Applied Language Sciences

RESEARCH Open Access

A Corpus-Based Analysis of Tourism Discourse in the Digital Era

Kenryo Fu[1]*

*Correspondence:
fujianliang@gmail.com

[1] Professional College of Arts and Tourism, Japan

Check for updates

Abstract

Although destination image has long been studied, little is known about how it is linguistically constructed across genres of English-language tourism discourse. This comparative corpus-based study examines three small specialized corpora for Takarazuka City, Japan: an Official Tourism Website Corpus (OTWC; 8,908 tokens), a User-Generated Tourism Corpus of online reviews (UGTC; 6,054 tokens), and a Travel Journalism Corpus from *Time's World's Greatest Places 2025* (TJC; 25,130 tokens), the last serving as a destination-general benchmark. Using frequency- and dispersion-based word lists, semantic categorization informed by the cognitive-affective model, 4-gram clusters, and log-likelihood keyness with effect sizes in AntConc 4.2.0, the study shows that the OTWC projects an institutional image centered on cultural heritage and the Takarazuka Revue; the UGTC constructs a perceived image through first-person reference, evaluative adjectives, and practical concerns such as accommodation and access; and the TJC constructs an editorially mediated image through novelty markers and quoted voices introduced by *says*. The comparison reveals a mismatch between projected and perceived images, especially the greater salience of manga in reviews and performing arts on the official website. It demonstrates how small comparative corpora link genre-specific lexical choices to destination image and offers implications for English-language destination communication.

Keywords: corpus-based discourse analysis; specialized corpora; keyness; log-likelihood; destination image; tourism discourse; online reviews; travel journalism

GloDAL Global Digital Applied Linguistics Association

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

1. Introduction

Corpus linguistics has become an established methodology for investigating linguistic variation across registers, genres, and discourse types through the systematic analysis of authentic language data (Biber et al., 1998; McEnery & Hardie, 2012). Advances in digital communication have expanded the possibility of specialized domain-specific corpus compilation, and further enabled researchers to examine how language is used in various contexts with both quantitative and qualitative methods. Comparative corpus analysis, in particular, provides a robust methodological framework for systematically identifying discourse-specific lexical, semantic, and phraseological patterns across different corpora. Despite the growing application of corpus linguistic methods to domain-specific discourse, comparatively little attention has been paid to how tourism discourse varies across different digital genres and how such variation contributes to the linguistic construction of destination image. From a register perspective (Biber & Conrad, 2019), official tourism websites, online reviews, and travel journalism differ in key situational characteristics, including the relationship between writer and reader, the communicative purpose, and the degree of institutional control over text production. Such situational differences are expected to be reflected in systematic lexical differences, which makes tourism discourse a productive site for applied corpus research: the same destination is described by different social actors for different purposes.

With the rapid development of digital communication, online tourism platforms have become one of the main communicative channels between tourism destinations and their potential visitors. Among these online platforms, official tourism websites, which present carefully curated information about local attractions, culture, history, events, and tourism services, have attracted considerable scholarly attention over the past few decades (Malenkina & Ivanov, 2018; Pagano, 2023; Pagano & De Cantis, 2018). Through these strategically designed language and other multimodal resources, these official websites aim not only to provide practical travel information but also to establish and promote a distinctive and attractive image of a destination. Such official online communication has been shown to influence tourists' perceptions, expectations, and destination choices in various ways, thereby playing a significant role in destination image formation and travel decision-making (Li & Du, 2020; Pagano, 2023). However, while the role of official tourism websites has been widely acknowledged, limited research attention has been paid to how destination images are constructed across different forms of digital tourism discourse (Pagano, 2023, p. 17).

In addition to official tourism websites, user-generated content (UGC), particularly online reviews, has become an increasingly influential factor in destination image construction in the digital tourism environment. Unlike official websites, which present officially designed information by institutions, UGC reflects visitors' or tourists' authentic experiences, personal evaluations, and recommendations, thereby providing alternative perspectives on destination image. Previous studies have shown that potential tourists heavily rely on user-generated reviews or UGC when making travel and accommodation decisions (Fu, 2024; Ye et al., 2011). At the same time, tourism practitioners, including hotel managers and hoteliers, actively analyze large volumes of online review data to better understand visitors' experiences, preferences, and expectations, and to improve the quality of tourism services (Fu, 2024; Zainol et al., 2022). Consequently, a combined approach integrating the above official tourism websites and user-generated tourism discourse is required to provide a more comprehensive understanding of destination image construction, while comparison with travel journalism offers a third, editorially mediated perspective that is neither wholly institutional nor wholly personal.

Travel journalism or travel writing occupies a distinctive position from official tourism websites and UGC discourses. In contrast to official tourism websites and UGC, travel journalism is produced by professional editors and journalists and is expected to maintain a degree of journalistic objectivity while simultaneously promoting destinations (Brett, 2023; Lekant & Palau-Sampio, 2022). *Time* magazine's *World's Greatest Places* feature, which has been published annually since its launch in 2018, is such a representative example of professional travel writing in international travel journalism[i]. Although the *Time* magazine feature is not devoted to one single tourism destination, travel journalism discourse provides an important benchmark for examining how destinations are represented in travel media in general.

Using Takarazuka City, Hyogo Prefecture, Japan, as a case study, the present study employs comparative corpus analysis to examine linguistic variation across three types of English-language tourism discourse: official tourism websites, UGC, and travel journalism. Through a comparison of three specialized corpora, the study identifies distinctive lexical, semantic, and discourse patterns associated with each discourse type and explores how these patterns contribute to the construction of destination image. Methodologically, the study demonstrates the usefulness of multi-corpus comparison for investigating tourism discourse and offers corpus linguistic insights into destination image construction in the digital era. In doing so, the study aims to contribute to applied corpus linguistics in three ways: it offers a replicable procedure for comparing small, purpose-built specialized corpora (Koester, 2010); it combines frequency, dispersion, and keyness measures with explicit reporting of significance thresholds and effect sizes; and it links genre-specific lexical choices to an established tourism model of destination image, thereby connecting corpus-based discourse analysis with tourism research.

2. Literature Review

The term *discourse*, as used in expressions such as *tourism discourse* and *the discourse of socialism*, has multiple meanings in the discipline of linguistics. Structurally, discourse refers to "any piece of connected language, written or spoken, which contains more than one sentence" (Thornborrow & Wareing, 1998, p. 240). Functionally, however, discourse is understood as "language in use" (Brown & Yule, 1983, p. 1). Following the structural and functional perspectives in the previous studies, the present study defines *tourism discourse* as the language used in tourism-related communicative contexts and represented by the three domain-specific corpora compiled in this study.

This section then reviews the basic concept of destination image in the tourism context in the previous studies. Although definitions of destination image vary slightly in wording across studies and disciplines, researchers generally adopt Crompton's (1979, p. 18) widely cited definition of destination image as "the sum of beliefs, ideas, and impressions that a person has of a destination." Following this definition, destination image is commonly regarded as consisting of two interrelated components: the cognitive factors, referring to visitors' knowledge and beliefs about a destination, and the affective factors, referring to their feelings and emotional responses toward a destination (Baloglu & McCleary, 1999). Baloglu and McCleary (1999) further argue that destination image is mainly shaped by two major forces: stimulus factors, i.e., what tourists know about a destination, such as information sources; and personal factors, i.e., how tourists feel about a destination, including psychological and social characteristics. Baloglu and McCleary's (1999) framework provides an important theoretical basis for examining how different forms of tourism discourse contribute to destination image formation. This perspective

provides the theoretical motivation for comparing official tourism websites, user-generated tourism discourse, and travel journalism discourse in the present study.

