SOUTH ASIAN JOURNAL OF SOCIAL SCIENCES & HUMANITIES

ISSN: 2582-7065 (Online) SAJSSH, VOL 5, ISSUE 6, PP. 13-25

The Use of Corpus Analysis in Analysing Tourism Texts

Bashar Abdulkareem Alali[1], Afiza Mohamad Ali[2], Ainul Azmin[3], Rafidah Sahar[4]

[1,3]Department of English Language & Literature, AbdulHamid A. AbuSulayman Kulliyyah of Islamic Revealed Knowledge and Human Sciences, International Islamic University Malaysia.

[2,4]Kulliyyah of Sustainable Tourism and Contemporary Languages, International Islamic University Malaysia

Corresponding Author: Bashar Abdulkareem Alali, **E-mail:** alali.bashar@hotmail.com

Received: 25th August 2024 **Accepted:** 15th October 2024 **Published:** 5th December 2024

ABSTRACT

This study investigates the application of corpus analysis in tourism studies, examining how researchers utilise this methodological approach to understand tourism texts' language patterns and features. By reviewing various journal articles, the paper explores the methods, tourism texts, and tools employed by researchers to analyse tourism discourse. Corpus analysis, a quantitative approach to linguistic inquiry, provides valuable insights into the language used in tourism-related materials. Researchers have employed corpus analysis to examine themes like destination image, cultural events, language patterns, and translation studies. The study highlights the use of concordance tools, such as Wordsmith and AntConc, in analysing tourism texts and uncovering linguistic patterns and rhetorical strategies. The findings demonstrate the versatility of corpus analysis in understanding the nuances of tourism discourse. Future research could explore further integrations of corpus analysis with other methodologies, such as ethnography and multimodality, to gain a more comprehensive understanding of the complex dynamics of tourism language.

Keywords: Corpus analysis, Tourism, ESP

DOI: 10.48165/sajssh.2024.5602

INTRODUCTION

Corpus analysis (plural corpora) is a linguistic approach to analyse collected and stored electronic ‘real-life’ language samples, which are gathered systematically or randomly, such as speeches, magazine articles, and text messages, to determine certain rules of language use, and grammatical or lexical patterns. Corpus studies are often classified to be a subset of discourse analysis because they study the usage of linguistic forms in different context (Biber, Connor, & Upton, 2007). For instance, words are defined in terms of them collocates (the words that frequently appear in the context of the speech). Grammatical variety I s also defined in terms of the words and other grammatical structures encountered in the context. As such, corpus linguistics research fits neatly within the first approach to discourse which is the examination of language use (Biber et al., 2007).

In the realm of English for Specific Purposes (ESP) and tourism field, corpus analysis is a valuable tool for examining tourism texts, as it provides a systematic and quantitative approach to understanding the language patterns and features of these texts (Sun, 2022; Vojnović, 2023). By harnessing computational linguistics techniques using tools such as Antconc, Sketch Engine, wordsmith, Wmartix and others, corpus analysis facilitates a meticulous exploration of textual data, enabling researchers to uncover underlying patterns, linguistic structures, and lexical choices prevalent within tourism discourse.

Furthermore, corpus linguistics can be used to investigate the cultural and social dimensions of tourism discourse. By analysing the language used in tourism texts, researchers can gain insights into the values, beliefs, and attitudes associated with tourism. For example, corpus analysis can identify how tourism discourse constructs and perpetuates stereotypes about different cultures and destinations.

In conclusion, corpus linguistics offers a valuable approach to understanding the language of tourism. By providing a systematic and quantitative method for analysing large corpora of tourism texts, researchers can identify key patterns, structures, and features that characterize this discourse domain. This information can be used to inform language teaching and learning, translation, and other applications related to tourism communication.

DOI: 10.48165/sajssh.2024.5602

This paper aims to explore how researchers use corpus analysis in tourism studies by looking at various journal articles. It shows methods, tourism texts and tools used by researchers to understand the language of tourism.

