Compiling Comparable Multimodal Corpora of Tourism Discourse

Ekaterina Ignatova

Lancaster University (UK)

Abstract

This paper describes work in progress on the design of two comparable multimodal corpora of written tourism discourse about London and Moscow. Multimodality is defined for the purposes of the current project as a combination of several discourse modes, including verbal and visual. The paper aims to contribute methodologically by providing a detailed description of the process and challenges of the multimodal corpora compilation. The building of the corpora is an essential precondition for using a multimodal corpus approach allowing to analyse a range of texts, to consider not only language but also images and layout, to search the data for patterns, to identify multimodal features of each set of texts and to compare these features across the two corpora. After introducing the project and its research questions, the paper highlights the principles of data selection. Then the planned structure of the corpora and data sources are described. The paper goes on by describing the constructed pilot corpora, as well as some technical moments of corpora building, arising problems and possible solutions. To conclude, I highlight the limitations of the article and its implications.

Keywords: tourism discourse, multimodality, corpus linguistics

Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching, Vol. 11. Papers from LAEL PG 2016–2017
Edited by Márton Petykó, Olena Rossi, and Shuo Yu

1. Introduction

Due to cheap airline fares and instant online booking services travelling is becoming more and more popular. The tourism industry is steadily growing and plays an important part in the economy of many countries (World Travel & Tourism Council 2017). Together with the advancement in technology and industry growth, the number of tourism-related websites and texts aimed at attracting tourists is increasing. Various scholars have studied the influence of such texts on travellers' behaviour (e.g. Dann 1996) and the image of a travel destination (e.g. Ip 2008; Pagano 2014; Stoian 2015). In addition, as the role of multimodality in tourism discourse is constantly developing and a variety of expressive solutions are used to promote destinations (Francesconi 2014), the interest to multimodal research of tourist texts is growing (e.g. Francesconi 2011, 2014; Manca 2016). However, there still remains a paucity of systematic multimodal analysis comparing the representation of various cities, that is, how cities are presented to the audience, in tourist texts. Therefore, the project seeks to obtain data which will help to address this research gap.

I have selected two cities for my research, London and Moscow. Both are capital cities and popular travel destinations attracting large numbers of international tourists. London has been placed as the third most visited city in the world whereas Moscow is ranked 46th in the 2017 City Destinations Ranking (Geerts, Popova, Bremner & Nelson 2017). Differences in the cultural and historical background make these destinations interesting for comparison. It should be underlined that this research project looks at original texts aimed not only at local tourists but at a wider international audience and, consequently, written in English. The research questions I have defined for the project are:

1. What are the similarities and differences of tourism texts about London and Moscow in terms of linguistic features?
2. What are the similarities and differences of tourism texts about London and Moscow in terms of visual features?
3. What are the similarities and differences of tourism texts about London and Moscow in terms of multimodality?

In line with Bednarek & Caple (2017), I define discourse as suggested by Brown & Yule (1983: xiii), namely, "language in use", but consider it as a multimodal notion. *Multimodality* can be defined for the purpose of the project as a combination of various discourse modes, or means for meaning-making, for example, language, images and layout, in a text (Van

Leeuwen 2015). It should also be noted that in this paper I do not draw any distinction between the notions *tourism discourse*, *travel discourse* and *travel-related discourse* and use these terms interchangeably.

One of the conspicuous features of present-day tourism discourse is a high impact of digital media. According to Dann (2007), the development of the internet has generated a number of new genres including customer-to-customer, or so-called C2C, communication, for instance, travelogues and online consumer reviews. It also led to the migration of traditional printed texts to the Internet, which allows to cut printing costs and what is more important to widen the audience. The boost of digital media has led to the increase in the role of multimodality in tourism discourse.

Therefore, multimodality is a distinct feature of modern travel-related texts. A range of modes is used to attract the attention of potential tourists and trigger positive emotions (Francesconi 2014) thus promoting various destinations.

There are various approaches to multimodal research, for instance, systemic functional multimodal discourse analysis, social semiotics, conversation analysis. Following Jewitt, Bessemer & O'Halloran (2016), I have chosen to use a corpus approach to multimodal analysis as it allows to search data for patterns, to identify multimodal features of each set of texts and to compare these features across the two corpora. The data for such an analysis consists of multimodal corpora, in other words, extensive computer-readable collections of multimodal documents (Bateman, Delin & Henschel 2004).