Destination image has long been recognized as one of the central concepts in tourism research because of its dual role in benefiting both tourism destinations and tourists. As Li and Du (2020, p. 13) have pointed out, destination image benefits both the supply side (tourism destinations) and the demand side (tourists) in the tourism industry. For tourism destinations, a favorable destination image enhances their competitiveness against their rivals, whereas for tourists, it increases travel expectations, and enriches the overall travel experience. At the same time, the ways in which destination images are constructed, communicated, and transformed have evolved considerably with the emergence of diverse forms of digital tourism discourse. In this regard, Gartner (1994) has argued that destination images are formed through multiple information sources, including induced, organic, and autonomous agents. However, relatively little emphasis has been placed on the digital and multimodal introduction of destinations on dynamic digital platforms that are continuously updated.

Among the various information resources that contribute to tourism destination image formation, linguistic resources are of particular importance. As Dann (1996, p. 2) has observed, "the language of tourism attempts to persuade, lure, woo, and seduce millions of human beings and, in so doing, convert them from potential into actual clients." Dann further argues that linguistic features, including strategically selected keywords, can "fire the imagination" of the potential tourists and thereby serve a powerful persuasive function (1996, p. 174). This perspective highlights the importance of examining lexical choices and recurring linguistic patterns in order to understand how destination images are constructed and promoted through tourism discourse.

Among various linguistic elements that exert influence on destination image formation, the prominence of first-person pronouns in travel journalism has particularly been observed by Brett (2023). Brett's travel subcorpus of the Sassari Newspaper Article Corpus (SNAC) was compiled from articles selected from the travel section of the online version of the British newspaper *The Guardian*, covering a wide range of travel topics, including weekend breaks, city breaks, skiing, golf, cruises, gap years, and green travel (Brett, 2023, p. 4)[ii]. By comparing the travel subcorpus with other subcorpora of SNAC, including Crime, Football, and Banking, Brett found that the first-person pronouns occurred far more frequently in their travel subcorpus—the equivalent of travel journalism in the present study—than those in the other subcorpora. He argues that this distinctive use of first-person reference reflects the experiential and personal nature of travel journalism, as detailed in (1).

(1) It is interesting to note how the most common form of reference in Travel is first person, whereas in Tourism it is second person. Such findings would tend to confirm a more literary nature of Travel Journalism, where the author makes frequent reference to him- or herself throughout the journey, and the texts take on some of the characteristics of the diary. (Brett, 2023, p. 7)[iii]

The travel journalism examined by Brett (2023) is characterized by its experiential and personal nature, as professional newspaper article writers or journalists narrate their own travel experiences predominantly from a first-person perspective. By contrast, the Travel Journalism Corpus (TJC) compiled and utilized in the present study consists of edited features that construct destination image through multiple voices. Although the two datasets have the same name of travel journalism, they represent different styles and purposes. From a discourse perspective, Brett's travel journalism shares important linguistic characteristics with another genre, i.e., user-generated tourism discourse, particularly its reliance on first-person narration and personal experience.

Corpus-based research on the language of tourism has itself developed considerably. Tourism discourse has been described as a specialized discourse with its own lexical and textual conventions (Dann, 1996; Gotti, 2006), and corpus studies have documented, for instance, the heavy reliance of promotional tourism texts on evaluative adjectives (Pierini, 2009) and the persuasive strategies of destination promotion (Manca, 2016). In parallel, research on online consumer reviews has shown that reviewers construct credibility and evaluation through personal narrative, first-person reference, and explicit evaluative lexis (Vásquez, 2014), a finding closely related to the wider study of evaluation in text (Hunston & Thompson, 2000). Methodologically, keyword analysis has become a central technique for identifying what is distinctive about one corpus relative to another (Scott & Tribble, 2006), although its results depend on the choice of reference corpus and of the keyness statistic (Baker, 2004; Gabrielatos, 2018). These strands provide the methodological background for the multi-corpus design adopted here.

Previous studies have considerably advanced our understanding of destination image by examining official tourism websites, user-generated tourism discourse, or travel journalism discourse from different perspectives. However, these three forms of tourism discourse have rarely been investigated within a single analytical framework. In particular, little attention has been paid to how they differ in their lexical and semantic characteristics and how these differences contribute to the construction of destination image. To address these gaps, the present study adopts a corpus-based case study approach to examine the destination image of Takarazuka City across three different discourses. More specifically, the study addresses the following three research questions (RQs).

(2) RQ1 What lexical and semantic features characterize the three specialized corpora of official tourism website texts (OTWC), user-generated reviews (UGTC), and travel journalism (TJC)?

(3) RQ2 How does the user-generated tourism corpus differ from the official tourism website corpus in the lexical construction of the destination image of Takarazuka City?

(4) RQ3 What does a comparison with a destination-general travel journalism corpus reveal about the genre-specific features of the two Takarazuka corpora, and what are the implications for Takarazuka's English-language tourism communication?

3. Methodology and Corpora

The paper presents a case study of Takarazuka City's destination image through the official tourism website of the Takarazuka International Tourism Association [iv] (TITA), Takarazuka-related UGC at Tripadvisor[v] and Trip[vi], and travel journalism discourse of *Time* magazine[vii]. The selection of Takarazuka City as the case study was based on the following three considerations. First, despite its rich cultural and tourism resources, including the internationally known Takarazuka Revue and the legacy of the manga artist Osamu Tezuka, often called the "God of Manga," the city has received relatively little attention in corpus-based and destination image studies. Second, Takarazuka City offers diverse tourism attractions, such as performing arts, hot springs, historical sites, and manga-related culture, which make it an appropriate case for examining how different tourism discourses represent a multifaceted destination image. Finally, reasonably scaled linguistic data are available from the official tourism website and UGC, which enables a combined approach of quantitative and qualitative analyses and motivates a systematic corpus-based comparison across different discourse types.

Three specialized domain-specific corpora were compiled by the author based on three different resources: the official tourism website of the TITA, Takarazuka-related UGC at Tripadvisor and Trip, and *Time* magazine's online *World's Greatest Places* 2025 feature.

These corpora represent three distinct forms of tourism discourse: official tourism communication, user-generated tourism discourse, and travel journalism. Together, they provide the basis for a corpus-based comparison of how the destination image of Takarazuka is constructed across different communicative contexts.

To ensure comparability, all three corpora were compiled using the same corpus-building procedures and subsequently processed using AntConc (version 4.2.0; Anthony, 2022)[viii] for lexical and semantic analyses. Based on Stefanowitsch (2006) and revised by the author, the following strategies were used for extracting examples and data. All texts were collected manually and saved as plain-text (UTF-8) files, one file per web page (OTWC), per review (UGTC), or per article (TJC). Reviews written in languages other than English and machine-translated reviews were excluded. AntConc's default token definition was used; all items were case-folded, and no lemmatization was applied, so that, for example, *museum* and *museums* were counted as separate types. The only exception is the manual merging of *theater* and *theatre* described in the note to Table 5. Given their size, the three corpora are best described as small specialized corpora (Koester, 2010), which are well suited to capturing the typical features of a narrowly defined discourse but call for caution when generalizing beyond it.