LITERATURE REVIEW

Corpus Analysis as a Research Method

One of the primary advantages of corpus analysis lies in its ability to provide empirical evidence and insights into the language usage within different genre texts such as tourism texts.(McEnery, Brezina, Gablasova, & Banerjee, 2019). Through the compilation and analysis of large-scale corpora comprised of diverse tourism-related documents such as brochures, websites, travel guides, and promotional materials, researchers can identify recurring themes, rhetorical strategies, and linguistic devices employed to engage and entice potential visitors (Ortikov, 2023). This empirical foundation allows researchers to move beyond anecdotal observations, offering a rigorous and systematic examination of the language employed within the tourism domain.

Moreover, corpus analysis enables researchers to adopt a quantitative approach towards linguistic inquiry, allowing for the measurement and comparison of linguistic phenomena across different tourism texts. By employing statistical methods and computational tools, researchers can quantify the frequency of specific lexical items, syntactic structures, or discourse features within the corpus, thereby illuminating salient linguistic patterns and trends. This quantitative dimension enhances the robustness of the analysis, facilitating a more nuanced understanding of the linguistic repertoire utilized within tourism discourse.

Furthermore, corpus analysis facilitates the exploration of both synchronic and diachronic aspects of tourism language, enabling researchers to examine linguistic variation and evolution over time. By compiling corpora spanning different periods or geographic regions, researchers can trace the development of linguistic norms, discursive conventions, and cultural representations within the tourism domain. This diachronic perspective not only enriches our understanding of the historical dynamics shaping tourism discourse but also sheds light on the socio-cultural contexts influencing language use in tourism communication.

In summary, corpus analysis offers a methodologically rigorous and empirically grounded approach to the examination of tourism texts, enabling researchers to uncover

DOI: 10.48165/sajssh.2024.5602

intricate linguistic patterns, quantify linguistic phenomena, and explore the evolution of tourism language over time. By leveraging computational techniques and large-scale corpora, corpus analysis enhances our comprehension of the language dynamics inherent within the multifaceted realm of tourism discourse.

Methods in Corpus Analysis

Corpus analysis is used to study tourism texts by assessing the keyness of adjectives in promotional materials of the adventure tourism area in English and comparing them to other texts (Durán-Muñoz, 2019). It is used to provide a genre-based analysis of brief tourist information (BTI) texts on tourism destinations' websites, by identifying the moves and steps of the corpus and studying the linguistic forms to realise each move (Huang, 2015). Corpus analysis is used in translation studies, such as translating tourism texts from one language to another, by analysing the translated language of tourism and identifying the language patterns and features of translated tourism texts (Gandin, 2013). Corpus analysis provides insights into the language patterns and features of tourism English, tourism websites, hotel online reviews, and translated tourism texts, which can help to improve the quality of tourism services, translation strategies, and tourism research. However, there are also challenges in using corpus analysis in tourism texts, such as the need for standardized large-scale multi-language tourism corpora and improved interpretability of deep learning techniques (Huang, 2015).

Corpus analysis of tourism texts can reveal common themes or topics related to tourism and travel. Some of these themes include:

1. Destination image: Corpus analysis can help to explore how the image of a tourism destination is discursively created and projected in BTI texts (Huang, 2015).

2. Cultural events and city image: Corpus analysis can be used to investigate the impact of cultural events on city image, such as Rotterdam, cultural capital of Europe 2001(Huang, 2015).

3. Language and linguistic patterns: Corpus analysis can be used to analyse the language and linguistic patterns of tourism texts, such as the use of adjectives in promotional texts of the adventure tourism domain in English (Gandin, 2013).

DOI: 10.48165/sajssh.2024.5602

4. Translation studies: Corpus analysis can be used in translation studies to analyse the translated language of tourism and identify the language patterns and features of translated tourism texts (Altameemi & Altamimi, 2023; Huang, 2015).

5. Genre analysis: Corpus analysis can be used to give a genre-based analysis of BTI texts on websites of tourism destinations, by identifying the moves and steps of the corpus and examining the linguistic forms to realize each move(Huang, 2015).

6. Thematic analysis: Corpus analysis can be used to understand themes or topics of a corpus through a classification process using long short-term memory (LSTM) models (Altameemi & Altamimi, 2023).

