Academic Education Publishing House

Higher Education and Practice Vol. 3 No. 3, 2026

A Corpus-Assisted Discourse Study on Identity Construction of Chinese Overseas Students in News Media

Luyu Yue, Jialing Wu, Ruinan Zhao*

School of Foreign Languages, Guangzhou Maritime University, Guangzhou, China

**Corresponding Author*

Abstract: This study examines the representation of Chinese overseas students in South China Morning Post (SCMP) coverage from 1992 to 2024 using a corpus-assisted discourse approach. Drawing on a purpose-built corpus, the analysis identifies patterns of media attention and recurring discursive domains, including education, transnational mobility, employment, social life, and public health. The findings show that Chinese students are constructed through multiple and sometimes competing identities, such as restricted learners, cultural exchange participants, vulnerable individuals, and economic contributors. These representations reflect shifting media priorities and broader global contexts, highlighting how discourse contributes to the construction of social identities.

Keywords: Corpus-Assisted Discourse Studies; Chinese Overseas Students; Media Discourse; Identity Construction

1. Introduction

With the intensification of globalization and the expansion of international higher education, Chinese students studying abroad have become one of the most visible and significant groups within global student mobility. As their numbers have increased, so too has their presence in public and media discourse, where they are frequently discussed in relation to academic performance, economic contribution, and cross-cultural interaction. Earlier research has noted the rapid growth of this group, particularly since the implementation of China's reform and opening-up policies, which enabled large-scale outbound mobility and contributed to their increasing global visibility [1].

At the same time, existing studies suggest that media representations of international students are not neutral reflections of reality but are shaped by broader discursive and ideological frameworks. For example, Nancy Abelmann and Jiyeon Kang [2] show that U.S. media discourse often frames Chinese international students through a market-oriented logic, emphasizing their economic value within higher education systems. This highlights the need to examine not only the presence of Chinese students in media reporting, but also the ways in which they are discursively constructed.

In this context, media discourse plays a crucial role in shaping public understanding of this social group. The representation of overseas Chinese students is closely linked to broader processes of image construction, extending beyond individual reputation to the level of national image and international perception. Media coverage therefore functions not only as a source of information but also as a site where meanings, evaluations, and identities are produced and circulated.

Against this background, the present study investigates how Chinese overseas students are represented in the *South China Morning Post* (SCMP) over the period from 1992 to 2024. Adopting a corpus-assisted discourse studies (CADS) approach [3,4], the study combines quantitative and qualitative analysis to examine patterns of frequency, keywords, and collocation. Rather than treating media texts as transparent reflections of social reality, this approach allows for an exploration of how recurring linguistic patterns contribute to the construction of social identities.

In doing so, this research contributes to the growing body of work in corpus-assisted discourse analysis by providing a systematic, diachronic examination of media representations of Chinese overseas students. It also offers insights into how media discourse may shape public perceptions of this group within broader contexts of globalization and international education.

2. Literature Review

2.1 Identity Construction in NewsMedia

Identity is not merely an individual's psychological attribute but a reflection of the intricate interplay between the individual, society, and groups. Its formation is a dynamic, ongoing process that must be examined within specific socio-cultural contexts [5]. Scholars have long argued that identity bridges individual psychology and broader social structures. From a discursive perspective, identity is understood as a social construct, where language, power, and interaction play constitutive roles [6]. This is particularly evident in the formation of collective or social identity, which serves to demarcate group boundaries and is forged through daily interactions and shared representations.

News media is a primary site where these identities are publicly negotiated, reinforced, or challenged. Media discourse does not neutrally reflect social reality but actively constructs it through selective representation. Discourse producers employ specific linguistic strategies to represent social actors, thereby positioning them within "in-group" or "out-group" categories and influencing public perception [7]. This power of media representation is evident in cross-national contexts.