(5) Strategies for extracting examples and data using AntConc (Version 4.2.0)

a. Manual searching

b. Searching for target vocabulary using the *KWIC* tab

c. Extracting target words using the *Word* tab (Min. Freq = 1, Min. Range = 1)

d. Extracting 4-gram clusters using the *N-Gram* tab (N-Gram Size = 4, Open Slots = 0, Min. Freq = 1, Min. Range = 1)

e. Extracting target words using the *Keyword* tab (Min. Freq = 1, Min. Range = 1; keyness statistic: Log-likelihood; effect-size measure: Dice; significance threshold: $p < .0001$ ($LL \geq 15.13$))

The descriptive statistics of the three corpora compiled by the author are elaborated, as shown in Table 1. First, the Official Tourism Website Corpus (OTWC) consists of 67 English-language texts collected from the official website of the TITA. These files cover the "our top 12 picks" and "Suggestions" sections of the official website. Second, the User-Generated Tourism Corpus (UGTC) comprises 58 English-language user reviews of "The Best Hotels in Takarazuka 2025," and "Top Attractions in Takarazuka" collected from Tripadvisor, and 18 English-language user reviews of tourism facilities and attractions in the city collected from the Trip website. Restricting the UGTC to 2025 would result in a dataset too small to support meaningful corpus-based analysis. For this reason, reviews published between 2011 and 2025 were included to obtain a sufficiently substantial body of user-generated discourse concerning Takarazuka[ix]. Third, the Travel Journalism Corpus (TJC) consists of 100 online articles featured in *Time* magazine's *World's Greatest Places* 2025. Unlike the OTWC and the UGTC, which are both Takarazuka-oriented, the TJC covers a broader range of destinations and is intended to serve as a destination-general benchmark against which the features of the two Takarazuka corpora can be evaluated. Table 1 summarizes the descriptive statistics of the three corpora, including their data collection periods, numbers of texts (files or reviews), tokens, types, Type-Token Ratio (TTR), and Standardized Type-Token Ratio (STTR). These statistics provide an overview of the size and composition of each corpus and serve as the basis for the subsequent lexical analyses.

Table 1. Descriptive Statistics of the OTWC, the UGTC, and the TJC

	OTWC	UGTC	TJC
Time Period	2025	2011-2025	2025
Files / Reviews	67	76	100
Tokens	8,908	6,054	25,130
Types	2,272	1,386	6,414
TTR (%) (Type-Token Ratio)	25.51	22.89	25.52
STTR[x] (500 tokens, %)	53.32	51.55	65.34

Note: Tabulated and compiled by the author from the OTWC, the UGTC, and the TJC using AntConc[xi]

4. Results and Discussion

4.1 *The Official Tourism Website Corpus (OTWC)*

To identify the prominent lexical features underlying the officially projected destination image of Takarazuka City, OTWC was analyzed using AntConc. The word list was finalized for analysis according to the following three procedures and screening criteria. First, a raw word list was generated from the OTWC using AntConc, with the parameters Min. Frequency = 1 and Min. Range = 1. Second, the 50 most frequent types were extracted from the raw word list because 50 words were deemed enough to provide a representative lexical overview of the corpus while maintaining a manageable dataset for qualitative semantic analysis. Third, function words (stop words) and semantically neutral or non-topical items were removed from the list so that only high-frequency content words[xii], namely, nouns, verbs, adjectives, and adverbs based on Biber et al. (1999, p. 62), were retained for subsequent semantic analysis, as presented in Table 2. Because a cut-off at the 50 most frequent types yields different numbers of content words in each corpus (13 in the OTWC, 15 in the UGTC, and 12 in the TJC), the finalized lists are used for qualitative profiling rather than for direct numerical comparison.

Table 2. The Finalized Word List from the OTWC

No.	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
1	*takarazuka*	7	144	42	16,165	0.627
2	*temple*	21	47	9	5,276	0.134
3	*revue*	26	37	14	4,154	0.209
4	*enjoy*	28	34	20	3,817	0.299
5	*hotel*	30	32	10	3,592	0.149
6	*shrine*	30	32	7	3,592	0.104

No.	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
7	*theater*	30	32	9	3,592	0.134
8	*area*	38	26	12	2,919	0.179
9	*japanese*	41	23	15	2,582	0.224
10	*grand*	43	22	7	2,470	0.104
11	*hall*	43	22	9	2,470	0.134
12	*time*	46	20	13	2,245	0.194
13	*museum*	50	18	8	2,021	0.119

Note: Tabulated and compiled by the author from the OTWC using AntConc

Table 2 presents the final set of 13 high-frequency content words from the OTWC, together with their distributional statistics, including raw frequency, raw range, normalized frequency, and normalized range. Normalized frequency, as noted in McEnery and Hardie (2012, pp. 49-50), refers to the number of occurrences of each lexical item per million words (pmw), as calculated using Equation (6)[xiii]. Normalized range represents the proportion of texts in the corpus in which a lexical item occurs, and is calculated by dividing its raw range by the total number of texts, as shown in Equation (7)[xiv]. Normalized frequency and normalized range facilitate comparative analyses across corpora of different sizes by indicating both the relative frequency and the textual distribution of each lexical item. These values therefore provide a more comprehensive account of the lexical features of the OTWC than raw frequency or raw range alone. Two caveats apply. First, given the small size of the corpora, frequencies per million words involve considerable extrapolation; they are reported here for comparability with previous studies, but frequencies per 10,000 words would express the same proportions on a more realistic scale (e.g., 16,165 pmw = 161.65 per 10,000 words). Second, normalized range is a simple measure of dispersion; more refined measures such as Gries's (2008) DP could be considered in future work.

$$\text{NormFreq} = \frac{Raw\ frequency}{Total\ number\ of\ tokens} \times 1{,}000{,}000 \quad (6)$$

$$\text{NormRange} = \frac{Raw\ range}{Total\ number\ of\ texts} \quad (7)$$

The 13 high-frequency lexical items presented in Table 2 were further classified into seven semantic categories by the author according to their functions in the tourism context. The classification was carried out manually on the basis of the concordance lines for each item, so that items were assigned according to their predominant use in context rather than their dictionary meaning. As shown in Table 3-1, these categories include destination identity, nation identity, cultural heritage, performing arts, tourism infrastructure, and general contextual terms (six cognitive factors), and promotional or evaluative language (one affective factor). These semantic categories reveal that the official tourism website constructs the destination image of Takarazuka primarily through its distinctive cultural

heritage (displayed by three lexical items, *temple, shrine,* and *museum*) and performing arts resources (displayed by four lexical items, *revue, theater, grand,* and *hall*). In particular, lexical items such as *temple*, *shrine*, and *museum* highlight the city's historical and cultural attractions, whereas *revue*, *theater*, *grand*, and *hall* collectively emphasize the symbolic importance of the Takarazuka Revue as a unique tourism asset[xv]. These semantic categories mainly contribute to the cognitive dimension of destination image (Baloglu & McCleary, 1999), while the frequent use of promotional expressions such as *enjoy* reflects the affective aspect of destination image (Baloglu & McCleary, 1999) and the persuasive function of tourism discourse (Dann, 1996). These findings answer RQ1 by demonstrating that official tourism discourse is characterized by destination-specific lexical choices that reflect cultural heritage and the performing arts in constructing the destination image of Takarazuka.