In the field of corpus approach, there are two main options for collecting data, either using an available ready-made corpus or creating your own. To my knowledge, there are no available multimodal corpora of tourism discourse about London and Moscow, therefore I have to build them in compliance with the guidelines set by the research question. In this paper, I discuss the design of the corpora. I will start by reviewing the existing literature on corpora of tourism discourse, including multimodal ones. Then, I will provide a description of the planned comparable multimodal corpora of tourism discourse about Moscow and London, namely, data, corpora structure and data sources. After that, I will move on to the pilot corpora that have been compiled and discuss some technical moments of corpora building, arising problems and possible solutions. Finally, I draw a conclusion about the limitations of the project and the impact of the work.

2. Literature review

Corpus linguistics and corpus-assisted approaches are frequently applied to study tourism discourse (e.g. Jaworska 2013, 2016, 2017; Manca 2008a, 2008b, 2013, 2018; Pierini 2009). However, most researchers account only for the size and topic of their corpora and sometimes sources of their texts. Only a few papers provide a detailed account of the corpus design. For instance, Jaworska (2013) describes the size and the structure of the corpora as well as the search terms and the sources of the data collected to analyse linguistic patterns used in English and German to represent local and international tourist attractions on popular British and German websites. Durán-Muñoz (2010) provides a detailed description of two comparable German and Spanish corpora of online promotional texts on adventure tourism, in particular, the selection criteria (e.g. complete and original texts, a reliable authorship), the size (both in words and files), the structure, the topics, the period of collection and the target audience. She also describes the corpora annotation scheme and gives an example of the corpora metadata record containing such information about the texts as, for instance, source, author, language. Moreover, she lists criteria adopted for the comparability of the corpora, namely, similar size, same domain, typology of texts and specialised level, same time period and limited geographical area. Whereas, Manca (2016) in her analysis of promotional discourse of official tourist websites of Great Britain, Italy and Australia applies different criteria for corpus comparability, namely, same communicative function, similar composition pattern and similar text type. In her recent work on analysing the keyness of adjectives in adventure tourism English promotional texts, Durán-Muñoz (2019/forthcoming) provides the protocol of semi-automatic compilation of corpora using the WebBootCat corpus building tool. However, all the abovementioned papers describe monomodal corpora containing only verbal texts. Moreover, most works using corpus approaches to analyse multimodal tourism discourse apply corpus techniques only to identify patterns in the writing mode while using qualitative approaches to study a relatively small sample of the visual mode (e.g. Cheng 2016; Francesconi 2014; Manca 2016). While Hiippala (2015) in his study of Helsinki tourist brochures provides a thorough description of the design of a multimodal corpus including writing, images and layout using Bateman's (2008) multi-layered Genre and Multimodality (GeM) scheme the focus of his research is the combination and interaction of modes in documents and not the linguistic and visual features of the representation of the city. Therefore, there is a clear lack of papers providing a detailed description of the process and

challenges of building multimodal corpora of tourism discourse containing verbal and visual modes and enabling a corpus analysis of linguistic and visual features. This paper aims to address the gap.

The described corpora of multimodal tourism discourse about London and Moscow are intended to enable the analysis of how the tourist destinations are presented to the audience both through verbal texts and images.

3. Corpora

In this part, I describe the data, data sources and the structure of the two multimodal corpora of travel-related texts about Moscow and London that I need to compile in order to conduct a multimodal corpus analysis and identify salient features of tourism texts about these two cities and see if there are any similarities and differences in the verbal and visual representation of the destinations.

3.1. Data

The section below looks at what data is required for the project. As we can see from the formulation of the question, in order to address it, I need to compare and contrast multimodal tourism texts about the two cities. As already mentioned, multimodality is an important aspect of modern travel-related texts. Consequently, the data appropriate for answering the research question should include more than one mode. Some researchers identify only more general modes in texts, for instance, textual and visual, others draw a more subtle distinction, recognizing writing, colour, image, font, layout for printed texts (Kress 2010) and hyperlinks for web pages (Lemke 2002). Moreover, geosemiotic mode, where discourses are viewed in space and time (Aboelezz 2014; Scollon & Scollon 2003), and socio-cultural context (Gillen 2011) can also be analysed within the framework of multimodal analysis.