The integration of corpus linguistics with critical discourse analysis (CDA) provides a powerful methodological synergy for empirically investigating these mediated identity constructions. Corpus tools allow for the systematic analysis of large volumes of text, revealing recurrent linguistic patterns and ideological leanings that might be opaque in close reading alone [8]. Sun and Gao [9] combined corpus-assisted discourse analysis with van Leeuwen's social actor framework to deconstruct the collective identity of Chinese international students in Japanese mainstream newspapers. Their study revealed that this identity was discursively constructed through four primary frames: "knowledge seekers and laborers", "volunteers", "grass-root messengers", and "challengers of social order". Crucially, they found that positive individual actions were often personalized, aligning them with the reader's in-group, while negative behaviors were generalized through collective "categorization", amplifying individual transgressions into a perceived group threat [9]. This study effectively demonstrates how corpus-assisted CDA can unpack the nuanced linguistic mechanisms behind media-driven identity formation.

2.2 Corpus-Assisted Discourse Study

Over the past two decades, Corpus-Assisted Discourse Studies (CADS) has evolved into a robust interdisciplinary methodology, effectively bridging the quantitative rigor of corpus linguistics with the qualitative depth of discourse analysis [4]. At its core, a corpus is defined as a large, structured set of texts stored electronically for linguistic analysis [3]. The construction of a specialized corpus involves systematic steps including data collection, organization, annotation, and storage, enabling researchers to move beyond intuitive claims to evidence-based analysis of language patterns across large datasets [8].

The synergy between corpus methods and discourse analysis, particularly Critical Discourse Analysis (CDA), forms the foundation of CADS. This approach allows researchers to uncover both dominant and subtle linguistic patterns—such as collocations, semantic prosody, and grammatical features—that reveal underlying ideologies and power dynamics within texts.

The analytical power of CADS is well-demonstrated in media studies. For instance, research comparing international news outlets, such as analyses of China Daily and The New York Times, utilizes CADS to deconstruct how lexical choices, grammatical structures, and metaphors strategically position differing editorial stances and construct social realities. Such studies illustrate how CADS can model the discursive frameworks employed by media institutions.

Methodologically, CADS is characterized by its hybrid approach, integrating statistical frequency analysis with detailed textual interpretation. This dual focus ensures that broad trends identified through corpus queries are grounded in close reading of their contextual usage. A meta-review of over a decade of research confirms that CADS provides effective and versatile analytical tools for discourse researchers, applicable across diverse topics from media bias to identity construction.

3. Methodology

3.1 Data Collection and Corpus Building

This study is based on a purpose-built news corpus compiled from the South China Morning Post (SCMP), a major English-language newspaper with extensive coverage of China-related issues. Articles were retrieved using the LexisNexis database, which provides access to archived news content over an extended time span.

To ensure relevance and consistency, the search was conducted using the keywords "Chinese student" OR "mainland student". These terms were selected because they are commonly used in English-language media to refer to students from China and allow for the systematic retrieval of relevant texts. The search covered the period from 1992 to 2024, enabling a diachronic analysis of media representation across more than three decades.

After removing duplicates, irrelevant items, and incomplete texts, the final corpus consists of 677,145 tokens. The corpus was then processed and analyzed using Wordsmith Tools, a widely used software package in corpus linguistics. The resulting dataset provides a sufficiently large and representative sample for identifying recurring linguistic patterns and examining how Chinese overseas students are represented in media discourse.

3.2 Analytic Procedure

The analysis adopts a corpus-assisted discourse studies (CADS) approach, combining quantitative measures with qualitative interpretation [3,4]. Frequency analysis is used to trace the distribution of the node terms *"Chinese students"* and *"mainland students"* across the corpus, with both raw and normalized frequencies examined to identify patterns of diachronic variation.

Keyword analysis is then conducted by comparing the SCMP corpus with BNC as a reference corpus in Wordsmith Tools, enabling the identification of statistically salient lexical items. These keywords are interpreted in context and grouped into broader discursive domains, providing insight into recurring thematic patterns in the data.

Collocation analysis further explores the lexical environment surrounding the node terms, identifying significant co-occurring words that contribute to meaning construction. These patterns are examined through concordance lines, allowing for close analysis of how Chinese overseas students are represented in specific contexts. Through this integrated procedure, the study links quantitative patterns with qualitative interpretation to reveal how recurring linguistic features shape the discursive construction of this social group.

4. Findings

4.1 Diachronic Analysis of Representation of Chinese Students

Using Wordsmith Tools, we calculated the yearly frequency of the terms *"Chinese students"* and *"mainland students"* across the corpus (see Figure 1). The results suggest a fluctuating rather than linear pattern of media attention. Peaks are observable around 2005, 2007, 2015, 2019, and 2023, while relatively low frequencies occur in earlier years such as 1992, 1995, 1999, as well as in 2009 and 2016.

