Linguistic Research 43(2): 621-646
DOI: 10.17250/khisli.43.2.202606.011

Discursive representations of immigration in *The New York Times*: Integrating computational and pragmatic approaches*

Seongmin Mun** · **Hye-Kyung Lee*****
(Kyungpook National University · Ajou University)

Mun, Seongmin and Hye-Kyung Lee. 2026. Discursive representations of immigration in *The New York Times*: Integrating computational and pragmatic approaches. *Linguistic Research* 43(2): 621-646. This study investigates how immigrants are represented in *The New York Times* by applying text mining techniques and critical discourse analysis to a corpus of 1,000 immigration-related news articles. Combining methods such as word frequency analysis, word embeddings, and clustering, the study identifies major thematic domains within immigration discourse, including political and legal processes, border control, socio-economic impacts, and urban life. The analysis reveals that while overtly negative terms are largely absent, linguistic choices frequently foreground immigrants' legal status, framing them in ways that implicitly reinforce institutional ideologies and marginalization. The study further argues that such delicate ideological representations often elude large-scale quantitative approaches, highlighting the limitations of automated, frequency-based text analysis. The absence of explicitly negative language is attributed to both methodological constraints and data bias, including sampling limitations and a narrow source base. As such, the findings emphasize the need for "smart big data", datasets that combine volume with contextual richness, and call for the integration of qualitative discourse analysis to reflect the pragmatic and ideological complexity of media language. By connecting computational and interpretive approaches, this research contributes to interdisciplinary scholarship on critical discourse analysis, immigration studies, and data methodologies. **(Kyungpook National University · Ajou University)**

Keywords immigration discourse, media representation, critical discourse analysis, text mining, smart big data

* An earlier version of this paper was presented at 2024 ELSOK Summer conference. The authors would like to thank the audience for their constructive comments. However, all remaining errors are ours.
** First Author
*** Corresponding author

1. Introduction

Immigration continues to be one of the most politically charged and socially consequential topics in contemporary media discourse. In this context, the media play an important role in shaping public perceptions of immigrants, not only through the topics they cover but also through the language they use. While previous research has extensively documented overtly negative representations of immigrants, less attention has been paid to more subtle and indirect forms of marginalization that persist even in ostensibly neutral or sympathetic reporting.

This study addresses that gap by investigating how immigrants are discursively represented in *The New York Times* (NYT), one of the most influential news outlets in the United States. Specifically, it focuses on the pragmatic implications of linguistic choices that foreground immigrants' legal status or associate them with political or socio-economic problems. Terms such as *undocumented* or *unauthorized* immigrants may appear descriptively neutral, but they often function to reduce complex human identities to legal categorizations. Such framing not only reproduces institutional ideologies but also contributes to what Goffman (1963) describes as "courtesy stigma", a subtle form of stigmatization that persists even in supportive narratives.

To examine these patterns, this study employs text mining techniques, namely word frequency analysis, word embeddings, and clustering, on a corpus of 1,000 recent NYT articles related to immigration. After collecting and preprocessing the data using standard NLP tools and compound word recognition algorithms, we identify high-frequency terms and group semantically related words using Word2Vec and clustering algorithms. The patterns are further visualized through Zipf's law graphs and t-SNE dimensionality reduction to reveal the structural contours of immigration discourse.

However, despite the strengths of these automated methods, they also present some limitations such as a striking absence of explicitly negative terms commonly identified in prior research. This absence, we argue, may be the result of sampling bias, temporal limitations, and the inherent constraints of frequency-based analysis. Importantly, the intricate ideological work of language—such as implicatures (à la Grice 1975), omissions, or framing devices—often eludes purely statistical techniques and requires context-sensitive interpretation. This limitation points to the need for greater methodological integration between quantitative and qualitative approaches. While text

mining provides valuable insights into large-scale patterns, it must be complemented by critical discourse analysis to uncover the complex ways in which language enacts ideology, stereotyping, and social marginalization. The development of "smart big data", data that is not only large in volume but also rich in interpretive depth, is crucial for this undertaking. Ultimately, this study aims to contribute to the ongoing academic endeavor on methodological hybridity and the ethical responsibilities of discourse analysis in an era of algorithmic research.

2. Literature review

2.1 Critical Discourse Analysis

Critical Discourse Analysis (CDA) examines the mechanisms through which social power, dominance, and inequality are enacted, perpetuated, and occasionally contested by analyzing various forms of communication within their broader social and political contexts (Fairclough 1989, 1995, 1998; van Dijk 1993, 2015; Wodak 2007). Rather than being a singular methodological approach, CDA represents a critical perspective on the use of discourse practices as tools to uphold existing power structures. Accordingly, CDA encompasses a multidisciplinary framework, integrating insights and methodologies from diverse areas of discourse studies, including conversation analysis, argumentation theory, pragmatics, multimodal discourse analysis, and sociolinguistics, among others (see van Dijk 2015 for an overview).

In CDA, discourse is understood as a form of social practice that arises from a bidirectional relationship between discursive events and their contexts, including situations, institutions, and social structures (Reisigl and Wodak 2009). Discourse also plays a significant role in the categorization of social actors into dominant groups (ingroups) and subordinate groups (outgroups). Immigrants, for example, are typically categorized as members of "weaker" outgroups. This process of categorization is closely tied to ideology, which functions as a framework for expressing political stances and interpreting reality. As van Dijk (1998) explains, ideologies inherently involve a polarization between ingroups and outgroups, promoting positive beliefs and attitudes toward ingroups while fostering negative perceptions of outgroups.

This ideological framing is implicated in the negative misrepresentation of

outgroup members, often in parallel with the positive misrepresentation of ingroup members. Importantly, such misrepresentation is not always intentional, which may stem from unconscious biases or unintentional misuse of language (Coshignano, Minnema, and Zanchi 2023: 108). As in Cogshigano et al's (2023) analysis, while most articles in our data rarely contain overtly anti-immigration rhetoric, they frequently employ routinized linguistic expressions that activate framing effects, often resulting in negative perceptions of outgroups and events related to them. For this reason, the present analysis adopts a critical perspective on discourse, as this study aims to examine how ideology is embedded and normalized to the extent that it may appear neutral and becomes widely accepted as common-sense (Fairclough 1989).