7. Online reviews: Corpus analysis can be used to analyse the textual features of online reviews, such as a polymerization topic sentiment model (Altameemi & Altamimi, 2023).

8. Keyword-based study: Corpus analysis can be used to study the thematic vocabulary in tourism texts, such as British weather news (Altameemi & Altamimi, 2023).

These themes or topics can help to improve the quality of tourism services, translation strategies, and tourism research. However, there are also challenges in using corpus analysis in tourism texts, such as the need for standardized large-scale multi-language tourism corpora and improved interpretability of deep learning techniques (Lee, 2016).

METHODOLOGY

This article seeks to examine the application of corpus analysis in the analysis of texts within the tourism genre. It aims to demonstrate the types of promotional texts that have undergone corpus analysis, along with the frameworks and tools utilized in these studies. The exploration of relevant literature employed a targeted research strategy focused on online databases and Google Scholar. The search criteria included terms such as promotion, corpus analysis, and tourism, with a specific emphasis on journal articles, and conference papers published exclusively in English.

RESULTS

Henry and Roseberry (1996) studied corpus analysis of 44 'Brief Tourist Information' (BTI) texts in Standard English aiming to compare the linguistic features of three obligatory moves:

DOI: 10.48165/sajssh.2024.5602

Location, Facilities/Activities, and Description. Using a computerised concordance program, they analysed these moves based on Halliday and Hasan's (1989), Swales' (1990) and Bhatia's (1993) frameworks. The researchers examined various linguistic features, including discourse functions, length, reader address, modality, idioms, lexical phrases, and common lexical items. Their findings revealed that the primary objectives of BTI texts were to promote the advertised destination and attract tourists. The analysis identified several notable linguistic features, such as the predominant use of active voice and simple tense. The 'Location' move was characterized by dependent clauses, a lack of modalities or idioms, and the use of distance terms and compass points. The 'Facilities/Activities' move employed imperatives, descriptive adjectives, pronouns, modal verbs, and numerous lexical phrase frames. Conversely, the 'Description' move featured existence verbs, pronouns, modal verbs, and pre-and post-modifier adjectives. While the study focused on the textual perspective of rhetorical structure and corpus analysis, it neglected other perspectives, such as ethnography and multimodality, which could have provided additional insights into the BTI genre.

Iborra and Garrido's (2001) analysis of twelve authentic travel leaflets written in English focused on identifying the moves and lexico-grammatical features within these texts. Using Swales' (1990) framework, they analysed the moves, while examining the lexico-grammatical features (nouns, adjectives, verbs, pronouns, and linking words) through frequency analysis and form-function correlations. The study found that all the leaflets shared only two obligatory moves, with the remaining moves being optional. Linguistic analysis of nouns revealed that 25.94% were compound nouns, while the majority (74.06%) were simple nouns. Over half of the compound nouns were proper names, such as names of museums, churches, streets, attractions, or local famous figures. The analysis also showed that adjectives played a more informative role than a persuasive one, while verbs were relatively less frequent compared to adjectives. Imperative language was primarily used to directly encourage readers to visit the attractions, while modal verbs were employed to indicate possibilities for visitors. Additionally, the active voice was used to convey enthusiasm and directness, making the text more understandable for readers.

Sinraksa's (2009) analysis of ten tourist leaflets published by the Tourism Authority of Thailand (TAT) aimed to identify the overall structure and communicative purposes of these texts. Utilizing the genre analysis theories of Swales (1990) and Bhatia (1993), the study identified seven moves within the leaflets. Five of these moves were prominent, while two were optional. Focusing on the 'Describing the Attraction' move, which was found to be

DOI: 10.48165/sajssh.2024.5602

significant in previous studies, Sinraksa analyzed the linguistic features based on Leech's (1966) advertisement language theory and the approaches of Iborra and Garrido (2001) and Boonchayaanant (2003), using Wordsmith Tool 5.0. The analysis revealed that modal verbs, such as 'can' and 'will,' were frequently used in the leaflets, primarily to describe possibilities and provide essential information to tourists. Pre-modifiers of adjectives were prominently employed with positive meanings to attract and entice readers to become tourists. Imperatives were used to persuade, inform, suggest, and encourage tourists to visit the promoted attractions. Finally, third-person pronouns were used as text references to locations, places, or people and as cohesive devices to represent preceding nouns.