The increases in 2005 and 2007 can be tentatively associated with China's growing integration into global systems and the expansion of international student mobility. As noted by Woo and Wang [10], media attention to international students often reflects their embeddedness in global economic flows. During this period, China's rapid economic growth and expanding middle class contributed to a rise in outbound student numbers. This demographic shift may have increased the newsworthiness of "Chinese students" as a social category, particularly in relation to themes such as education, mobility, and transnational exchange. Importantly, the corpus data does not in itself demonstrate causality; rather, it suggests that media discourse became more attentive to this group during a period of intensified global engagement.

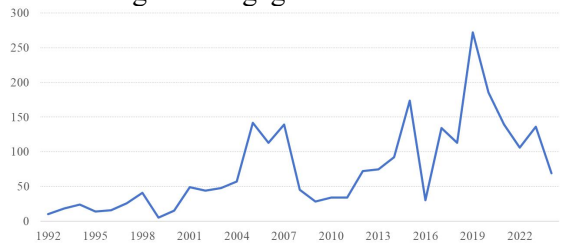

Figure 1. Frequency of Chinese Students

A similar pattern can be observed in 2015 and 2019, where higher frequencies may reflect the intersection of internationalization policies and media agendas. For instance, initiatives such as the Belt and Road Initiative and high-level international forums likely contributed to a

discursive environment in which cross-border mobility—and by extension, overseas students—became more salient topics. In CADS terms, this indicates a potential shift in discursive prominence, where "Chinese students" are more frequently invoked within broader narratives of globalization, diplomacy, and economic cooperation.

The peak in 2023 can be interpreted in relation to the post-pandemic reopening and renewed emphasis on international education and talent mobility. As China re-engaged with global academic exchanges following COVID-19 restrictions, media discourse may have reactivated the category of "Chinese students" as part of narratives about recovery, mobility, and future development. Again, the frequency increase should be understood as reflecting changing discursive priorities rather than a direct index of real-world student numbers.

In contrast, the relatively low frequencies in the early 1990s and late 1990s (e.g., 1992, 1995, 1999) likely reflect both the smaller scale of outbound student mobility at the time and the lower salience of this group in media discourse. Although 1992 marked a significant moment in China's economic reform, overseas education had not yet become a widespread social phenomenon, and thus did not constitute a prominent media topic.

Overall, the diachronic pattern indicates that the representation of Chinese overseas students in the media is closely tied to shifting discursive contexts, particularly those related to globalization, economic development, and international mobility. Rather than reflecting a simple increase in attention, the data points to periods in which "Chinese students" become more or less discursively salient, depending on how they are positioned within broader news narratives.

4.2 Topics of Chinese Students in the Media

This study uses Wordsmith Tools to generate keyword lists to examine how they cluster into broader discursive domains and contribute to the construction of Chinese overseas students as a social group in media discourse.

A prominent cluster of keywords relates to education, including items such as *IELTS*, *TOEFL*, *enrolment*, *course*, *academic*, *graduation*, and *diploma*. These lexical patterns suggest that Chinese students are consistently represented through what can be described as an educational trajectory discourse, in which their experiences are structured as a linear progression from preparation to completion. Within this discourse, the presence of keywords such as *scholarship* and *internship* further reinforces a meritocratic framing, positioning students as competitive individuals striving for academic and professional advancement. This does not merely reflect educational realities but constructs an identity in which Chinese students are primarily understood through achievement, mobility, and institutional participation.

Another significant pattern emerges in the geographical distribution of keywords, which include both Chinese regions (e.g., *Shanghai*, *Hubei*) and international destinations (e.g., *Australia*, *U.S.*, *Seoul*). These references point to the transnational nature of Chinese student mobility, but more importantly, they function discursively to position these students within a broader global framework. When considered alongside keywords such as relations and geopolitical, it becomes evident that Chinese students are often embedded within narratives of international relations. In this sense, they are constructed not only as individuals pursuing education abroad but also as figures whose presence is intertwined with diplomatic, economic, and geopolitical contexts. This reflects a key insight in CADS that social actors are frequently represented through their relation to wider ideological structures rather than in isolation [3].