2.2 Previous research on media representation of immigrants

The past decade has witnessed an increasing body of research on media representation of immigration. A comprehensive meta-analysis of media portrayal of Refugees, Asylum Seekers, and Immigration (RASIM) by Seo and Kavakli (2022) identify several characteristics of existing research on the topic. Their findings reveal a rapid expansion of research in this field over the last decade, with increasing contributions from interdisciplinary perspectives such as anthropology and regional studies. It is also pointed out that research remains largely concentrated on the Global North. Other key issues include the scarcity of comparative studies across different regions and the problematic nature of media representation. Media narratives often construct a binary opposition between host-country citizens and immigrants, emphasizing an "us vs. them" discourse, as discussed in CDA-based research.

In the literature, three major types of media discourse of immigration have been identified (Fuller 2024): othering discourses, threat discourses, and deservedness discourses. First, discourses on immigration inherently engage in othering, shaping notions of belonging in host countries through racial hierarchies that frequently dehumanize others (Giorgi and Vitale 2017; Böröcz 2021; Lirola 2022; Morera-Bello et al. 2026). Immigrant voices are largely muted in media narratives, as immigration is discussed about them rather than by them, leading to stereotypes and oversimplified portrayals. Second, in threat discourses, immigrants are often framed as a societal threat through both overt and subtle linguistic strategies, including metaphors likening

them to parasites or natural disasters (e.g., Baker et al. 2008; Jaspal and Cinnerella 2010; Caviedes 2015; Musolff 2021). These threat discourses ultimately reinforce a monolithic conception of host communities, positioning immigrants as external and incompatible with the host societies. Third, deservedness discourses differentiate between refugees, who are seen as worthy of protection due to forced displacement, and economic immigrants, who are stigmatized despite their pursuit of financial stability (Ullmann 2023). These discourses often fluctuate over time; initial sympathy for refugees often gives way to threat narratives that frame immigrants as economic and security risks (Vollmer and Karakayali 2018; Trost 2023). Additionally, immigrant worthiness is stratified by economic utility, with those serving the wealthy viewed as more deserving than asylum seekers who are perceived as burdens on society (Giorgi and Vitale 2017; Morris 2025).

The influence of media discourse on immigration and immigrants in shaping public perception has been widely studied (Eberl et al. 2016; Kroon et al. 2016; Conzo et al. 2021; Urman et al. 2022; Madrigal and Sorika 2023). For example, Kroon et al. (2016) find that tabloid newspapers predominantly focused on perpetrator frames, which likely reinforce negative perceptions of minority groups among their readers. Given that tabloid audiences are more inclined to view minority issues as threats (Vergeer et al. 2000), this framing may further amplify existing biases and anxieties. Conzo et al. (2023), using surveys and an experiment, examine how positive and negative discourses on immigration affect prosocial behavior across different levels of ethnic diversity. Their findings reveal that negative media depictions provoke physiological and emotional antagonism toward the outgroup while reinforcing ingroup favoritism in economic interactions, which may lead to efficiency declines in ethnically diverse markets. In an experimental study of news photos, Madrigal and Sorika (2023) evidence that the selection of news photos of immigrants affects readers' reaction to news content, thereat sensitivity, and attitudes. They also suggest that news readers may move in pro-immigration direction when immigrants are represented as individuals rather than en masse, providing the impact of person positivity in the context of visual news images.

Collectively, these studies indicate the significant role media plays in shaping public perceptions of immigrants, emphasizing the power of linguistic choices in either reinforcing stereotypes or challenging negative perceptions. Given this influence, they advocate the importance of critically analyzing media content to understand and

mitigate potential biases, proposing a more inclusive and accurate representation of immigrant communities. Building on this foundation, the present study investigates how immigration-related issues are represented by the NYT, by examining the specific lexical patterns employed in its articles.

2.3 Text mining analysis

Text mining is a powerful data analysis method used to discover meaningful patterns, trends, and insights from large collections of natural language texts (Hotho et al. 2005). The main goal is to transform unstructured text data into structured, analyzable information. This process combines techniques from natural language processing (NLP) and data mining to identify hidden meanings and generate new knowledge from language-based data (Cohen and Hunter 2008).

Text mining typically involves several key steps: collecting text data from sources like social media, news articles, or research papers; cleaning and preprocessing the text by removing errors and irrelevant words; breaking the text into tokens (such as words or morphemes); and extracting features like word frequency, TF-IDF scores, or word embeddings. These features are then used in various types of analysis, including sentiment analysis, clustering, and topic modeling. The results are often visualized using tools like word clouds or graphs to make the findings easier to interpret (Kim et al. 2024).

Some commonly used techniques in text mining include frequency analysis (which helps identify important keywords), word embeddings (which represent the meanings of words in a numerical vector space), clustering (which groups similar texts or words together), topic modeling (which automatically finds recurring themes in documents), and sentiment analysis (which classifies emotional tone in text) (Shin and Mun 2022). Compared to traditional qualitative research, text mining enables researchers to analyze large-scale text data quantitatively. This approach is now widely used in fields like opinion analysis, customer feedback evaluation, document classification, and policy trend analysis. It helps uncover patterns that might otherwise go unnoticed and provides insights that support decision-making in research, business, and government (Shin et al. 2024).

In this study, we apply text mining methods, such as frequency analysis, word

embeddings, and clustering, to examine how immigrants are represented in articles from the NYT. These methods allow us to identify key terms, thematic patterns, and the structure of public discourse surrounding immigration.

3. Data and methodology

3.1 Data collection and preprocessing

This study analyzes how immigrants are portrayed in the media by examining articles published in the NYT. Specifically, we searched for articles using the keyword "immigration" on the newspaper's official website and collected the top 1,000 most recent articles that appeared in the search results. Since the website limits the number of search results to 1,000, this set formed the basis of our text corpus. The collected articles span a period from May 31, 2019 to May 19, 2024.

To gather the articles, we used a web scraping technique, which allowed us to automatically extract data from online sources. With the help of Python's selenium library, we automated navigation through dynamic web pages, and used the BeautifulSoup library to extract text content from the HTML code. When access issues such as IP restrictions occurred, we also saved the HTML files and manually extracted the article text as a backup method. In addition to the article content itself, we collected relevant metadata for the analysis, including the title, lead sentence, and the webpage URL.

Once the data was collected, we carried out a text preprocessing phase to ensure reliable analysis. First, we cleaned the raw text using Python by removing unnecessary characters such as punctuation marks, line breaks, and extra spaces, and by standardizing the format across all documents. We then applied morphological analysis and stopword removal to focus only on the meaningful words in the articles. For this, we used the Natural Language Toolkit (NLTK), a widely used Python library for natural language processing. With NLTK, we extracted content words such as nouns, verbs, adjectives, and adverbs, while removing stopwords like articles, prepositions, and conjunctions that do not contribute much to the core meaning of the text.

