (2004) *(Mis)representing Islam: The Racism and Rhetoric of British Broadsheet Newspapers*. Amsterdam: John Benjamins Publishing Co.

Riff, D., Lacy, S., and Fico, F. (2014). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. New York, NY: Routledge.

Santa Ana, O. (2002). *Brown Tide Rising. Metaphors of Latinos in Contemporary American Public Discourse*. Austin: University of Texas Press.

Scott, M. (2017). *Wordsmith Tools Manual*. Gloucestershire: Lexical Analysis Software.

Simmons, K. and Lecouteur, A. (2008). Modern racism in the media: constructions of 'the possibility of change' in accounts of two Australian 'riots'. *Discourse and Society*, **19**(5): 667–687.

Seol, D-H. (2015). Population Ageing and International Migration Policy in South Korea. *Economy and Society*, **106**: 73–114.

Shin, G-W. (2013). Racist South Korea? Diverse but not tolerant of diversity. In Kowner R. and Demel W. (eds), *Race and Racism in Modern East Asia: Western and Eastern Constructions*. Leiden, The Netherlands: Brill. pp. 369–390.

Stubbs, M. (1997) Whorf's children: critical comments on critical discourse analysis. In Ryan, A. and Wray, A. (eds), *Evolving Models of Language*. Clevedon: Multilingual Matters, pp. 100–16.

Tanev, H. (2014). Learning textologies: networks of linked word clusters. In Biemann. C. and Mehler, A. (eds), *Text Mining: From Ontology Learning to Automated Text Processing Applications*. London: Springer, pp. 25–39.

Talbot, M. (2007). *Media Discourse: Representation and Interaction*. Edinburgh: Edinburgh University Press.

Teo, P. (2000). Racism in the news: a critical discourse analysis of news reporting in two Australian newspapers. *Discourse and Society*, **11**(1): 7–49.

Törnberg, A. and Törnberg, P. (2016a). Muslims in social media discourse: combining topic modeling and critical discourse analysis. *Discourse, Context and Media*, **13**: 132–42.

Törnberg, A. and Törnberg, P. (2016b). Combining CDA and topic modeling: analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse and Society*, **27**(4): 401–22.

Van Dijk, T. A. (1987) *Communicating Racism: Ethnic Prejudice in Thought and Talk*. Newbury Park, CA: SAGE.

Van Dijk, T. A. (1991). *Racism and the Press*. London: Routledge.

Van Dijk, T. A. (1993). *Elite Discourse and Racism*. Newbury Park, CA: Sage Publications

Van Dijk, T. A. (2000). New(s) racism: a discourse analytical approach. In Cottle, S. (ed.), *Ethnic Minorities and the Media*. Buckingham: Open University Press, pp. 33–49.

Van Dijk, T. A. (2002). Discourse and racism. In Goldberg, D. T. and Solomos, J. (eds), *A Companion to Racial and Ethnic Studies*. Oxford: Blackwell, pp. 145–59.

Van Dijk, T. A. (2004). Racist discourse. In Cashmore, E. (ed.), *Routledge Encyclopedia of Race and Ethnic Studies*. London: Routledge, pp. 351–55.

Van Dijk, T. A. (2005). *Racism and Discourse in Spain and Latin America*. Amsterdam: John Benjamins.

Van Dijk, T. A. (2012). The role of the press in the reproduction of racism. In Messer, M., Schroeder, R. and Wodak, R. (eds), *Migrations: Interdisciplinary Perspectives*. New York; London: Springer-Verlag Wien, pp. 15–29.

Waldherr, A., Heyer, G., Jähnichen, P., Niekler, A. and Wiedemann, G. (2016). Mining big data with computational methods. In Vowe, G. and Henn, P. (eds), *Political Communication in the Online World: Theoretical Approaches and Research Designs*. London: Routledge, pp. 201–217.

Wiedemann, G. (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. London: Springer.

Widdowson, H. G. (1995). Discourse analysis: a critical view. *Language and Literature*, **4**(3): 157–72.

Widdowson, H. G. (2004) *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell Publishing.

Wodak, R. and Meyer, M. (2009) Critical discourse analysis: history, agenda, theory and methodology. In Wodak, R. and Meyer, M. (eds), *Methods of Critical Discourse Analysis*. London: Sage, pp. 1–33.