Öztürk and Şiklar's (2014) analysis of a Turkish brochure advertising the tourist destination of Kemer utilized Bhatia's (2004) move structural model (Headline, Tour features, Highlights, Basic information, and Introduction) to investigate discourse patterns and features. The study aimed to uncover the communicative purposes of the brochure and the lexico-grammatical features that attracted customers. The analysis revealed that the use of Bhatia's moves, lexico-grammatical features, and visuals served both communicative and persuasive purposes. The study identified the use of present and past tenses within the brochure. Present tense was used to present the real world and state general facts, while past tense was employed to discuss past stories and events. The analysis also found that some essential information, such as telephone numbers, was missing. Additionally, numerous positive adjectives were used, with only two negative adjectives present. Modal verbs were primarily used to inform readers. All of these linguistic features were employed by the brochure's writers to create persuasive communication. Furthermore, the chosen images were related to the brochure's theme, serving the overall purpose of the writers.

Huang's (2015) genre-based analysis of 30 Brief Tourist Information (BTI) texts on tourism destination websites utilized Swales' (1990) model to identify the moves and steps within the corpus. Subsequently, he examined the linguistic forms to understand how each move was realized. The findings revealed that the 'establishing credentials' move played a crucial role in introducing and promoting the country, making it an obligatory move in BTI texts. The lexical linguistic features predominantly used in this move included adjectives, declarative sentences in simple present tense, and the imperative use of the personal pronoun 'you.'

DOI: 10.48165/sajssh.2024.5602

Cesiri's (2018) corpus analysis of travel guidebooks for Venice City utilized Wordsmith 6.0 and the framework of Kang and Yu (2011) to investigate the lexico-grammatical characteristics within a digital English-language corpus of travel guides. The purpose of the analysis was to understand how the distinctive aspects of Venice and its local culture were described to tourists and to examine the authors' strategies for balancing technical and promotional language. The findings revealed that the corpus featured complex sentences and a formal register. The tourism discourse reflected in the terminology and use of content words, particularly nouns and adjectives, showcased the complexity of tourism discourse and its interdisciplinary nature. The analysis also identified that digital travel guidebooks employed a specialized language distinct from the colloquial language used on tourism websites. A keyword analysis indicated that the most frequently occurring nouns referred to specific landscape features of the city, while adjectives highlighted the intrinsic qualities, dimensions, or other characteristics of the described elements. The authors' use of verbs suggested their intention to control tourists' actions during their stay in Venice by referring to typical experiences.

Hui, Santhi, and Mungthaisong's (2020) comparative analysis of natural and man-made tourism discourses in Lijiang, China, focused on move structures and linguistic features. Sixty-three promotional texts were selected from the top five tourism websites in the region, including 25 for natural attractions and 38 for man-made attractions. Using the move-step frameworks of Swales (1990) and Bhatia (2008), the researchers analysed the move structures and linguistic features of these texts. AntConc 3.5.8 was employed for the linguistic feature analysis. The results indicated that natural and man-made tourism discourses shared eleven moves, with one additional move exclusive to man-made texts. While the 'Headline' move was mandatory for both categories, the 'Detailing the Product' move was specific to man-made attractions. Similarities between the two discourses stemmed from their shared communicative purposes, which included informing, attracting, and persuading tourists. Modal verbs like 'can' and 'will' were commonly used in both categories, serving as symbols of promise, opportunity, and information to entice visitors. The analysis also revealed that moves 4, 5, and 9 received the most attention from adjectives in both categories. Both groups primarily used general descriptive adjectives to describe and rate the attractions' features. Superlative adjectives were utilized to highlight characteristics, emphasize value and significance, and make recommendations. Additionally, adjectives in both categories served a secondary purpose of implying meanings.