Finally, we applied the compound-word recognition algorithm proposed by Mun

et al. (2021) to merge multiword expressions such as *United States* into a single token (e.g., United_States). This process helped ensure that words like United and States, which might otherwise be treated separate high-frequency words, were instead recognized as a single meaningful unit. As a result, the accuracy of topic identification across the documents were enhanced. These preprocessing procedures provided the foundation for the subsequent analyses, including word frequency analysis, word embedding, and clustering.

3.2 Data analysis: Word frequency analysis, word embedding, and clustering

To understand how immigrants are described in the NYT articles, we used three main analysis methods: checking word frequency, mapping word meanings with word embedding, and grouping similar words using clustering. These methods helped us find common words, understand how they are used in context, and discover major themes in the articles.

First, we looked at how often certain words appeared using word frequency analysis. This showed us which words were used the most in the articles, giving us a quick view of the main topics. We also used a graph based on Zipf's Law, which shows that a few words are used very often, while most words appear only a few times (Zipf 1932; Li 1992).

Next, to examine how words relate to each other in meaning, we used Word2Vec with the Skip-gram with Negative Sampling (SGNS) algorithm (Mikolov et al. 2013; Goldberg 2014). The SGNS model learns word representations by predicting surrounding context words for a given target word within a defined window. Through this process, each word is mapped onto a point in a high-dimensional continuous vector space. In this space, semantic relationships between words are operationalized in terms of distance or similarity metrics between their vectors. In particular, words that occur in similar contexts are positioned closer together, and their semantic similarity can be quantitatively measured using metrics such as cosine similarity, which captures the angle between vectors regardless of their magnitude.

The "negative sampling" component improves computational efficiency by updating only a subset of weights, contrasting observed context words with randomly sampled noise words. As a result, semantically related words are represented by vectors

that are close to one another in the vector space. For example, words such as *immigrant*, *refugee*, and *migrant* are located in close proximity, indicating that they share similar contextual usage patterns in the corpus.

Then we used K-means clustering to group these word vectors into categories. K-means is an unsupervised learning algorithm that partitions data points into k distinct clusters by minimizing the distance between each point and the centroid of its assigned cluster (MacQueen 1967; Hartigan and Wong 1979). The process begins by initializing k centroids randomly, assigning each word vector to the nearest centroid, and iteratively updating the centroids until convergence. This method effectively groups together words that appear in similar semantic and syntactic environments.

In addition to K-means clustering, hierarchical clustering was applied to further examine the semantic relationships among embedded words. Prior to clustering, cosine similarity between word vectors was calculated and then transformed into a distance metric to construct a similarity-based distance matrix. To determine an appropriate linkage method, we compared several hierarchical clustering strategies (e.g., single, complete, and average linkage) and selected the method that most clearly preserved the overall structure of the data. Based on this evaluation, the complete linkage method —which defines the distance between clusters as the maximum distance between their elements—was adopted for the final analysis.

Finally, to analyze the data more intuitively, we used t-Stochastic Neighbor Embedding (t-SNE), a visualization technique that reduces multidimensional data to two dimensions. t-SNE is a dimensionality reduction method that visually organizes complex data by similar structures, helping us better understand the data structure (Van der Maaten and Hinton 2008).

4. Results and discussion

4.1 Word frequency analysis, word embedding, and clustering

As explained in Section 3.3, we first examined the word frequencies using a Zipf's law graph to understand how immigrants are portrayed in the NYT articles. We found that words like *immigration* (3,622 times) and *border* (3,295 times) appeared much

more frequently than other words in the NYT articles related to immigrants, as shown in Figure 1.

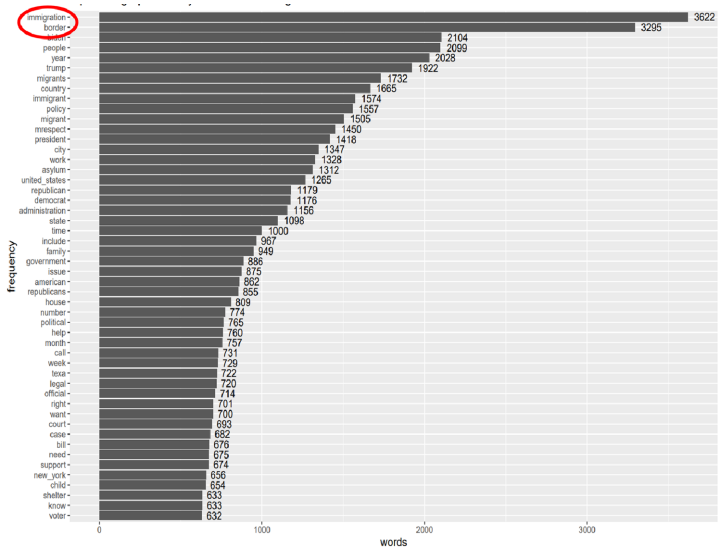

Figure 1. Zipf's law graph of the NYT on immigration

While this result confirms that immigration is the central topic of the corpus, the high frequency of the word *border* is particularly noteworthy from a discourse-analytic perspective. Rather than simply referring to a geographic boundary, border is closely associated with issues of control, surveillance, and enforcement. Its prominence suggests that immigration is frequently framed not as a human or social phenomenon, but as a matter of territorial management and national security.

Next, a dendrogram was constructed to visualize the hierarchical organization of the word clusters. This representation facilitates the interpretation of semantic proximities and distinctions among key terms in the corpus across multiple levels of granularity. The resulting structure is presented in Figure 2.

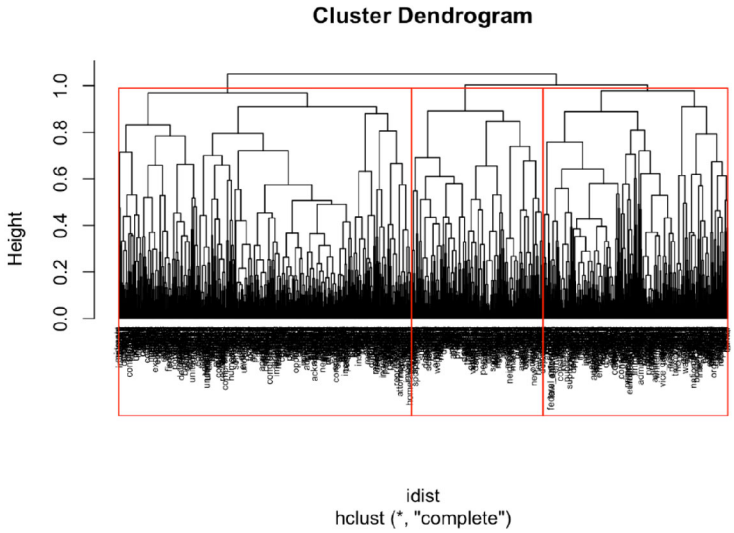

Figure 2. After analyzing the distance relationships by using k-means clustering

Due to the large number of data points, the resulting dendrogram presents limitations as a visualization tool for detailed interpretation. To address this, we highlighted three major clusters in Figure 2 using bounding boxes. These clusters represent higher-level semantic groupings that correspond broadly to distinct thematic domains in the corpus. The boxed regions therefore serve as an interpretive aid, allowing us to focus on macro-level patterns rather than fine-grained hierarchical distinctions. They also provide a basis for subsequent qualitative interpretation of the clustered lexical items.