The keyword set also reveals a strong emphasis on employment and future prospects, with items such as *career*, *job*, *employment*, and *skills*. These patterns suggest that media discourse frames Chinese students in terms of their anticipated economic roles, aligning them with a broader human capital discourse. In this representation, overseas education is implicitly constructed as an investment that enhances employability and productivity. Such a framing positions Chinese students as future contributors to both national and global economies, reinforcing a utilitarian perspective on education that extends beyond individual development.

In contrast, another cluster of keywords introduces a more complex and sometimes negative dimension to representation. Words such as *protest*, *vigil*, *racism*, *warning*, *scammers*, and *theft* indicate that Chinese students are also associated with themes of

social tension, risk, and vulnerability. These patterns suggest a shift in semantic prosody, where the surrounding discourse carries evaluative meanings that may problematize this group. This dual positioning produces a more ambivalent identity, combining elements of agency with vulnerability.

Finally, the presence of keywords such as *virus*, *masks*, and *lockdown* highlights the impact of the COVID-19 pandemic on media representations. These terms indicate that Chinese students have been incorporated into broader crisis-related discourses, where issues of mobility, health, and regulation become central. In this context, their identity is reshaped through what can be seen as a process of discursive recontextualization, as global health concerns intersect with existing narratives of international mobility. As a result, Chinese students are not only represented as learners or economic actors but also as subjects navigating restrictions, uncertainties, and, at times, heightened scrutiny.

4.3 Analysis of Identity Constructed

Building on the collocational and concordance analysis, this section examines how Chinese overseas students are represented in the corpus. The recurrent collocations and concordance patterns are understood to reflect not only linguistic regularities but also systematic ways of representing social groups [3,4]. From this perspective, identity is not treated as fixed, but as emerging from repeated patterns of language use across texts.

One prominent pattern in the data is the construction of Chinese students as restricted learners. Collocational analysis shows frequent co-occurrence with terms such as *visa*, *restriction*, *policy*, *review*, and *refusal*, suggesting that their educational experiences are often framed in relation to institutional control. Concordance lines further indicate that these restrictions are associated with visa limitations, administrative delays, and policy changes. This pattern contributes to a representation in which Chinese students are positioned as subjects whose mobility is constrained by external forces, rather than as fully autonomous participants in international education. The repeated association with bureaucratic processes creates a discourse in which limitation and dependency become salient features of their identity.

At the same time, the corpus also constructs Chinese students as participants in cultural exchange. Collocates such as *international*, *exchange*, *contribution*, and s*ociety* point to a more positive evaluative environment. Concordance evidence suggests that Chinese students are frequently described in terms of their interaction with local communities, engagement in academic activities, and contribution to cultural diversity. This pattern reflects a more agentive representation, in which students are positioned as active participants in transnational communication and mutual exchange. The contrast with the "restricted learner" pattern highlights the coexistence of different evaluative orientations within the corpus.

Another significant representation emerges through collocations related to risk and vulnerability, including words such as *racism*, *attack*, *abuse*, *safety*, and *warning*. These lexical patterns indicate a consistent association between Chinese students and experiences of insecurity or discrimination. The concordance lines often foreground incidents of violence or social tension, which contributes to a negative semantic prosody surrounding this group. As a result, Chinese students are discursively constructed as vulnerable individuals navigating potentially hostile environments. This pattern does not simply describe isolated events; rather, the repetition of such associations reinforces a broader perception of risk linked to their presence abroad.

In addition, the corpus reveals a strong tendency to represent Chinese students as economic contributors and consumers. Collocates such as *fees, revenue*, *market*, *rent*, and spending highlight their financial significance within host countries. Concordance analysis shows that Chinese students are frequently quantified in terms of their economic impact, particularly in relation to tuition income and local consumption. This creates a representation in which their value is framed in economic terms, emphasizing their role within the global education market. The repeated aggregation of students into large numerical groups further strengthens this effect, shifting attention from individual experiences to collective economic influence.