DOI: 10.48165/sajssh.2024.5602

Thu's (2021) analysis of a specialized corpus of tourism promotional texts focused on investigating the use of adjectives in English tourism writing to understand their contribution to persuasive text creation. The corpus, self-compiled from the official Vietnam tourism website, aimed to examine adjectival usage within a discourse known for its hyperbolic language. TermoStat Web 3.0 and Antconc were utilized to identify adjectives within the corpus. The findings revealed a high percentage of adjectives in the analyzed texts. The extensive and selective use of adjectives contributed to providing readers with a comprehensive picture of the described locations. Additionally, the frequent employment of compound adjectives enabled concise and detailed expressions.

The following Table 1 summarise the above studies that have been done on tourism texts using corpus analysis methods:

Table 1. *Summary on the previous corpus studies on tourism texts*

Study Title	Author(s) and Year	Data Analyzed	Tools/Methods Used	Concordance Tool
Genre Analysis of Brief Tourist Information (BTI)	Henry and Roseberry (1996)	44 samples of BTI genre in Standard English	Concordance program, Halliday and Hasan (1989), Swales (1990), Bhatia (1993)	Not specified
Moves and Lexico-Grammatical Features in Travel Leaflets	Iborra and Garrido (2001)	12 authentic travel leaflets	Swales (1990) framework, frequency analysis, form-function correlations	Not specified
Genre Analysis of Tourist Leaflets	Sinraksa (2009)	10 tourist leaflets by Tourism Authority of Thailand	Swales (1990), Bhatia (1993), Leech (1966)	Wordsmith Tool 5.0
Discourse Patterns in Turkish Tourist Brochures	Öztürk and Şiklar (2014)	1 Turkish brochure for tourist destination 'Kemer'	Bhatia's (2004) move structural model	Not specified
Genre-based Analysis of BTI Texts on Tourism Websites	Huang (2015)	30 BTI texts on tourism destination websites	Swales (1990) model, genre-based analysis	Not specified

DOI: 10.48165/sajssh.2024.5602

Corpus Analysis of Travel Guidebooks	Cesiri (2018)	English-language travel guides for Venice	Wordsmith 6.0, corpus analysis	Wordsmith 6.0
Comparative Analysis of Natural and Man-made Tourism Discourses	Hui, Santhi, and Mungthaisong (2020)	63 promotional texts for Lijiang tourist attractions	Swales (1990), Bhatia (2008), AntConc 3.5.8	AntConc 3.5.8
Use of Adjectives in English Tourism Writing	Thu (2021)	Specialized corpus from Vietnam tourism website	TermoStat Web 3.0, Antconc	TermoStat Web 3.0, Antconc

Table 1 provides an in-depth analysis of several studies focusing on genre analysis within tourism texts, highlighting the specific concordance tools employed in each study and their impact on the analysis of linguistic patterns and rhetorical strategies.

Wordsmith was used by Sinraksa (2009) to analyse 10 tourist leaflets from the Tourism Authority of Thailand, focusing on moves such as 'Describing the Attraction' and emphasizing the frequent use of modal verbs and pre-modifier adjectives to persuade readers to visit tourist attractions. Similarly, Wordsmith 6.0 was employed by Cesiri (2018) to study English-language travel guides for Venice. This study uncovered complex sentences, formal registers, and the strategic use of technical and promotional terms to describe Venice's cultural and historical landmarks.

AntConc 3.5.8 was utilized by Hui, Santhi, and Mungthaisong (2020) in analysing 63 promotional texts for Lijiang tourist attractions, emphasizing modal verbs and descriptive adjectives to entice tourists. Additionally, TermoStat Web 3.0 and Antconc were used by Thu (2021) to investigate adjectival usage in English tourism writing, focusing on the extensive and selective use of adjectives, including compound adjectives, to provide detailed descriptions and create persuasive texts promoting tourist destinations.

CONCLUSION

This study demonstrates the diverse methodologies and concordance tools used in analysing tourism texts. Wordsmith 6.0, AntConc, TermoStat Web 3.0, and Antconc have been essential in uncovering linguistic patterns, rhetorical strategies, and persuasive language techniques within tourism genres. Future research could explore further integrations of these tools with

DOI: 10.48165/sajssh.2024.5602

ethnographic and multimodal approaches to enhance our understanding of how tourism discourse shapes audience perceptions and decisions.