Then to analyze the data more intuitively, we used t-Stochastic Neighbor Embedding (t-SNE), which is a visualization technique that reduces multidimensional data to two dimensions. The result is presented in Figure 3.

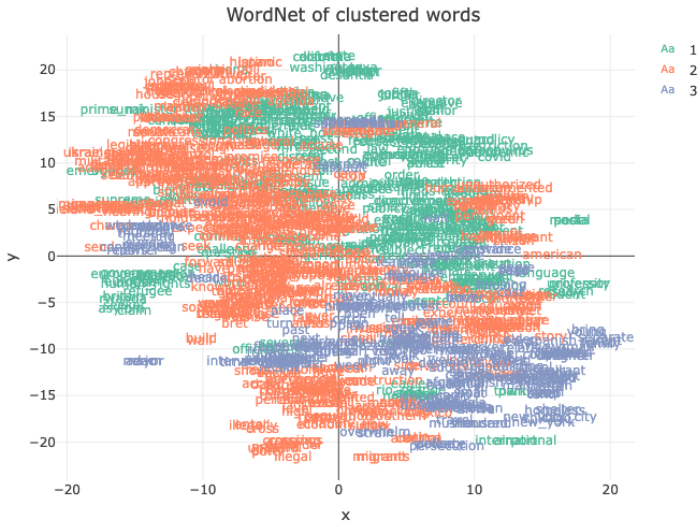

Figure 3. The word embeddings reduced to two dimensions through t-SNE

Figure 3 shows the word embeddings reduced to two dimensions through t-SNE, revealing that the words clearly divide into three main clusters. The first group includes words like 'immigration,' 'asylum,' 'migration,' 'deportation,' 'seeker,' 'border_patrol,' 'enforcement,' and 'refugee.' The second group contains 'border,' 'migrants,' 'immigrant,' 'migrant,' 'united_states,' 'mexico,' 'cross,' 'illegal,' 'undocumented,' 'crossing,' 'entry,' 'illegally,' 'southern,' and 'port.' The third group consists of location-related terms such as 'city,' 'new_york,' 'shelter,' 'housing,' 'street,' 'neighborhood,' 'apartment,' 'building,' 'camp,' 'tent,' 'crowd,' 'manhattan,' 'queens,' and 'brooklyn.'

Beyond this descriptive grouping, the t-SNE visualization reveals a meaningful discursive structuring of immigration-related language. The first cluster primarily reflects institutional and policy-oriented discourse, the second cluster foregrounds border control and movement across national boundaries, and the third cluster represents the urban and socio-spatial contexts in which immigrants are situated. This division suggests that immigration discourse is not organized around immigrants as central social actors, but rather distributed across distinct thematic domains such as governance, regulation, and settlement.

4.2 Structuring of word clusters

The words included in each cluster were further subcategorized into smaller groups, as shown in Tables 1, 2, and 3.

Table 1. Cluster A: Political/Legislative measures

Sub-group		Examples
A-1	Political Leadership and Governmental Actions	"biden," "trump," "president," "administration," "government," "state," "official," "party," "conservative," "white_house," "department," "governor," and "vice_president"
A-2	Political and Legislative Processes	"republican," "democrat," "house," "senate," "congress," "bill," "vote," "election," "candidate," "legislation," "reform," "measure," "proposal," "law," "committee," "representative," "senator," and "congressional"

Table 2. Cluster B: Immigration policy and border control

Sub-group		Examples
2-1	Immigration Policy and Administration	"immigration," "asylum," "migration," "deportation," "seeker," "border_patrol," "enforcement," and "refugee“
2-2	Immigration and Border Control	"border," "migrants," "immigrant," "migrant," "united_states," "mexico," "cross," "illegal," "undocumented," "crossing," and "entry"

Table 3. Cluster C: Socio-economic impacts

Sub-group		Examples
3-1	Socio-Economic Impacts	"work," "american," "population," "worker," "support," "community," "voter," "percent," "increase," "change," "economic," "labor," "opportunity," "funding," "growth," and "poll"
3-2	Family and Social Support	"family," "child," "home," "parent," "mother," "father," "daughter," "wife," "husband," "relative," "brother," "friend," "adult," "older," "newcomer," "social," "welcome," and "assistance"

A close examination of the words in Tables 1, 2, and 3 reveals the presence of proper nouns referring to specific places and individuals, such as *New York* and *Trump*. In this study, proper nouns were included in the word lists on the grounds that geographic locations and public figures play a significant role in immigration discourse. Although automatic clustering and subgrouping were performed using Word2Vec, some redundancy and overlap were evident in the resulting categories. For instance,

subcategories 1-1 and 2-1 both relate to immigration, yet the word *immigration* appears in 1-1 while *migrant(s)* is placed in 2-1. Similarly, subcategories 1-2 and 2-2 share the thematic link of politics. To address these issues, the automatically generated clusters were reorganized into four categories, as shown in Tables 4, 5, 6, and 7 based on the researchers' expertise in the subject matter and their linguistic intuition.