As already mentioned, many previous works on multimodal analysis of tourism discourse use monomodal textual corpora to identify patterns in the writing mode while using qualitative approaches to study a relatively small sample of the visual mode (e.g. Cheng 2016; Francesconi 2014; Manca 2016). Hiippala (2015) in his study of Helsinki tourist brochures conducts a corpus analysis of writing, images and layout using Bateman's (2008) multi-layered Genre and Multimodality (GeM) scheme, however, the focus of his research is the combination and interaction of modes in documents and not the representation of the city. As

the major aim of my research is to see how the tourist destinations are presented to the audience through writing, images and a combination of these modes, my approach is closer to that of Bednarek & Caple (2017), who study how the news is "sold" to public with the help of writing and visual resources utilizing "corpus-assisted multimodal discourse analysis". Therefore, I am interested in the following modes: verbal texts, images and either the static layout for offline texts or the hypertextual structure for online texts. Metadata, which is background information about the text, for instance, the genre, the source and the date of collection (McEnery & Hardie 2012), is also required for the purpose of comparison in order to be able to identify where the text comes from and what its genre-specific features are.

Next, as the focus of my research question is a comparison of texts about the two cities, I need two multimodal corpora, a London corpus and a Moscow corpus. These corpora should be specialised, meaning they contain only texts belonging to a certain domain (Koester 2010), namely, tourism-related texts.

To be suitable for comparison, the corpora should be comparable, in other words, the sampling frame should be the same in terms of text genres and their proportions (McEnery & Hardie 2012), size of the texts and the time period (Kenning 2010).

Regarding the size of the corpora, as they are aimed at examining specific peculiarities of tourism discourse and not rare linguistic features of the English language in general, the size of the corpora can be secondary, provided they comply with the aforementioned criteria (Koester 2010).

As for the genre of texts, the main focus of the current project is discourse aimed at a wider audience. Therefore, a variety of tourism texts can be collected including so-called business-to-consumer (B2C) genres, in other words, traditional genres written by tourist industry participants for prospective travelers, such as travel magazines, travel guides, city overviews, descriptions of accommodation, restaurants and sights, as well as new C2C genres, for instance, online consumer reviews of sights, places to eat and stay, and travelogues. The main challenge here is to ensure that the structure of both corpora is the same and therefore comparable.

3.2. Corpora structure

In the section that follows I describe the planned structure of the two corpora. It should be noted that there is a difference in the modes that constitute online and offline texts, namely, that offline texts have static layouts, in other words, their page elements are organized in a certain way (Bateman 2008), whereas online texts usually also have a hypertextual structure, or links connecting various elements of the site (Lemke 2002). Therefore, a decision has been made that each corpus will consist of two subcorpora: a subcorpus of online tourism discourse and a subcorpus of offline tourism discourse. Furthermore, each subcorpus will have three parts corresponding to the modes analyzed, namely, writing, images and a hypertextual structure for online discourse and writing, images and static layout for offline discourse. Figure 1 displays the planned structure of the two multimodal corpora.

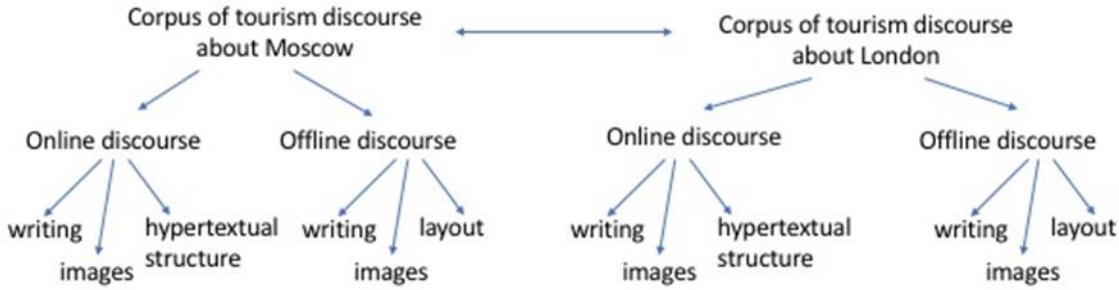

Figure 1. ***Planned structure of multimodal corpora of tourism discourse about Moscow and London***