Finally, the presence of pandemic-related collocates such as *virus*, *lockdown*, and *quarantine* introduces a health and crisis

dimension to their representation. These patterns indicate that Chinese students are incorporated into broader discourses surrounding COVID-19, where issues of mobility, restriction, and uncertainty become central. The co-occurrence of these terms with references to travel and study suggests that the pandemic reshaped the way this group is discussed, linking their identity more closely to disruption and regulation. This adds another layer to their representation, reinforcing the idea that their experiences are shaped by external conditions beyond their control.

5. Conclusion

This study has examined the representation of Chinese overseas students in *South China Morning Post* coverage from 1992 to 2024 through a corpus-assisted discourse approach. By combining diachronic frequency analysis, keyword-based thematic analysis, and collocational investigation, the study demonstrates how this social group is discursively constructed across time rather than simply described.

The diachronic analysis shows that references to "Chinese students" and "mainland students" fluctuate rather than follow a linear trend. The keyword analysis reveals that media coverage is organized around several recurring discursive domains, including education, transnational mobility, career development, social life, and public health. These domains do not function merely as neutral "topics"; instead, they structure how Chinese students are understood within the media. In particular, the prominence of education- and employment-related keywords reflects a tendency to frame students in terms of achievement and future economic value, while the presence of geopolitical and pandemic-related vocabulary situates them within broader global and crisis-oriented narratives.

Building on these patterns, the collocational analysis demonstrates that Chinese overseas students are constructed through multiple identity positions. They are frequently represented as restricted learners whose mobility is shaped by institutional constraints, as participants in cultural exchange who contribute to host societies, as vulnerable individuals exposed to risks such as discrimination and violence, and as economic contributors whose financial impact is foregrounded. These representations are not mutually exclusive; rather, they coexist within the corpus, producing a complex and sometimes contradictory discursive image of this group.

Taken together, the findings highlight that the identity of Chinese overseas students in media discourse is dynamic and context-dependent. It emerges from recurring linguistic patterns that reflect broader concerns related to globalization, education, and social change. This underscores the importance of examining not only what is said about a social group, but also how repeated patterns of language contribute to shaping its public representation.

This study is not without limitations. The impact of the COVID-19 pandemic introduces a degree of disruption to the diachronic analysis, as pandemic-related vocabulary (e.g., lockdown, mask) becomes highly salient after 2020 and may temporarily skew patterns of representation. In addition, the analysis is based on a single news outlet, which limits the generalizability of the findings. Future research could address these limitations by constructing a comparative corpus across different media sources and national contexts, allowing for a more systematic examination of how geopolitical factors influence the representation of international students. Further work could also integrate audience reception data to explore how such representations are interpreted and negotiated by readers, as well as explore computational methods for analyzing multimodal discourse.

Acknowledgments

This work is supported by Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau [Grant ID: 2024311985].

References

[1] Simon Marginson (2014).Student self-formation in international education. Journal of Studies in International Education, 18(1), 6–22.

[2] Nancy Abelmann, & Jiyeon Kang (2014).A fraught exchange? U.S. media discourse on Chinese international undergraduates. Journal of Studies in International Education, 18(4), 382–397.

[3] Paul Baker (2006). Using Corpora in Discourse Analysis. London: Continuum.

[4] Alan Partington, Duguid, A., & Taylor, C. (2013). Patterns and Meanings in Discourse. Amsterdam: John Benjamins.
[5] Koller, V. (2012). How to analyze collective identity in discourse: Textual and contextual parameters. Critical Approaches to Discourse Analysis across Disciplines, 5(2), 19–38.
[6] Reisigl, M., & Wodak, R. (2009).The discourse-historical approach (DHA). In R. Wodak & M. Meyer (Eds.), Methods of Critical Discourse Analysis (2nd ed., pp. 87–121). London: Sage.
[7] Van Leeuwen, T. (2008). Discourse and practice: New tools for critical discourse analysis. Oxford University Press.
[8] McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.
[9] Sun, C., & Gao, H. (2021). The collective identity of ‘Chinese overseas students’ in Japan’s mainstream newspapers: A corpus-assisted discourse study. Foreign Languages and Their Teaching, 318(3), 53-61. https://doi.org/10.13458/j.cnki.flatt.004775
[10]Woo, E., & Wang, L. (2024). Crisscrossing scapes in the global flow of elite mainland Chinese students. Higher Education, 87, 551–566.