Table 3-1. Classification of the Finalized Word List of the OTWC

No.	Factor Type	Attribute	Lexical Items
1	Cognitive	Destination Identity	*takarazuka*
2	Cognitive	Nation Identity	*japanese*
3	Cognitive	Cultural Heritage	*temple, shrine, museum*
4	Cognitive	Performing Arts	*revue, theater, grand, hall*
5	Cognitive	Tourism Infrastructure	*hotel*
6	Cognitive	General Contextual Terms	*area, time*
7	Affective	Promotional / Evaluative Language	*enjoy*

Note: Tabulated and compiled by the author from the OTWC using AntConc

Further, a 4-gram cluster analysis was conducted on the OTWC. A 4-gram cluster refers to a recurrent sequence of four consecutive words and four-word sequences are the length most frequently investigated in research on lexical bundles (Biber et al., 2004; Partington & Morley, 2004). The results reveal that eight of the ten most frequent 4-gram clusters in the OTWC are related to the performing arts, which is consistent with the lexical patterns presented in Tables 2 and 3-1. The most frequent 4-gram cluster is *the Takarazuka Grand Theater* (Rank = 1, Frequency = 16, Range = 5). The remaining two clusters, *is said to have* (Rank = 6, Frequency = 5, Range = 2) and *dinner is served in* (Rank = 10, Frequency = 4, Range = 2), fall into the categories of general contextual terms and tourism infrastructure, respectively. These findings further demonstrate that recurrent lexical bundles in the OTWC reinforce the prominence of performing arts in the linguistic construction of Takarazuka's destination image. Because several of these clusters occur in only two to five texts, they are best interpreted as recurrent phrasing on particular pages rather than as corpus-wide lexical bundles, which are usually required to be dispersed across at least five texts (Biber et al., 2004).

4.2 *The User-Generated Tourism Corpus (UGTC)*

4.2.1 The Finalized Word List of the UGTC

Following the same analytical procedure, the UGTC was analyzed to identify the lexical features that characterize visitors' perceived destination image of Takarazuka City. The finalized word list of the UGTC is presented in Table 4. The 15 high-frequency content words from the UGTC were subsequently divided into the seven semantic categories (Table 3-2). An overall comparison of the OTWC and the UGTC reveals three major differences in three semantic categories: cultural heritage, performing arts (both cognitive factors), and promotional / evaluative language (an affective factor), which are further elaborated in the subsequent sections.

Table 4. The Finalized Word List of the UGTC

No.	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
1	*hotel*	16	45	21	7,433	0.276
2	*very*	16	45	27	7,433	0.355
3	*nice*	27	29	17	4,790	0.224
4	*takarazuka*	27	29	20	4,790	0.263
5	*good*	31	27	18	4,460	0.237
6	*great*	34	26	20	4,295	0.263
7	*room*	35	25	17	4,130	0.224
8	*place*	36	24	19	3,964	0.25
9	*station*	37	23	19	3,799	0.25
10	*beautiful*	46	19	15	3,138	0.197
11	*japanese*	46	19	17	3,138	0.224
12	*manga*	49	18	8	2,973	0.105
13	*museum*	49	18	9	2,973	0.118
14	*temple*	49	18	6	2,973	0.079
15	*walk*	49	18	13	2,973	0.171

Note: Tabulated and compiled by the author from the UGTC using AntConc

One striking lexical difference between the OTWC and the UGTC concerns the lexical item *manga*. *Manga*, an important tourism resource in Takarazuka through the legacy of Osamu Tezuka and the Takarazuka Osamu Tezuka Manga Museum, occurs 18 times in eight texts (NormFreq = 2,973 pmw[xvi]; NormRange = 0.105[xvii]) in the UGTC but only eight times in three texts (NormFreq = 898 pmw; NormRange = 0.045) in the OTWC. Osamu Tezuka is widely recognized as the "God of Manga" (Encyclopedia Britannica[xviii]). A further analysis identified two examples (8) and (9) referring to Osamu Tezuka as the "*God of Manga*" in the UGTC, whereas no corresponding examples were found in the OTWC. Although the absolute frequencies are small, the difference is statistically significant (LL = 8.77, $p < .01$), with *manga* occurring approximately 3.3 times as frequently in the UGTC as in the OTWC. This discrepancy suggests that visitors perceive manga as a more salient destination attribute than is reflected in the official tourism website, which indicates a gap between the officially projected destination image and the perceived

destination image by tourists. More broadly, the finding implies that manga culture, despite its strong association with Takarazuka City, remains underrepresented in the city's official English-language tourism communication.

Table 3-2. Classification of the Finalized Word Lists of the OTWC and the UGTC

No.	Factor Type	Attribute	OTWC	UGTC
1	Cognitive	Destination identity	*takarazuka*	*takarazuka*
2	Cognitive	Nation identity	*japanese*	*japanese*
3	Cognitive	Cultural Heritage	*temple, shrine, museum*	*manga, museum, temple*
4	Cognitive	Performing Arts	*revue, theater, grand, hall*	—[xix]
5	Cognitive	Tourism Infrastructure	*hotel*	*hotel, room, station*
6	Cognitive	General Contextual Terms	*area, time*	*place, walk*
7	Affective	Promotional / Evaluative Language	*enjoy*	*very, nice, good, great beautiful*

Note: Tabulated and compiled by the author from the OTWC and UGTC using AntConc

(8) The entire museum is dedicated to Osamu Tezuka, "*the god of manga*" and they've added a lot of decorations and architecture for his various works. (UGTC, 2014, italicized by the author)

(9) Our tour group stopped for a short restroom break here on the way to Awaji Island, so was not able to fully explore what was available at this rest area. Overall, seemed very clean and ultra-modern. Noticed a lot of Harry Potter promotional merchandise (a tie-in with Universal Studios Osaka?) and the characters of the famed "*God of Manga*," Osamu Tezuka. Not sure what to think about the Dr. Blackjack toilet paper that was on sale. (UGTC, 2024)

A second lexical difference is observed in the category of performing arts between the OTWC and the UGTC. While lexical items related to the Takarazuka Revue (*revue, theater, grand,* and *hall*) constitute one of the most prominent semantic categories in destination image formation in the OTWC, no corresponding high-frequency lexical items appear in the UGTC. As shown in Table 5, the lexical evidence suggests that the performing arts, although strongly emphasized in the official tourism website, are considerably less salient in visitors' online reviews. For example, the normalized frequency of *grand* in the OTWC (NormFreq = 2,470) is 7.48 times that of *grand* in the UGTC (NormFreq = 330); the normalized range of *grand* in the OTWC (NormRange = 0.104) is 4 times as high as that in the UGTC (NormRange = 0.026). Log-likelihood tests confirm that all four items are significantly less frequent in the UGTC than in the OTWC: *revue* ($LL = 9.18$, $p < .01$), *theater* ($LL = 7.63$, $p < .01$), *grand* ($LL = 12.67$, $p < .001$), and *hall* ($LL = 22.82$, $p < .0001$). These findings indicate a discrepancy between the officially projected destination image and the perceived destination image with respect to one of Takarazuka's most distinctive tourism resources.

Table 5. Comparison of Performing Arts-Related Lexical Items in the OTWC and the UGTC

Corpus	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
OTWC	*revue*	26	37	14	4,154	0.209
	theater	30	32	9	3,592	0.134
	grand	43	22	7	2,470	0.104
	hall	43	22	9	2,470	0.134
UGTC	*revue*	110	9	7	1,487	0.092
	theater	127	8	8	1,321	0.105
	grand	387	2	2	330	0.026
	hall	—	—	—	—	—

Note: Tabulated and compiled by the author from the OTWC and the UGTC using AntConc. The UGTC contains one occurrence of the American spelling "theater" and seven occurrences of the British spelling "theatre." To ensure consistency in the lexical analysis, these spelling variants were manually combined by the author, and their frequencies and distributional measures (Rank, Raw Frequency, Raw Range, Normalized Frequency, and Normalized Range) were calculated as a single lexical item. By contrast, no instances of the British spelling *theatre* were found in the OTWC.