3.3. Data sources

In this section, I describe the planned sources of texts for the corpora. Firstly, the inclusion of both online and offline texts means that various sources of data collection are to be used. The online texts will be collected from travel-related websites. For instance, *www.booking.com* can serve as a good source of customer reviews about hotels and apartments. Descriptions and reviews of tourist sights, restaurants and accommodation can be found on such popular tourist portals as *www.tripadvisor.com*, whereas official websites of travel destinations provide texts written by tourist industry participants for travellers.

As for offline texts, some of them, such as travel guides or travel magazines can be found in libraries. While field trips to London and Moscow might be required in order to obtain the data unavailable in libraries, for example, posters and tourist information brochures.

4. Pilot project

4.1. TripAdvisor corpora

The current section provides an overview of the pilot corpora that are being compiled at the moment. The data for the pilot project consists of two comparable corpora of multimodal texts about Moscow and London collected from a popular tourist website *www.tripadvisor.com* (hereinafter referred to as TripAdvisor).

I have selected the website for the pilot study as it is one of the most popular tourist portals and a travel community attracting users at all stages of their trip. At the pre-trip stage, when they are only planning their journey, they can read reviews of accommodation, sights and restaurants, compare rankings and prices and make their choice. During the trip, the website can be consulted for itinerary, timetables and directions, whereas at the post-trip stage, travellers share their experience, give recommendations and post favourable or negative reviews and photographs. The portal has become a powerful player in the industry claiming to attract more than 400 million unique visitors each month on average (TripAdvisor 2017). Nowadays it also offers online booking of hotels, flight tickets, guided tours and other tourist services, which means that consumers can research and book their trips on one website. The TripAdvisor mobile app makes the content easily accessible even when on the move. The portal contains information provided both by travellers and by tourist industry participants, for instance, hotels and restaurants management, who represent themselves on the site in order to be able to reach potential travellers.

All the texts included in the corpora belong to the same time period (between 2017 and 2018) and are originally written in English (no translations included). The structure of the two pilot corpora is identical and balanced in the sense that the texts from both corpora are on similar topics corresponding to the sections of the website (namely, places to stay, places to eat and things to do) and the same proportion of texts from each section of the website is represented. Both corpora consist of three parts each representing a different mode, namely, hypertextual structure, writing and images. Figure 2 shows the sources of data for each of the three models.