Table 4. Revised cluster 1

Sub-group		Examples
1-1	Political Leadership and Governmental Actions	"biden," "trump," "president," "administration," "government," "state," "official," "party," "conservative," "white_house," "department," "governor," and "vice_president"
2-2	Political Leadership and Legislative Processes	"republican," "democrat," "house," "senate," "congress," "bill," "vote," "election," "candidate," "legislation," "reform," "measure," "proposal," "law," "committee," "representative," "senator," and "congressional"
1-3	Legal and Judicial Processes	"legal," "court," "case," "judge," "claim," "justice," "lawyer," "deport," "detention," "supreme_court," "lawsuit," "constitutional," and "law_enforcement

Table 5. Revised cluster 2

Sub-group		Examples
1-1	Immigration Policy and Administration	"immigration," "asylum," "migration," "deportation," "seeker," "border_patrol," "enforcement," and "refugee"
2-1	Immigration and Border Control	"border," "migrants," "immigrant," "migrant," "united_states," "mexico," "cross," "illegal," "undocumented," "crossing," and "entry"

Table 6. Revised cluster 3

Sub-group		Examples
2-3	Socio-Economic Impacts	"work," "american," "population," "worker," "support," "community," "voter," "percent," "increase," "change," "economic," "labor," "opportunity," "funding," "growth," and "poll".
3-2	Family and Social Support	"family," "child," "home," "parent," "mother," "father," "daughter," "wife," "husband," "relative," "brother," "friend," "adult," "older," "newcomer," "social," "welcome," and "assistance"

Table 7. Revised cluster 4

Sub-group		Examples
3-1	Urban Life and Challenges	"city," "new_york," "shelter," "housing," "street," "neighborhood," "apartment," "building," "camp," "tent,"

		"crowd," "manhattan," "queens," and "brooklyn"
3-3	Migration and (Re)settlement	"move," "arrive," "live," "travel," "journey," "flee," "venezuelan," "venezuela," "haitian," "persecution," "refugee," "escape," "homeless," "shelter," "new_york_city," "new_yorker," "route," "desperate," and "smuggler"

4.3 Lack of negative words

A body of previous research on immigration and immigrants has found that immigration-related issues are frequently framed in negative terms, such as crimes, predators, or threats. For example, Harris and Gruenewald (2022) find that most news reports linking immigration and crime portray immigrants as predisposed to criminal behavior or as contributing to rising overall crime rates. They also demonstrate that such framing has become more common over time, alongside misleading narratives that treat undocumented immigration itself as a criminal offense. Unlike previous studies, this study's high-frequency word extraction method revealed a notable absence of negative terms associated with immigrants or immigration. In particular, words linking undocumented immigrants to crime or socioeconomic burdens were largely missing, with only a few exceptions such as *smuggle*, *homeless*, and *persecution*. This result is likely attributable to data bias (Hammersley and Gomm 1997; Liu 2023; Shahbazi et al. 2023). According to Shahbazi et al. (2023: 9), data bias can be categorized into three types.

(1) Three types of data bias
   a. **Historical Bias** that arises from pre-existing societal inequalities that are reflected in data. For example, image searches for "CEO United States" mostly show men because only a small percentage of CEOs are women, mirroring real-world disparities.
   b. **Underlying Distribution Skew** that occurs when the data reflects an imbalanced real-world distribution without any discriminatory intent. For instance, since only 7% of the U.S. population is Asian, data collected uniformly may underrepresent this group, potentially leading to biased outcomes in certain applications.
   c. **Sampling/Selection/Self-Selection Bias** that results from flaws in how data samples are chosen. Selection bias happens when the sample is not

representative due to flawed selection methods, while self-selection bias arises when participation is voluntary, skewing the results. Sampling bias refers more generally to any non-random sampling that fails to capture the true population distribution.

The absence of negative representations observed in this study's data—contrary to findings from previous research—appears to stem primarily from sampling and selection bias. First, the analysis is based solely on articles from a single media outlet, the NYT, rather than a diverse range of sources. Second, the dataset includes only articles from the past decade, rather than a comprehensive collection spanning a longer period, which increases the likelihood of selection bias. Additionally, since this study focuses on frequently occurring words, it did not analyze words that co-occur with these high-frequency terms, which may have contributed to the limited appearance of negative words.

Given these factors, a qualitative analysis of the collected corpus is necessary to assess whether negative portrayals of immigrants are truly absent or merely overlooked by the quantitative method. This is particularly important because findings that suggest a lack of negative content in immigration-related news may mask the realities of immigration challenges and obscure systemic inequalities (cf. Halman 2011). Moreover, as Cheng (2002) notes, news discourse often leaves much unsaid and relies heavily on implicatures (à la Grice 1975) to convey meanings indirectly—meanings that are grounded in shared beliefs, opinions, or contextual knowledge. This study therefore critiques the absence of explicitly negative expressions, which may impede clear communication and, unintentionally, reinforce stereotypes or misinform readers.

5. Discussion: Qualitative analysis

The effects of unintended exposure to negative language have been widely reported in the literature (e.g., Shattell 2009; Conzo et al. 2021; Lutz and Bitschnau 2022). For example, Shattell (2009) argues that unintended meanings may reinforce existing social power structure and become normalized to the point of invisibility. Shattell (2009) further emphasizes the importance of recognizing such implicit meanings, particularly in discourse concerning marginalized groups. Against this backdrop, the

following closely examines a selection of sampled articles form the NYT.

Excerpt (2) is part of an article about a federal judge who temporarily blocked part of a Florida law that criminalized transporting undocumented immigrants into the state, citing constitutional concerns. The content of the article is generally favorable to undocumented immigrants, as it temporarily halts the enforcement of a law that could criminalize those who assist them in traveling into Florida. However, in the process of reporting on undocumented immigrants, the language used to describe their current condition or legal status often includes negatively charged terms. This tendency is evident in the examples provided in (2), which illustrate how such word choices may contribute to unfavorable representations.

(2) **Judge Temporarily Blocks Florida From Criminalizing Transport of Undocumented Immigrants**

A federal judge on Wednesday temporarily blocked part of a Florida law that criminalized transporting into the state anyone ***who lacked lawful immigration status****, raising new legal questions for other states pursuing similar measures.*

The Florida law was intended to discourage ***unauthorized immigrants*** *from living and working in the state, and organizations that work with immigrants say many* ***undocumented*** *workers have left the state in recent months. (by Miriam Jordan, May 22, 2024)* [1]

Since the article itself focuses on legal actions involving undocumented immigrants, it is inevitable that words such as *undocumented* and *unauthorized* frequently co-occur with *immigrant*. As these collocations become more frequent and readers are repeatedly exposed to them, the likelihood increases that readers may form or reinforce associations between immigrants and negative terms.[2]

1 https://www.nytimes.com/2024/05/22/us/florida-undocumented-immigrant-transport.html

2 Within the corpus, the frequencies of collocations involving *undocumented* and *unauthorized* with *immigration* and *immigrant*(*s*) are as follows:

word 1	word 2	n
undocumented	immigrant(s)	268
unauthorized	immigrant(s)	61
unauthorized	immigration	12

(3) was excerpted from an article that challenges Donald Trump's anti-immigrant rhetoric, particularly his claim that immigrants are harming the country. It highlights Queens, New York, where Trump grew up, as a counterexample. According to the article, Queens is the most ethnically and racially diverse county in the continental U.S., with immigrants making up nearly half the population. Far from being dangerous, Queens has lower rates of serious crime than the rest of New York City, illustrating that immigrant communities contribute positively to society.