One possible explanation for the gap mentioned above is the limited accessibility of theater attendance. Although the Takarazuka Revue is one of the city's most distinctive tourism resources, references to the performing arts occur far less frequently in the UGTC than in the OTWC. Unlike museums, temples, or other attractions that visitors can access almost throughout the day, attending a theatrical performance is often constrained by fixed performance schedules, ticket availability, and advance reservations, as noted by Nozu et al. (2023). As a consequence, many visitors may experience other more accessible attractions without necessarily attending a Takarazuka Revue performance, resulting in the relatively limited representation of performing arts in user-generated tourism discourse. This explanation remains tentative, as it cannot be tested with the present corpus data alone.

A third lexical difference concerns promotional and evaluative language between the OTWC and the UGTC. In contrast to the OTWC, which primarily provides factual and descriptive information, the UGTC contains a considerably larger number of evaluative lexical items such as *nice*, *good*, *great*, and *beautiful*. These expressions reflect visitors' personal evaluations and emotional responses, corresponding to the affective aspect of destination image (Baloglu & McCleary, 1999). Overall, the comparison demonstrates that official tourism discourse emphasizes destination promotion and branding, whereas user-generated tourism discourse foregrounds visitors' authentic experiences and evaluations.

The differences between the OTWC and the UGTC are further demonstrated through a comparison executed in the larger unit of 4-gram clusters. Unlike the OTWC, none of the ten most frequent 4-gram clusters in the UGTC are related to the performing arts. Overall, six of the ten most frequent clusters fall into the tourism infrastructure category, whereas the remaining four belong to the general contextual terms category. The most frequent 4-gram cluster is *from the train station* (Rank = 1, Frequency = 5, Range = 5), which belongs

to the tourism infrastructure category and represents a cognitive factor. These findings suggest that visitors to Takarazuka focus more on practical aspects of travel, particularly tourism infrastructure such as hotels and transportation, than on destination-specific cultural resources.

4.2.2 Keyness Analysis in the UGTC

Keyness analysis was conducted using AntConc to identify lexical items that are significantly overrepresented in the UGTC (target corpus) relative to the OTWC (reference corpus). As noted by Partington et al. (2013, p. 18), a word list and a keyword list serve different analytical purposes. A word list (as presented in Table 4) provides information about the frequency of individual lexical items within the target corpus, identifying words that occur relatively frequently in that corpus. In contrast, a keyword list (as presented in Table 6) identifies lexical items that occur with statistically significantly higher frequencies in the target corpus than would be expected on the basis of their frequencies in a reference corpus. The finalized keyword list of the UGTC (Target Corpus: the UGTC; Reference Corpus: the OTWC) is presented in Table 6. Using the log-likelihood (LL) statistic (Dunning, 1993; Rayson & Garside, 2000), all eight keywords yielded the LL values exceeding the threshold of 15.13 ($p < .0001$), indicating highly significant lexical differences between the two corpora. To complement the statistical significance indicated by LL, the keyness (effect) values reported by AntConc, calculated using the Dice measure selected in AntConc's keyword settings, are also presented as indicators of the practical magnitude of lexical uniqueness. As effect size is a descriptive measure rather than a statistical significance test, it is interpreted as supplementary evidence in conjunction with the LL measures. Because keyness is calculated for every type in the corpus, many comparisons are performed; the stringent threshold of $p < .0001$ was therefore adopted partly to reduce the risk of false positives arising from multiple testing (Gabrielatos, 2018). Only positive keywords, i.e., items overrepresented in the target corpus, are reported in Table 6.

Table 6. Keyness Analysis (Target Corpus: the UGTC; Reference Corpus: the OTWC)

No.	Type	Keyness (Likelihood)	Keyness (Effect)
1	*i*	161.84	0.029
2	*my*	36.23	0.007
3	*nice*	44.83	0.01
4	*great*	24.89	0.009
5	*good*	23.84	0.009
6	*stay*	20.72	0.005
7	*beautiful*	17.30	0.006
8	*room*	17.02	0.008

Note: 3.84, $p < .05$; 6.63, $p < .01$; 10.83, $p < .001$; 15.13, $p < .0001$[xx]

First, the keyness analysis demonstrates the overrepresentation of first-person pronouns (*I* and *my*) in the UGTC relative to the OTWC. As shown in Table 6, among all keywords, the lexical item *I* exhibits the highest LL value (LL = 161.84, $p < .0001$) as well

as the largest effect size (effect size = 0.029). Unlike the OTWC, which adopts an institutional and relatively impersonal style in promoting the destination, the UGTC constructs destination image primarily through visitors' first-hand experiences and personal perspectives. This finding is broadly consistent with Brett (2023, pp. 6–7), who reports that first-person pronouns are among the most prominent forms of personal reference in the travel subcorpus of the SNAC[xxi], compiled in his study. In the present study, the frequent use of first-person pronouns enhances the authenticity and experiential nature of user-generated tourism discourse, thereby contributing to the construction of a more personal and experience-based destination image. This interpretation is consistent with research on online consumer reviews, in which first-person narrative is a central resource for constructing credibility (Vásquez, 2014).

The keyness analysis further shows that the UGTC is characterized by lexical items associated with subjective evaluation (*nice*, *great, good,* and *beautiful*) and personal experience (*stay*), i.e., affective factors, and tourism infrastructure-related terms (*room*), i.e., cognitive factors. These lexical items reflect both the affective and cognitive components of destination image (Baloglu & McCleary, 1999), which indicates that user-generated tourism discourse constructs destination image primarily through personal experience, subjective evaluation, and practical travel concerns. By contrast, the official tourism website constructs destination image mainly by presenting tourism resources and destination attributes from a promotional but impersonal perspective.

4.3 *The Travel Journalism Corpus (TJC)*

4.3.1 The Finalized Word List of the TJC

Following the same analytical procedure, the TJC was analyzed to identify the salient lexical features underlying the editorially mediated destination image presented in international travel journalism. Unlike the OTWC and the UGTC, which are destination-specific corpora focusing on the specific destination of Takarazuka City, the TJC, which is destination-general in nature, comprises 100 texts describing 49 *places to visit* and 51 *places to stay* featured in *Time* magazine's *World's Greatest Places 2025*. The TJC serves as an external benchmark for the OTWC and UGTC, both of which focus on Takarazuka. As a destination-general corpus covering multiple tourist destinations, the TJC provides a broader, cross-destination perspective on tourism discourse, enabling comparisons of lexical diversity and linguistic characteristics across the three corpora in this study. In general, the finalized word list demonstrates that the high-frequency content words (Table 7) from the TJC are more diverse and are distributed unevenly across the seven categories shown in Table 3-3.

Table 7. The Finalized Word List of the TJC

No.	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
1	*new*	20	94	55	3,741	0.55
2	*hotel*	26	65	30	2,587	0.3
3	*art*	31	55	30	2,189	0.3
4	*world*	32	54	38	2,149	0.38

No.	Type	Rank	Raw Frequency	Raw Range	Normalized Frequency (pmw)	Normalized Range
5	*first*	34	52	40	2,069	0.4
6	*guests*	40	49	36	1,950	0.36
7	*local*	42	48	34	1,910	0.34
8	*museum*	43	46	20	1,830	0.2
9	*says*	43	46	32	1,830	0.32
10	*opened*	45	45	35	1,791	0.35
11	*resort*	48	43	19	1,711	0.19
12	*country*	49	40	29	1,592	0.29

Note: Tabulated and compiled by the author from the TJC using AntConc

One notable characteristic of the TJC is the balanced use of promotional and evaluative language, representing an intermediate pattern between the OTWC and the UGTC. Unlike the OTWC, which mainly employs promotional language to market the destination, or the UGTC, which predominantly contains visitors' personal evaluations, the TJC combines descriptive, evaluative, and journalistic language to present destination image in an informative yet engaging manner. Specifically, the distinctive feature of the TJC is the occurrence of lexical items such as *new, first, local,* and *opened,* none of which appear among the high-frequency lexical items in either the OTWC or the UGTC. Strictly speaking, these items are not evaluative in the same sense as *nice* or *beautiful*. They function rather as markers of newsworthiness, foregrounding novelty and locality, and carry an implicit positive appraisal; it is in this sense that they are grouped with promotional and evaluative language in Table 3-3. These words foreground novelty, authenticity, and newsworthiness, which are characteristic features of editorial tourism discourse. As illustrated in (10) – (13), *new* is used to introduce a recently developed resort in (10), *first* highlights a pioneering dome in (11), *local* emphasizes authentic cultural experiences with the local wildlife conservationist in (12), and *opened* reports a newly established museum in (13). Together, these lexical choices distinguish travel journalism from both official tourism and user-generated tourism discourses. They also suggest that incorporating greater emphasis on novelty and authenticity could further enhance the English-language tourism communication of Takarazuka.