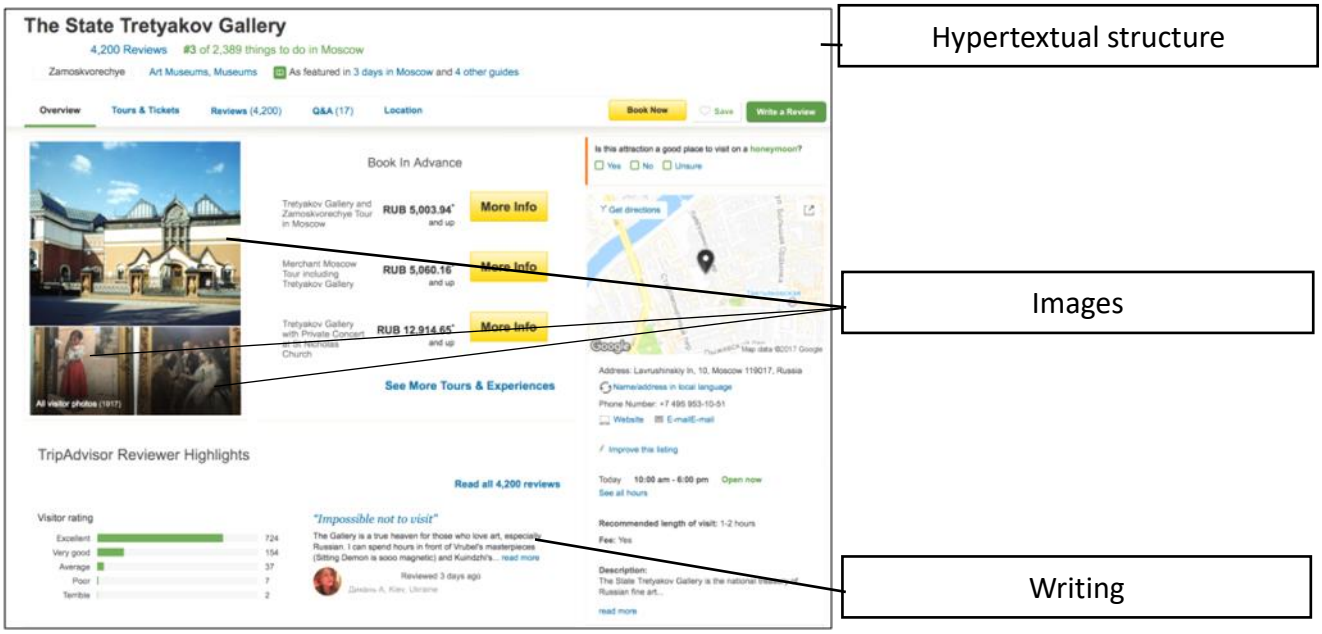

Figure 2. ***Sources of data for different discourse modes***

As can be seen from Figure 2, the data for the hypertextual structure part of the corpora contain whole webpages. Each webpage was saved as a webarchive file. Unlike screenshots, webarchive files store not only the visible part of the page but the whole of it making the page scrollable and links clickable. The hypertextual structure part of each corpus contains a total of 44 webarchive files:

- A city overview – one file
- Rankings of top 10 places to stay, such as hotels, apartments and other lodgings (hereinafter all top items are determined in accordance with the traveller ranking) – one file
- Pages featuring each of the top 10 places to stay containing 10 most recent traveller reviews on the date of collection – 20 files (one page contained only five reviews, so in order to capture 10 reviews I had to save two pages for each place)
- Rankings of top 10 places to eat, namely, restaurants or cafes – one file
- Pages featuring each of the top 10 places to eat containing 10 most recent traveller reviews on the date of collection – 10 files (one page contained 10 reviews, so only one page per place was included)
- Rankings of top 10 things to do, such as landmarks, sights – one file

- Pages featuring each of the top 10 things to do containing 10 most recent traveller reviews on the date of collection – 10 files (one page contained 10 reviews, so only one page per place was included).

The writing part of each corpus contains a total of 331 plain text files with written texts from the abovementioned pages:

- A city overview – one file
- Descriptions of top 10 places to stay – 10 files
- 10 most recent reviews on the date of collection of the 10 top hotels – 100 files altogether;
- Descriptions of top 10 places to eat – 10 files
- 10 most recent reviews on the date of collection of the 10 top places to eat – 100 files altogether.
- Descriptions of top 10 things to do – 10 files
- 10 most recent reviews on the date of collection of the 10 top things to do – 100 files altogether.

Due to the fact that the texts comprising the corpora are different in size, the corpora are also slightly unequal. The London writing part contains 26,973 words whereas the Moscow writing part contains 24,722 words.

The image parts of the corpora include all photos from the pages listed above. The number of photos on each page is different and ranges from 21 to 73. Such visual objects as maps, advertisements and icons, also contained by the pages, are not included in the image data set. Although all these elements contribute to the multimodal representation of the destinations, in case of TripAdvisor they are similar for all destinations and, therefore, are not an area of interest in the current study. Due to the fact that not all images for dining places have been saved in separate files and counted, it is impossible to give the exact figures. However, according to the preliminary count, the London corpus will contain above 1,200 images and the image part of the Moscow corpus will be roughly 10% larger. Table 1 gives an overview of the pilot corpora size.

Table 1. ***Size of London and Moscow pilot corpora***

	London	Moscow
Hypertextual structure	44 webarchive files	44 webarchive files
Writing	26,973 words	24,722 words
Images	approx. 1,200 images	approx. 1300 images

4.2. Technical moments of corpora building

In this section, I write about some technical moments of corpora building which are rarely discussed in research papers on corpus-based analysis of tourism discourse. The building of multimodal corpora is still at the very early stages of development, so there is considerable variation in how researchers address it (e.g. Bednarek & Caple 2017; Francesconi 2014; Hiippala 2015). However, some of the ideas might be useful for future research.