DOI: 10.48165/sajssh.2024.5602

REFERENCES

Altameemi, Y., & Altamimi, M. (2023). Thematic Analysis: A Corpus-Based Method for Understanding Themes/Topics of a Corpus through a Classification Process Using Long Short-Term Memory (LSTM). *Applied Sciences, 13*(5), 3308.

Bhatia, V. (1993). *Analysing Genre: Language Use in Professional Settings*: Routledge Taylor & Francis Group.

Bhatia, V. (2008). Genre analysis, ESP and professional practice. *English for Specific Purposes, 27*(2), 161-174. doi:10.1016/j.esp.2007.07.005

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the Move: Using corpus analysis to describe discourse structure*: John Benjamins Publishing Company.

Boonchayaanant, V. (2003). *A genre-based analysis of tourist leaflets produced and distributed in the United States of America*: Kasetsart University.

Cesiri, D. (2018). Balancing Tourism Promotion and Professional Discourse: A Corpus-based Analysis of Digital Travel Guidebooks Promoting Venice in English. *EPiC Series in Language and Linguistics, 2*, 247-237. doi:10.29007/2npc

Durán-Muñoz, I. (2019). Adjectives and their keyness: A corpus-based analysis of tourism discourse in English. *Corpora, 14*(3), 351-378.

Gandin, S. (2013). Translating the language of tourism. A corpus based study on the Translational Tourism English Corpus (T-TourEC). *Procedia-Social and Behavioral Sciences, 95*, 325-335.

Halliday, M. A. K., & Hasan, R. (1989). *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective*: Oxford University Press.

Henry, A., & Roseberry, R. L. (1996). A Corpus-Based Investigation of the Language and Linguistic Patterns of One Genre and the Implications for Language Teaching. *Research in the Teaching of English, 30*(4), 472-489. doi:10.2307/40171553

Huang, S. (2015, 2015). *A Genre-based Analysis of Brief Tourist Information Texts*.

Hui, W., Santhi, N., & Mungthaisong, S. (2020). A genre analysis of online English tourist attraction promotional texts in Lijiang Yunnan province. *Interdisciplinary social sciences and communication journal, 3*(1), 83-106.

Iborra, S. A., & Garrido, F. R. M. (2001). The Genre Of Tourist Leaflets. *Pasaa, 32*, 71-81.

Lee, C. (2016). A corpus-based investigation of theme choice in English translations of Korean online tourist texts–with focus on interactional themes. *Perspectives, 24*(2), 294-318.

Leech, G. (1966). *English in advertising: A linguistic study of advertising in Great Britain* (Vol. 2): London: Longmans.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics, 39*, 74-92.

Ortikov, U. (2023). Practical uses of corpus analysis in designing language teaching materials. *Oriental renaissance: Innovative, educational, natural and social sciences, 3*(7), 304-309.

DOI: 10.48165/sajssh.2024.5602

Öztürk, B., & Şiklar, E. (2014). Türkçe Bir Turizm Broşür Çözümlemesi Örneği : Kemer. *The Journal of Academic Social Science*(2004), 321-333.

Sinraksa, D. (2009). *A genre Based Approach To 'Describing the Attraction' Move in Tourist Leaflets of Tat.*

Sun, J. (2022). *A Corpus-Based Multi-dimensional Study of Tourism English Register Features.* Paper presented at the International Conference on Cognitive based Information Processing and Applications (CIPA 2021) Volume 2.

Swales, J. (1990). *Genre analysis: English in academic and research settings*: Cambridge university press.

Thu, D. Q. A. (2021). Adjectives in Destination Promotion Texts. *Indonesian Journal of EFL and Linguistics, 6*(1), 187-187. doi:10.21462/ijefl.v6i1.354

Vojnović, D. V. (2023). 'Experience Norfolk! Experience Fun!'vs.'Doživi više od očekivanog'–A Corpus-Based Contrastive Study of Reader Engagement Markers on the Web. *ELOPE: English Language Overseas Perspectives and Enquiries, 20*(1), 133-150.

DOI: 10.48165/sajssh.2024.5602