Table 3-3. Classification of the Finalized Word Lists of the Three Corpora

No.	Factor Type	Attribute	OTWC	UGTC	TJC
1	Cognitive	Destination identity	*takarazuka*	*takarazuka*	*world, country*

2	Cognitive	Nation identity	*japanese*	*japanese*	—
3	Cognitive	Cultural Heritage	*temple, shrine, museum*	*manga, museum, temple*	*museum*
4	Cognitive	Performing Arts	*revue, theater, grand, hall*	—	*art*
5	Cognitive	Tourism Infrastructure	*hotel*	*hotel, room, station*	*hotel, resort,*
6	Cognitive	General Contextual Terms	*area, time*	*place, walk*	*guests, says*
7	Affective	Promotional / Evaluative Language	*enjoy*	*very, nice, good, great beautiful*	*new, first, local, opened*

Note: Tabulated and compiled by the author from the OTWC, the UGTC, and the TJC using AntConc

(10) "Since Madagascar is [largely] untouched, it's a *new* destination for a lot of people," says Philippe Kjellgren, founder and owner of the *new* luxury resort, Voaara Madagascar, on Sainte-Marie's northwestern coast, where you can spot humpback whales migrating from June to September. (Voaara)

(11) Aiming to be the NBA's *first* carbon-neutral arena includes solar panels and water preservation efforts like low-flow faucets, drought-tolerant landscaping, and water-efficient irrigation. (Intuit Dome)

(12) Other tour highlights include night safaris, a tiger-painting workshop, and drinks with a *local* wildlife conservationist. (Oberoi Vindhyavilas Wildlife Resort, Bandhavgarh)

(13) The two-story museum *opened* in October in Uji, about 30 minutes south of Kyoto, on the site of the company's former Uji Ogura plant, where the Hanafuda trading cards that put the company on Japan's map were once printed. (Nintendo Museum)

4.3.2 Keyness Analysis in the TJC

Keyword analysis was also conducted by comparing the TJC (target corpus) with the OTWC (reference corpus). Because the TJC comprises travel journalism covering 100 destinations worldwide rather than a single destination, its lexical profile is more diverse and less destination-specific. Consequently, the comparison between the TJC and OTWC yielded far fewer statistically significant keywords (Table 8) than the comparison between the UGTC and OTWC (Table 6). This result is attributable to the broader topical coverage and greater lexical diversity of the TJC, which is not restricted to a single destination. It should be noted, however, that the TJC differs from the OTWC in both genre and topical scope, so the keywords obtained from this comparison cannot be attributed to genre alone; the result is therefore treated as exploratory.

Specifically, one notable keyword in the TJC is *says* (LL = 27.936, $p < .0001$; effect = 0.004), a reporting verb that introduces direct speech and contributes to the authenticity of travel journalism. In contrast to Brett (2023), who argued that direct speech is "relatively rare in the Travel" corpus and "is practically non-existent" in the Language of Tourism corpus (p. 6), a concordance analysis of the TJC in this study revealed that all 46 occurrences of *says* in 32 texts (see Table 7) introduce direct quotations from local residents, business owners, experts, or other stakeholders, as presented in (14), (15), and (16). The frequent use of direct speech contributes to authenticity and immediacy by presenting destination information through multiple voices rather than through the journalist's narration alone.

(14) “It’s been a long journey,” Van Andel *says*, “but a worthwhile and rewarding one.” (Peter Island Resort)

(15) “Being the older sibling, I’d sacrifice it for my brother; now I can make enough for everyone to share,” *says* the Eleven Madison Park alum who helms the kitchen at the 12-seat chef’s counter Papa’s Bombay. (Papa’s Bombay)

(16) “Travelers to the region can cross borders easily, spending less money on visas, time on checkpoints, and exploring more of the conservation area on a single trip,” *says* Barbara Murasiranwa-Hughes, regional representative for the Tourism Business Council of Zimbabwe. (Mbano Manor Hotel)

Table 8. Keyness Analysis (Target Corpus: the TJC; Reference Corpus: the OTWC)

No.	Type	Keyness (Likelihood)	Keyness (Effect)
1	*says*	27.936	0.004
2	*new*	27.797	0.007

Note: 3.84, $p < .05$; 6.63, $p < .01$; 10.83, $p < .001$; 15.13, $p < .0001$[xxii]

4.4 *Across Corpora Analysis*

4.4.1 Type-Token Ratio (TTR) and Standardized Type-Token Ratios (STTR) across the Three Corpora

The TTR and STTR values provide an indication of lexical diversity across the three corpora. Because the raw TTR decreases as corpus size increases (McEnery & Hardie, 2012, p. 50), the raw TTR values in Table 1 are not directly comparable: the TJC is roughly three to four times larger than the OTWC and the UGTC. The almost identical raw TTRs of the OTWC (25.51%) and the TJC (25.52%) therefore conceal a real difference, because the TJC achieves the same ratio over a much larger number of tokens. The STTR, calculated over consecutive 500-token segments, controls for this effect and is therefore used as the primary measure here. On this measure, the TJC (STTR = 65.34%) is clearly the most lexically diverse corpus, followed by the OTWC (53.32%) and the UGTC (51.55%). The relatively low value for the UGTC is consistent with the repeated use of a small set of evaluative and experience-based items that characterizes user-generated reviews, whereas the high value for the TJC reflects its coverage of 100 different destinations. Even so, the STTR values are interpreted descriptively: the UGTC yields only about 12 segments of 500 tokens, compared with about 18 for the OTWC and 50 for the TJC, so the small difference between the OTWC and the UGTC (1.77 percentage points) should not be over-interpreted.

4.4.2 Convergence and Divergence across Corpora

Beyond the individual corpus analyses, a cross-corpus comparison reveals that the three discourse types construct destination image through different combinations of content, stance, and voice. The OTWC foregrounds cultural heritage and performing arts in Takarazuka City through a promotional voice, thereby projecting an officially constructed destination image. By contrast, the UGTC constructs a perceived destination image through first-person experience, subjective evaluation, and practical visitor concerns, which reflects an experiential and highly personal communicative stance. The TJC, meanwhile, combines destination description with novelty, locality, and direct quotations, resulting in an editorially mediated, multi-voiced representation of destinations. Taken together, these findings suggest that destination image is not simply projected by tourism organizations

but is co-constructed through the interaction of institutional, experiential, and editorial discourses.