For the pilot corpora, each corpus item is stored in a separate file, so that various sections of the corpora can be compared, for instance, only hotel consumer reviews. As texts from the website are considered as online data, verbal, visual and hypertextual modes are captured for each text. Therefore, the verbal data is stored as plain text files (.txt), which is a format required by some corpus software, for instance, AntConc (Tang 2013). Then, images are saved in JPEG files (.jpg), so that they can later be tagged and analysed. As regards the hypertextual structure, which is required to analyse the context of the usage of texts and images, I store web pages as webarchive files (.webarchive). Unlike screenshots, webarchive files store not only the visible part of the page but the whole of it including the hypertextual structure and hyperlinks, thus making the page scrollable and links clickable. Finally, two Excel databases, one for the London corpus and the other for the Moscow corpus, contain lists of all files with metadata.

F44 f_x

	A	B	C	D	E	F
1	**Item**	**file type**	**file name**	**retrieved from**	**date of retrieval**	**comments**
129	Hotel Indigo London-Paddington	webarchive-1	Hotel Indigo London-Paddington 050517-1.webarchive	https://www.tripadvisor.com/Hotel_Review-g186338-d1139866-Reviews-Hotel_Indigo_London_Paddington-London_England.html	05/05/2017	reviews 1-5, Sponsored
130	Hotel Indigo London-Paddington	webarchive-2	Hotel Indigo London-Paddington 050517-2.webarchive	https://www.tripadvisor.com/Hotel_Review-g186338-d1139866-Reviews-or5-Hotel_Indigo_London_Paddington-London_England.html#REVIEWS	05/05/2017	reviews 6-10
131	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-1.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r480982497-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response
132	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-2.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r479960538-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response
133	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-3.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r479520496-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response
134	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-4.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r478441465-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response
135	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-5.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r477814562-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response
136	Hotel Indigo London-Paddington	review	Hotel Indigo London-Paddington 050517-6.txt	https://www.tripadvisor.com/ShowUserReviews-g186338-d1139866-r477715650-Hotel_Indigo_London_Paddington-London_England.html#CHECK_RATES_CONT	05/05/2017	hotel response

Figure 3. ***Screenshot of London corpus Excel database***

Figure 3 presents a screenshot of the London corpus Excel database, which includes item description (here it is the name of a hotel), file type, file name, source and date of retrieval. As you can see, the file name also contains the name of the place and date of retrieval, whereas the file extension tells us about the file type. Although Reppen (2010) warns against having long file names and suggests that by using not more than eight characters in a name one can avoid problems with software for analysis and data backup, I prefer names that do not require deciphering and can be easily understood by other researchers. Moreover, I have not encountered any issues while processing the corpora with the AntConc and LancsBox software and while backing up the data onto the AirPort Time Capsule device.

In the comments section, I add such information as what type of search results sorting has been applied on the website, whether the search results were sponsored, meaning a place appears on top of search results because it has been paid for, and whether the management of a place posted a response to the consumer review. This metadata provides some useful context for a more detailed analysis and interpretation of the findings.

As regards corpus analysis, the main aim of the study is the comparison of the representation of two tourist destinations in texts in terms of three aspects, namely, language, images and their combination. Therefore, Baker's (Baker 2010; Baker & Levon 2015) approach for corpus-based analysis and comparison of media representation is applicable to the verbal mode of my corpora. The following techniques can be applied for analysis. Firstly,

keywords comparison, which means analysing how often all the words appear in one corpus compared to the other in order to identify the most noticeable ways of representing the two cities. Secondly, concordance analysis, that is studying lines of texts from the corpus showing the keywords in context (McEnery & Hardie 2012), which allows to "explain why certain words occurred as keywords, what their most common uses were and whether there were similarities or differences" (Baker 2010: 317) between the corpora. The third technique, collocation analysis, explores which words co-occur more often and therefore are more associated with the search terms, in this study, London and Moscow. And finally, concordance analysis of these collocations looks at how they were used in context.

The described framework of verbal text analysis requires only part-of-speech annotation allowing to conduct such advanced searches as, for instance, adjectives most commonly co-occurring with London.