(3) **No, Immigrants Aren't 'Poisoning the Blood of Our Country'**

Does Donald Trump ever visit Queens, the land of his youth? If he did, he would presumably be horrified. According to the census, Queens is the most racially and ethnically diverse county in the continental United States; it's hard to think of a nationality or culture that isn't represented there. Immigrants are almost half the borough's population and more than half its work force.

. . .

And no, Queens isn't an urban hellscape. It may not be leafy and green, but it has less serious crime per capita than the rest of New York City, and New York, although nobody will believe it, is one of the safest places in America. It's also relatively healthy, with life expectancy around three years higher than that of the United States as a whole.

*But Trump has declared that migrants are "**poisoning the blood of our country**" – a phrase that, to steal from the late, great Molly Irvins, might sound better in the original German. Look, I know there's a debate over whether the MAGA movement fully meets the classic criteria for fascism, but can we at least agree that its language is increasingly fascist-adjacent?*[3]

Although the article criticizes Trump's inappropriate and politically incorrect expression, it still needs to quote his words in order to report and evaluate them. However, once such inflammatory language is exposed to the public, there is a risk that it may become memeified and widely circulated. In that process, it could

undocumented	immigration	10
immigrant	undocumented	1

3 https://www.nytimes.com/2024/05/22/us/florida-undocumented-immigrant-transport.html

unintentionally create or reinforce an association between immigrants and the idea of "poisoning the blood of the U.S." (e.g., Zannettou et al. 2020; Layne 2023; Pandiani et al. 2025). Figure 3 illustrates the Google Trends data for the expression "poisoning the blood." The data reveal a significant spike in search frequency in November 2023, coinciding with widespread media coverage of Donald Trump's use of the phrase. This peak is followed by sustained levels of more heightened circulation, indicating continued public engagement with the expression.

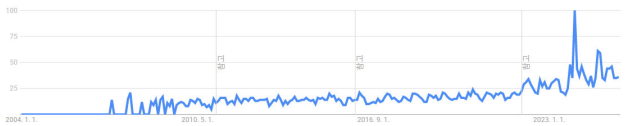

Figure 4. The Google Trends data for the expression "poisoning the blood"

(4) was excerpted from an article that reports on a case that, while not reflected in overall statistics, fuels concerns about a perceived rise in crimes involving migrants. It covers a shocking incident in which a 15-year-old migrant allegedly shot at a police officer.

(4) **'Migrant Crime Wave' Not Supported by Data, Despite High-Profile Cases**
In the past month, the New York Police Department has described alarming crimes involving young men living in ***the city's migrant shelters.***

A 15-year-old boy, the police said, ***shot at an officer*** *in Times Square and hit a tourist. Two officers were kicked and punched on West 42nd Street. A Venezuelan man oversaw* ***a ring of criminals*** *who rode mopeds and snatched purses and cellphones from more than 60 people, most of them women walking alone.*

During an early-morning police raid last week in the Bronx, Mayor Eric Adams, dressed in a bulletproof vest over a Fendi scarf, joined officers as they arrested five people accused of perpetrating the robbery spree. "A ***migrant crime wave*** *is washing over our city," Police Commissioner Edward Caban told reporters hours later. Some of the crimes were captured on videos*

> *that have since gone viral, leading Republican politicians and their allies to say that* ***migrant criminals*** *are besieging New York.*
>
> *Quantifying* ***crimes committed by migrants*** *is nearly impossible, because the police are not allowed to ask about a suspect's immigration status, said Kenneth Corey, a former chief of the department who retired in 2022. But police data indicate that there has been no surge in crime since April 2022, when Gov. Greg Abbott of Texas started sending buses of migrants to New York to protest the federal government's border policy.*[4]

In reporting this incident, the article inevitably provides background information about the individuals involved in the crime, including their age, country of origin, and immigration status. As a result, expressions such as *migrant shelter, migrant crime/criminals*, and *crimes committed by migrants*, highlighted in article (4), are repeated in ways that reinforce a collocational link between *migrants* and *crime*. This pattern aligns with findings from previous studies that point out how immigrants are frequently framed through a crime-related lens (e.g., Merolla et al. 2013; Caviedes 2015; Chouliaraki and Stolic 2017; Eberl et al. 2018; Farris and Mohamed 2018; Coschignano, Minnema, and Zanchi 2023). Moreover, although statistical data do not support the claim that migrant crime is on the rise, the article's title, '*Migrant Crime Wave' Not Supported by Data, Despite High-Profile Cases*,' implies an experiential or perceived increase in such crime, especially in light of factors such as regulations preventing police from inquiring about immigration status and the symbolic act of Texas's governor sending buses of migrants to New York City.

As this section has shown, in an effort to attract readers' attention, news outlets often report incidents using dramatic and emotionally charged language. When reporting on immigrants, particularly those in vulnerable situations, journalists may inevitably rely on negative portrayals. Such portrayals, whether intentional or unintentional, can reinforce associations between immigrants and negative attributes.

These patterns, commonly observed in qualitative research, may be overlooked in large-scale linguistic analyses. To address this limitation, there is a growing need to collect more comprehensive and balanced datasets, along with the development of more complicated analytical tools. In this context, the concept of smart big data,

4 https://www.nytimes.com/2024/02/15/nyregion/migrants-crime-nyc.html

data that is not only large in volume but also rich in contextual and interpretive depth, becomes increasingly important. Close collaboration between data scientists and discourse analysts must also be essential in achieving this goal.

6. Conclusion

This study employed text mining techniques in conjunction with a critical discourse analytic framework to systematically investigate the representation of immigrants in a corpus of immigration-related articles published by the NYT. The automated analysis revealed that immigration discourse in this mainstream media outlet is organized around several key thematic domains, including immigration policy, border control, socio- economic impacts, and urban settlement. These clusters provide a macro-level view of how the topic of immigration is structured and prioritized in journalistic narratives.