The cross-corpus comparison also identifies both convergence and divergence in the lexical display of destination image. As shown in Table 3-3, only two high-frequency lexical items, *hotel* and *museum*, are shared across all three corpora, indicating that accommodation or tourism infrastructure and cultural attractions constitute common cognitive attributes of destination image regardless of discourse type. By contrast, the greater prominence of *manga* in the UGTC relative to the OTWC and performing arts-related lexical items *(revue*, *theater*, *grand*, and *hall*) in the OTWC relative to the UGTC reveals a mismatch between visitors' perceived destination image and the officially projected destination image. From a practical perspective, this divergence suggests that official tourism websites could place greater emphasis on Takarazuka's manga-related resources while providing more accessible and user-friendly information on theater performances. Such improvements may help narrow the gap between destination promotion and visitors' actual interests and experiences.

5. Conclusion, Limitations, and Future Prospects

This study has presented a corpus-based case study of destination image construction in the digital era by comparing three English-language tourism corpora compiled by the author: the Official Tourism Website Corpus (OTWC), the User-Generated Tourism Corpus (UGTC), and the Travel Journalism Corpus (TJC). Through word lists, semantic categories, and keyness analyses, the study examined how destination image is linguistically constructed across different types of tourism discourse.

Regarding RQ1, the analysis demonstrated that each corpus exhibits distinctive lexical and semantic characteristics reflecting its communicative purpose and style. The OTWC is characterized by destination-specific lexical items related to cultural heritage, performing arts, tourism infrastructure in Takarazuka City, and promotional language, officially projecting the city primarily as a culturally distinctive destination. In contrast, the UGTC contains a greater proportion of evaluative expressions, first-person pronouns, and accommodation-related vocabulary, reflecting visitors' or tourists' experiences and their perceptions of the destination. The TJC is characterized by a more diverse lexical profile, including markers of novelty and locality, and direct quotations introduced by the reporting verb *says*, revealing distinctive features of travel journalism in the online tourism environment.

Regarding RQ2, the comparison across corpora demonstrated that the destination image of Takarazuka City constructed by the UGTC is different from that formed by the OTWC. The OTWC projects destination image from an institutional and promotional perspective, whereas the UGTC constructs a perceived destination image through first-hand experiences and subjective evaluation. The comparison further reveals both convergence and divergence between the two discourse types. While cultural attractions and tourism infrastructure such as *museums* and *hotels* are commonly represented across the three corpora, notable discrepancies are also observed in the prominence of manga-related discourse and performing arts, suggesting that the officially projected destination image does not always fully correspond to visitors' perceptions.

Regarding RQ3, the comparison with the international travel journalism corpus suggests several implications for improving English-language tourism communication on official tourism websites. In particular, incorporating local perspectives, direct quotations from residents and tourism practitioners, and greater emphasis on novelty may enhance the

authenticity, credibility, and international appeal of official tourism websites. These findings indicate that effective destination communication may benefit from integrating institutional promotion with visitors' experiences and editorial storytelling.

Theoretically, this study contributes to tourism discourse research and applied corpus linguistics by demonstrating that destination image is not constructed solely through official tourism promotion but is co-constructed through institutional, experiential, and editorial discourses. Methodologically, it illustrates the value of combining word list analysis, semantic categorization, and keyness analysis in examining projected and perceived destination images from complementary perspectives. It also shows that small, purpose-built specialized corpora, when analyzed with dispersion information, explicit significance thresholds, and effect sizes, can reveal interpretable differences between closely related genres of the same discourse domain. Overall, the findings suggest that destination image in contemporary tourism communication should be understood as a dynamic and negotiated outcome emerging from the interaction of institutional, experiential, and editorial voices rather than as a unidirectional product of official destination marketing.

Several limitations should be acknowledged. First, the study focuses on a single Japanese destination and therefore the findings cannot be generalized to all tourism destinations. Second, although the three corpora represent different discourse types, their sizes and sources are not identical, which may influence lexical distribution despite the use of normalized frequency and keyness analysis. Third, the UGTC spans 2011–2025, whereas the OTWC and the TJC represent 2025, so diachronic change in visitors' reviews and in the website itself cannot be ruled out. Fourth, the semantic categorization was carried out manually by a single analyst, and the 4-gram analysis relied on low frequency and range thresholds. Fifth, the analysis is restricted to verbal text and does not consider images or other multimodal resources. Future research could investigate multiple destinations across different countries, incorporate additional discourse genres and multimodal resources, test the manual categories with automatic semantic tagging (e.g., the USAS tagger in Wmatrix; Rayson, 2008), compare each corpus with a large general reference corpus of English, and employ more advanced corpus techniques such as collocation and concordance-based analyses of evaluation to provide a more comprehensive understanding of destination image construction in digital tourism communication.

Declaration regarding usage of generative AI (Artificial Intelligence) tools

This manuscript made use of ChatGPT to assist in searching for previous studies in related fields and in paraphrasing and improving the clarity of text. No AI tools were used for the generation of research hypotheses or the writing of conclusions. All suggestions generated by ChatGPT were carefully reviewed, selectively adopted, and revised as necessary by the listed author, who takes full responsibility for the final content of the manuscript.

Conflict of Interest

The authors declare no conflicts of interest regarding this manuscript.

References

Anthony, L. (2022). *AntConc* (Version 4.2.0) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in. keywords analysis. *Journal of English Linguistics, 32*(4), 346–359. https://doi.org/10.1177/0075424204269894

Baloglu, S., & McCleary, K. W. (1999). A model of destination image formation. *Annals of Tourism Research, 26*(4), 868–897. https://doi.org/10.1016/S0160-7383(99)00030-4

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Brett, D. F. (2023). Is Travel Journalism more similar to Newspaper Language or the Language of Tourism? A corpus-based study. *Studies in Travel Writing*, *26*(2), 152-167. https://doi.org/10.1080/13645145.2023.2261635

Brown, G., & Yule, G. (1983). *Discourse Analysis*. Cambridge University Press.

Crompton, J. L. (1979). An assessment of Mexico as a vacation destination and the influence of geographical location upon that image. *Journal of Travel Research, 17*(4), 18–23. https://doi.org/10.1177/004728757901700404

Dann, G.M.S. (1996). *The Language of Tourism: A Sociolinguistic Perspective*. CABI.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19* (1), 61–74.

Fu, K. (2024). Text Analysis on Toyooka-based Hotel Customers' Online Reviews. *Journal of Arts and Tourism, 3*(1), 193-200. https://doi.org/10.60292/artsandtourism.3.0_194

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 225–258). Routledge.

Gartner, W. C. (1994). Image Formation Process. *Journal of Travel & Tourism Marketing*, *2*(2–3), 191–216. https://doi.org/10.1300/J073v02n02_12

Gotti, M. (2006). The language of tourism as specialized discourse. In O. Palusci & S. Francesconi (Eds.), *Translating tourism: Linguistic/cultural representations* (pp. 15–34). Editrice Università degli Studi di Trento.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Hardie, A. (2014, April 28). *Log Ratio: An informal introduction*. ESRC Centre for Corpus Approaches to Social Science (CASS), Lancaster University.

Hunston, S., & Thompson, G. (Eds.). (2000). *Evaluation in text: Authorial stance and the construction of discourse*. Oxford University Press.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66–79). Routledge.