As for the visual mode, I would like to apply a corresponding methodology for analysing and comparing corpora containing images in terms of tags and their frequency. I plan to tag the images with words describing the represented objects (a place, people or thing, including abstract things, represented in the picture) and analyse which of the tags are more frequent in the London corpus compared to the Moscow corpus and vice versa. This technique is aimed at exploring the patterns, as well as similarities and differences in the visual representation of the two cities. Looking at how these images are used in context, namely, which texts they accompany, might be the second stage of the analysis aimed at the interpretation of these patterns.

4.3 Problems and possible solutions

This section describes some problems that I experienced while compiling the pilot corpora. First of all, online data is constantly changing. The standard approach is to include the source and the date of access into metadata in order to provide the information about when and where the data has been retrieved from. However, this does not solve the problem as new reviews and images are added to the website, hotels and restaurants ratings are influenced by these reviews and, therefore, the pages might look completely different in just a few months. For instance, 10 of the 24 images on the London overview page of TripAdvisor have changed in November 2017 compared to April 2017. Moreover, the website might be redesigned. My approach here was building both corpora simultaneously and parallel. I started with the top

sights of both cities, then accommodation and places to eat. It was also useful capturing webarchives of the main pages first, which then allowed saving embedded texts and images in separate files.

Another issue was the different size of the writing and image parts of the corpora. As for writing parts, in spite of the equal number of text documents in both corpora, the length of the verbal texts comprising the corpora was different and, as a result, the London pilot corpus was almost 5% larger than the Moscow pilot corpus. Actually, this is a common problem for corpora based on document count and not on word count. Due to the fact that the original reviews were very short and some interesting features could appear at the end or at the beginning of the texts, I have made a decision at the start of the project not to take sample parts from the texts but to use relative frequencies. In this respect, relative frequencies per 10,000, showing how many times a search term occurs in a corpus per 10,000 words, can be used to compare the results despite the difference in the corpora size (McEnery & Hardie 2012). As regards image parts, according to rough estimates, the size of the Moscow corpus is 10% larger than of the London corpus. Here relative frequencies of images can also be applied to compare the two corpora.

The next issue was spelling irregularities and mistakes, for example, "*wil*" instead of *will*, and "*mist"* instead of *must.* As such irregularities were not the primary concern of the study, they were manually detected in the data and standardized, in other words, corrected to ensure that all the occurrences of a search term are included in the search results. However, this approach is time-consuming and not suitable for a larger corpus. Therefore, in the main project, I plan to use a software tool for finding spelling variants of a search term, such as VariAnt.

Another question is what reference corpus to use for analysis. One of the corpus techniques I plan to use in my research of the writing mode is keywords analysis, which allows identifying words that occur more frequently in a corpus in question than in another, usually larger, reference corpus (McEnery & Hardie 2012). Due to the comparative nature of the project, first of all, I will be looking at the keywords of the Moscow corpus in comparison with the London corpus and vice versa in order to identify the differences. However, it might also be interesting to compare the compiled corpora to a larger corpus. According to Scott (2010), the choice of a reference corpus influences the results of the keyword analysis. Whereas, Culpeper (2009) notes that content, size and date are the three main aspects that should be taken into account when selecting a reference corpus. However, regarding the size,

Xiao and McEnery (2005) compared the results of a keyword analysis based on the one-million-word Freiburg-LOB Corpus of British English (FLOB) and on the 100-million-word British National Corpus (BNC) and concluded that the results were very similar and, consequently, the size of a reference corpus is not the most important factor in keywords analysis.

Some of the previous studies use a reference corpus of general British English, for instance, BNC (Cesiri 2017) or FLOB (Cesiri 2017; Kang & Yu 2011) for verbal texts. On the one hand, such comparison might provide some interesting findings, for instance, the difference in the usage of modals in tourism discourse and general English. On the other hand, using a corpus of general English as a reference for keywords analysis means that most of the keywords in the compiled corpora will relate to tourism and the results of such analysis are quite predictable. However, the focus of the current study is not on the keywords characteristic of the tourism domain in general but on those used to represent city destinations. In this respect, conducting an additional comparison with a specialised reference corpus of tourism discourse might provide valuable insights into how the texts about the two cities are different from texts about other destinations. According to preliminary search, the one-million-word Tourism English Corpus (Jiansheng 2012) containing a variety of texts from brochures, guides, forum posts, journal articles, ordinances and travelogues might be interesting for comparison if available. As regards a reference corpus of images, the problem is that, to my knowledge, there is no available multimodal corpus of tourism discourse containing images and including a comparable variety of text genres. Building a larger reference corpus myself is a very challenging task going far beyond the scope of this project. Therefore, this question still remains open and requires further consideration.