Notably, while the surface-level lexical choices did not include overtly negative terms commonly associated with immigrant rhetoric in previous studies, this absence should not be interpreted as evidence of ideological neutrality. Rather, it revealed the methodological limitations of purely quantitative, frequency-based approaches, which may not detect implicit biases, subtle forms of stereotyping, or pragmatic implicatures (à la Grice 1975) embedded in context. This was particularly relevant in light of Goffman's (1963) notion of "courtesy stigma," where even ostensibly supportive representations can encode assumptions of deviance by foregrounding legal status or other marginalizing frames.

The lack of explicit negativity in the dataset likely resulted from various forms of data bias, especially sampling bias, given the reliance on a single media outlet and the limited temporal scope. In this respect, the findings pointed to the importance of supplementing computational analyses with qualitative discourse approaches. While text mining offers valuable scalability and pattern recognition, it must be complemented with interpretive depth to examine how media discourse is interrelated with ideology and power.

Overall, this study advocated for the development of "smart big data", datasets that are not only extensive in volume but also curated for contextual richness and analytical granularity. The integration of computational tools with critical discourse

analysis is essential to fully grasp the sociopolitical functions of language in media. Achieving this requires close inter-disciplinary collaboration between discourse analysts, computational linguists, and data scientists.

References

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michal Krzyżanowski, Tony McEnery, and Ruth Wodak. 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 & Society* 19(3): 273-306. https://doi.org/10.1177/0957926508088962.

Böröcz, József. 2021. "Eurowhite" conceit, "dirty white" ressentment: "Race" in Europe. *Sociological Forum* 36(4): 1116-1134. https://doi.org/10.1111/socf.12752.

Caviedes, Alexander. 2015. An emerging 'European' news portrayal of immigration? *Journal of Ethnic and Migration Studies* 41(6): 897-917. https://doi.org/10.1080/1369183x.2014.1002199.

Cheng, Maria. 2002. The standoff- What is unsaid? A pragmatic analysis of the conditional marker `if'. *Discourse & Society* 13(3): 309-317. https://doi.org/10.1177/0957926502013003050.

Chouliaraki, Lilie and Tijana Stolic. 2017. Rethinking media responsibility in the refugee 'crisis': A visual typology of European news. *Media, Culture & Society* 39(1): 016344371772616. https://doi.org/10.1177/0163443717726163.

Cohen, Kevin B. and Lawrence Hunter. 2008. Getting started in text mining. *PLoS Computational Biology* 4(1): e20. https://doi.org/10.1371/journal.pcbi.0040020.

Conzo, Pierluigi, Giulia Fuochi, Lorenzo Anfossi, Francesca Spaccatini, and Carlo Orazio Mosso. 2021. Negative media portrayals of immigrants increase ingroup favoritism and hostile physiological and emotional reactions. *Scientific Reports* 11: 16407. https://doi.org/10.1038/s41598-021-95800-2.

Coschignano, Serena, Gosse Minnema, and Chiara Zanchi. 2023. Explaining the distribution of implicit means of misrepresentation: A case study on Italian immigration discourse. *Journal of Pragmatics* 213: 107-125. https://doi.org/10.1016/j.pragma.2023.06.002.

Eberl, Jacob-Morit, Christin E. Meltzer, Tobias Heidenreich, Beatriz Herrero, Nora Theorin, Fabienne Lind, Rosa Berganza, Hajo G. Boomgaarden, Christian Schemer, and Jesper Strömbäck. 2018. The European media discourse on immigration and its effects: A literature review. *Annals of the International Communication Association* 42(3): 207-223. https://doi.org/10.1080/23808985.2018.1497452.

Fairclough, Norman. 1989. *Language and power*. London: Longman.

Fairclough, Norman. 1995. *Critical discourse analysis*. London: Longman.

Fairclough, Norman. 1998. *Discourse and social change*. Cambridge: Polity Press.

Farris, Emily M. and Heather S. Mohamed. 2018. Picturing immigration: How the media criminalizes immigrants. *Politics Groups Identities* 6(4): 814-824. https://doi.org/10.1080/21565503.2018. 1484375.

Fuller, Janet M. 2024. Media discourses of migration: A focus on Europe. *Language and Linguistics Compass* 18(4): e12526. https://doi.org/10.1111/lnc3.12526.

Giorgi, Alberta and Thommaso Vitale. 2017. Migrants in the public discourse: Between media, policy and public opinion. In Stefania Marino and Judith Roosblad (eds.), *Trade unions and migrant workers. New contexts and challenges in Europe*, 66-89. New York: Edward Elgar Publishing.

Goffman, Erving. 1963. *Stigma: Notes on the management of spoiled identity*. New York, NY: Simon & Schuster.

Goldberg, Yoav. 2014. *Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method*. arXiv preprint arXiv:1402.3722.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole and Jerry Morgan (eds.), *Syntax and semantics*, 41-58. New York: Academic Press.

Halman, Helena. 2011. Political correctness, euphemism, and language change: The case of 'people first'. *Journal of Pragmatics* 43(3): 828-840.

Hammersley, Martyn and Roger Gomm. 1997. Bias in social research. *Sociological Research Online* 2(1): 7-19.

Harris, Casey and Jeff Gruenwald. 2022. News media trends in the framing of immigration and crime, 1990–2013. *Social Problems* 67(3): 452-470. https://doi.org/10.1093/socpro/spz024.

Hartigan, John A. and M. Anthony Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C* (*Applied Statistics*) 28(1): 100-108.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. 2005. A brief survey of text mining. *LDV Forum* 20(1): 19-62.

Jaspal, Rusi and Marco Cinnirella. 2010. Media representations of British Muslims and hybridised threats to identity. *Contemporary Islam* 4(3): 289-310.

Kim, Hye-Jeong, Min-Seo Kim, and Seongmin Mun. 2024. Analysis of descriptive course evaluations applying word ontology: Focusing on student-centered narrative lecture evaluation platform, Everytime. *The Journal of the Korea Contents Association* 24(4): 560-570. https://doi.org/10.5392/JKCA.2024.24.04.560.

Kroon, Anne, Alena Kluknavská, Rens Vliegenthart, and Hajo Boomgaarden. 2016. Victims or perpetrators? Explaining media framing of Roma across Europe. *European Journal of Communication* 31(4): 375-392.

Layne, Nathan. 2023. Trump repeats 'poisoning the blood' anti-immigrant remark. Reuters.

https://www.reuters.com/world/us/trump-repeats-poisoning-blood-anti-immigrant-remark-2023-12-16/.

Li, Wentian. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38(6): 1842-1845.

Lirola, Maria M. 2022. A critical discourse study of the portrayal of immigrants as non-citizens in a sample from the Spanish press. *Language and Migration* 14(1): 69-91. https://doi.org/10.37536/LYM.14.1.2022.1053.