Lekant, M., & Palau-Sampio, D. (2022). Reinventing travel journalism in the digital age: Quality, specialization, technology, and a unique perspective. *Comunicación y sociedad*, *19*. https://doi.org/10.32870/cys.v2022.8250

Li, W., & Du, W. (2020). Understanding destination Image in UGC: A lexicon approach. *Journal of Travel, Tourism and Recreation*, *2*(1), 13-28. https://doi.org/10.22259/2642-908X.0201002

Malenkina, N., & Ivanov, S. (2018). A linguistic analysis of the official tourism websites of the seventeen Spanish Autonomous Communities. *Journal of Destination Marketing & Management*, *9*, 204-233. https://doi.org/10.1016/j.jdmm.2018.01.007

Manca, E. (2016). *Persuasion in tourism discourse: Methodologies and models*. Cambridge Scholars Publishing.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nozu, N., Yao, Y., Kawamura, T., Fu, K., & Takahashi, N. (2023). A comparative study on temporary linguistic landscapes in local art festivals: Toward application to the Toyooka Theater Festival. *Journal of Arts and Tourism, 2*, 41-57. https://doi.org/10.60292/artsandtourism.2.0_41

Pagano, N. (2023). Keywords in UNESCO websites: The role of image in destination marketing. *International Journal of Islands Research*, *4* (1), 17-32. https://doi.org/10.21427/9GEX-1A64

Pagano, N., & De Cantis, S. (2018). Tourism destination image, tourism discourse and UNESCO sites: A contrastive analysis. In P. Calò & G. Ruggieri (Eds.), *10th International Conference on Islands Tourism: Book of full papers* (pp. 160–170). Observatory on Tourism for Islands Economy (OTIE).

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS).* John Benjamins Publishing Company. https://doi.org/10.1075/scl.55

Partington, A., & Morley, J. (2004). At the heart of ideology: Word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk (ed.), *PALC 2003: Practical Applications in Language Corpora*, 179–192. Peter Lang.

Pierini, P. (2009). Adjectives in tourism English on the web: A corpus-based study. *Círculo de Lingüística Aplicada a la Comunicación, 40*, 93–116.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519–549. https://doi.org/10.1075/ijcl.13.4.06ray

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117730

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins.

Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. In Stefanowitsch & Gries (eds), *Corpus-based approaches to metaphor and metonymy*, 1–16. Walter de Gruyter.

Thornborrow, J., & Wareing, S. (1998). *Patterns in Language: An Introduction to Language and Literary Style.* Routledge.

Vásquez, C. (2014). *The discourse of online consumer reviews*. Bloomsbury.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, *27*(2), 634-639. https://doi.org/10.1016/j.chb.2010.04.014

Zainol, Z., Lee, A. S., Nohuddin, P. N. E., Ibrahim, N. F., & Hijazi, M. H. A. (2022). Examining the relationship of keyword analysis using online traveller hotel reviews. *International Journal on Perceptive and Cognitive Computing*, *8*(1), 47-52.

Websites and Resources

https://chatgpt.com/ (Retrieved July 26, 2026)

https://kanko-takarazuka.jp/english/ (Retrieved July 23, 2026)

https://kanko-takarazuka.jp/english/lp_introduction/ (Retrieved March 24, 2025)

https://time.com/ (Retrieved July 25, 2026)

https://time.com/5375033/2018-worlds-greatest-places-choice/ (Retrieved July 21, 2026)

https://time.com/collections/worlds-greatest-places-2025/ (Retrieved July 23, 2026)

https://ucrel.lancs.ac.uk/llwizard.html (Retrieved July 25, 2026)

https://www.britannica.com/search?query=%22god+of+manga%22 (Retrieved July 25, 2026)

https://www.city.takarazuka.hyogo.jp/1060683/index.html (Retrieved July 25, 2026)

https://www.laurenceanthony.net/software/antconc/ (Retrieved July 25, 2026)

https://www.theguardian.com/uk/travel (Retrieved July 2, 2026)

https://www.tripadvisor.com/ (Retrieved July 26, 2026)

https://www.trip.com/ (Retrieved July 26, 2026)

Author biodata

Kenryo Fu is an Associate Professor at Professional College of Arts and Tourism in Hyogo, Japan. He holds a Ph.D. in Language, Communication, and Culture from Kwansei Gakuin University. His research interests include English linguistics, semantics, contrastive linguistics, corpus linguistics, and linguistic landscapes, with particular attention to language use in tourism and intercultural communication. He is the author of *The Present Perfect in English: From Semantic, Evolutionary, and Contrastive Perspectives* (Kaitakusha, 2021). His recent research focuses on linguistic features of tourism discourse, online reviews, and language use in public spaces.

Email: fujianliang@gmail.com
ORCID: 0000-0001-6773-8788

[i] Retrieved July 21, 2026, from https://time.com/5375033/2018-worlds-greatest-places-choice/.

[ii] Refer to https://www.theguardian.com/uk/travel (Retrieved July 2, 2026)

[iii] In Brett (2023), the data excluded the occurrences of the first-person pronouns occurring in direct quotations from the travel subcorpus in order to focus on the journalist's own narrative voice.

[iv] Refer to https://kanko-takarazuka.jp/english/.

[v] Refer to https://www.tripadvisor.com/. Reviews written in languages other than English or translated using machine translation were excluded.

[vi] Refer to https://www.trip.com/. Reviews written in languages other than English or translated using machine translation were excluded.

[vii] Refer to https://time.com/.

[viii] AntConc (Version 4.2.0) was released on December 25, 2022. For more information, refer to https://www.laurenceanthony.net/software/antconc/.

[ix] The 76 reviews in the UGTC are distributed across the collection period as follows: 18 reviews from 2011–2015 (23.7%), 32 from 2016–2020 (42.1%), and 26 from 2021–2025 (34.2%).

[x] To compare lexical diversity across the three corpora in this study while minimizing the influence of differences in corpus size, standardized type-token ratios (STTRs) were calculated using fixed segments of 500 tokens.

[xi] OTWC, retrieved March 24, 2025, from https://kanko-takarazuka.jp/english/lp_introduction/; UGTC, retrieved July 15, 2026, from https://www.tripadvisor.com/ and https://www.trip.com/; TJC, retrieved July 23, 2026, from https://time.com/collections/worlds-greatest-places-2025/.

[xii] The auxiliary verbs *have* (raw frequency = 32) and *has* (raw frequency = 18) were also removed from the raw word list because they primarily serve grammatical functions and contribute little to the lexical or semantic characterization in tourism discourse.

[xiii] NormFreq in (6), also labeled as nf, short for normalized frequency, indicates the number of the occurrences of a lexical item per million words in a corpus in the present study.

[xiv] NormRange in (7), short for normalized range, indicates the proportion of texts in which a lexical item occurs in a corpus.

[xv] Refer to https://www.city.takarazuka.hyogo.jp/1060683/index.html.

[xvi] A NormFreq value of 2,973 pmw means that the lexical item manga appears 2,973 times per million words. Similarly, a NormFreq value of 898 pmw indicates that manga appears 898 times per million words.

[xvii] A NormRange value of 0.105 means that the lexical item manga appears in 10.5% of the texts in the UGTC. Similarly, a NormRange value of 0.045 indicates that manga occurs in 4.5% of the texts in the OTWC.

[xviii] Retrieved July 25, 2026, from https://www.britannica.com/search?query=%22god+of+manga%22.

[xix] "–" in Table 3-2 indicates that no item in this category was found among the high-frequency content words.

[xx] Refer to https://ucrel.lancs.ac.uk/llwizard.html. (Retrieved July 25, 2026)

[xxi] In Brett (2023, p. 4), the Travel subcorpus of SNAC consists of texts downloaded from the travel section of the online version of the well-known British newspaper, *The Guardian*. On the other hand, the Tourism corpus in Brett "is comprised of texts that appear on the websites of public organisations entrusted with the task of attracting tourists to a destination, or providing them with basic information about things to see and do once there" (ibid.).

[xxii] Refer to https://ucrel.lancs.ac.uk/llwizard.html. (Retrieved July 25, 2026)