Liu, Zhaoming. 2023. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication* 3(2): 224-244. https://doi.org/10.1515/jtc-2023-0019.

Lutz, Philipp and Marco Bitschnau. 2022. Misperceptions about immigration: Reviewing their nature, motivations and determinants. *British Journal of Political Science* 53(2): 1-16. https://doi.org/10.1017/S0007123422000084.

MacQueen, James. 1967. Some methods for classification and analysis of multivariate observations. In Lucien le Cam and Jerzy Neyman (eds.), *The Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 281-297. Berkeley, CA: University of California Press.

Madrigal, Guadalupe and Stuart Soroka. 2023. Migrants, caravans, and the impact of news photos on immigration attitudes. *The International Journal of Press/Politics* 28(1): 49-69. https://doi.org/10.1177/19401612211008430.

Merolla, Jennifer, S. Karthick Ramakrishnan, and Chris Haynes. 2013. "Illegal," "undocumented," or "unauthorized": Equivalency frames, issue frames, and public opinion on immigration. *Perspectives on Politics* 11(3): 789–807. https://doi.org/10.1017/S1537592713002077.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26.
https://doi.org/10.48550/arXiv.1310.4546.

Morera-Bello, Maryurena Lorenzo, Daniel Buraschi, and Naira Delgado. 2026. Same but different: The dehumanised perception of immigrants, attitudes towards interpersonal relationships, and inclusion policies. *International Journal of Intercultural Relations* 111: 102369.

Morris, Lydia. 2025. *Citizen rights, migrant rights and civic stratification*. London: Routledge.

Musolff, Andreas. 2021. The scenario of (im-)migrants as scroungers and/or parasites in British media discourses. In Barbara Schmidt-Haberkamp, Marion Gymnich, and Klaus P. Schneider (eds.), *Representing poverty and precarity in a postcolonial world*, 246-260. Leiden: Brill.

Pandiani, Delfina S., Erik Tjong Kim Sang, and Davide Ceolin. 2025. Toxic memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Computation and Language*. arXiv:2406.07353. https://doi.org/10.48550/arXiv.2406.07353.

Reisigl, Martin and Ruth Wodak. 2009. The discourse-historical approach (DHA). In Ruth Wodak and Michael Meyer (eds.), *Methods for critical discourse analysis*, 87-121. London: Sage Publishing.

Seo, Soomin and Sebgi Başak Kavakli. 2022. Media representations of refugees, asylum seekers and immigrants: A meta-analysis of research. *Annals of the International Communication Association* 46(3): 159-173. https://doi.org/10.1080/23808985.2022.2096663.

Shahbazi, Nina, Yin Lin, Abolfazi Abolfazi, and Hosagrahar V. Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys* 55(13s): 1-39. https://doi.org/10.1145/ 3588433.

Shattell, Mona M. 2009. Stigmatizing language with unintended meanings: "Persons with mental illness" or "mentally ill persons". *Issues in Mental Health Nursing* 30(3): 199. https://doi.org/10.1080/01612840802694668.

Shin, Gyu-Ho, Boo-Kyung Jung, and Seongmin Mun. 2024. Transformer-based text similarity and second language proficiency: A case of written production by learners of Korean. *Natural Language Processing Journal* 6(3): 100060. https://doi.org/10.1016/j.nlp.2024.100060.

Shin, Hye-Jung and Seongmin Mun. 2022. A text mining analysis of translator's style: Comparing translating styles of 'as if' constructions. *T&I Review* 12(1): 97-120. https://doi.org/10.22962/tnirvw.

Trost, Igor. 2023. The representation of refugees in the German media discourse from 2015 to 2017: A corpus-based approach on frames and deep cases in constructions. In Annamária Fábián (ed.), *The representation of REFUGEES and MIGRANTS in European national media discourses from 2015 to 2017: A contrastive approach* (*Corpus linguistics*), 121-137. Berlin: Springer.

Ullmann, Stefanie. 2023. The representation of refugees, migrants and migration in the British media discourse of 2015: A contrastive, corpus-based approach. In Annamári Fábián (ed.), *The representation of REFUGEES and MIGRANTS in European national media discourses from 2015 to 2017: A contrastive approach* (*Corpus linguistics*), 139-162. Berlin: Springer.

Urman, Aleksandra, Mykola Makhortykh, and Roberto Ulloa. 2022. Auditing the representation of migrants in image web search results. *Humanities and Social Sciences Communications* 9: 1-16. https://doi.org/10.1057/s41599-022-01144-1.

van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(86): 2579-2605.

van Dijk, Teun A. 1993. Principles of critical discourse analysis. *Discourse & Society* 4(2): 249-283. https://doi.org/10.1177/0957926593004002006.

van Dijk, Teun A. 1998. *Ideology: A multidisciplinary approach*. London: Sage Publications.

van Dijk, Teun A. 2015. Critical discourse analysis. In Deborah Tannen, Heidi E. Hamilton, and Ddeborah Schiffrin (eds.), *The handbook of discourse analysis*, 466-485. Oxford: Wiley-Blackwell.

Vergeer, Maurice, Marcel Lubbers, and Peer Scheepers. 2000. Exposure to newspapers and attitudes toward ethnic minorities: A longitudinal analysis. *The Howard Journal of Communication* 11(2): 127-143.

Vollmer, Bastian and Serhat Karakayali. 2018. The volatility of the discourse on refugees in Germany. *Journal of Immigrant & Refugee Studies* 16(1-2): 118-139. https://doi.org/10.1080/15562948.2017. 1288284.

Wodak, Ruth. 2007. Pragmatics and critical discourse analysis. A cross-disciplinary inquiry. *Pragmatics and Cognition* 15(1): 203-225.

Zannettou, Savvas, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In Emilio Ferrara and Pauline Leonard (eds.), *The Proceedings of the 12th ACM Conference on Web Science*, 125-134. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3394231.3397902.

Zipf, George K. 1932. *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Seongmin Mun
Assistant Professor
Department of English Language and Literature
Kyungpook National University
80 Daehak-ro, Buk-gu
Daegu, 41566, Korea
E-mail: seongminmun@knu.ac.kr

Hye-Kyung Lee
Professor
Department of English Language and Literature
Ajou University
206 Worldcup-ro, Yeongtong-gu, Suwon
Gyeonggi-do, 16499, Korea
E-mail: hklee@ajou.ac.kr

Received: 2026. 03. 17.
Revised: 2026. 04. 09.
Accepted: 2026. 04. 24.