5. Conclusion

To summarise, in this work I discussed the design of the two multimodal corpora of tourism discourse about such travel destinations as Moscow and London aimed at studying the representation of the cities. Firstly, I briefly introduced the research project which the corpora are intended for. In the overview of the existing works on specialized corpora of tourism discourse, I demonstrated that there is a lack of papers on the process and challenges of compiling multimodal corpora of tourism discourse. Next, the paper provided a description of the data to be included in the corpora, the planned structure including such modes as writing,

images and static layout/ hypertextual structure and possible sources of data. I have also shared my experience of working on the pilot multimodal corpora of tourism discourse about Moscow and London sourced from the TripAdvisor website. In this respect, a detailed description of the data, corpora structure and size has been provided. Moreover, such technical moments of corpora building have been considered as metadata storage, file organization, file formats and naming. Finally, the problems I encountered while compiling the pilot corpora have been highlighted. Satisfactory solutions have been found for some of the issues, namely the changing online data can be captured by saving webarchive files. As regards the difference in the size of the corpora, relative frequencies allow conducting a comparison of unequal datasets. Whereas, spelling irregularities can be detected and corrected either manually or with the VariAnt software depending on the corpus size. The question of selecting a reference corpus requires further consideration.

It should be underlined, that the described corpora design and solutions are only one of the many possible approaches to multimodal corpora building. Another limitation is that compiling the corpora is only the first step in the analysis of the representation of the cities in multimodal tourism texts and the methodology and results of such analysis are beyond the scope of the present article and are to be addressed in a separate paper. However, it is hoped that this work will contribute to the challenging process of developing multimodal corpora of tourism discourse and will help other researchers to produce their own ideas.

References

Aboelezz, M. (2014). The Geosemiotics of Tahrir Square: A study of the relationship between discourse and space. *Journal of Language and Politics, 13*(4), 599-622.

Baker, P. (2010). Representations of Islam in British broadsheet and tabloid newspapers 1999-2005. *Journal of Language and Politics, 9*(2), 310-338. doi:10.1075/jlp.9.2.07bak

Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication, 9*(2), 221-236. doi:10.1177/1750481314568542

Bateman, J. (2008). *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*: Springer.

Bateman, J., Delin, J., & Henschel, R. (2004). Multimodality and empiricism. *Perspectives on multimodality, 6*, 65-87.

Bednarek, M., & Caple, H. (2017). *The discourse of news values: how news organizations create newsworthiness*. New York: Oxford University Press.

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

Cesiri, D. (2017). Balancing Tourism Promotion and Professional Discourse: A Corpus-based Analysis of Digital Travel Guidebooks Promoting Venice in English. In C. Vargas-Sierra (Ed.), *Professional and Academic Discourse: an Interdisciplinary Perspective* (pp. 247-255).

Cheng, F.-W. (2016). Constructing hotel brands: A multimodal analysis of luxury hotel homepages. *Iberica, 31*, 83-108.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics, 14*(1), 29-59.

Dann, G. (1996). *The Language of Tourism: A Sociolinguistic Perspective*. Oxon: CAB INTERNATIONAL.

Dann, G. (2007). Revisiting the language of tourism: What tourists and tourees are saying. In C. De Stasio & O. Palusci (Eds.), *The Languages of Tourism: Turismo e meediazione* (pp. 15-32). Milan: Unicopli.

Durán-Muñoz, I. (2010). A Corpus-based Ontoterminological Tool for Tourist Translations. *International Journal of Translation, 22*(1-2), 149-165.

Durán-Muñoz, I. (2019/forthcoming). Adjectives and their Keyness. A Corpus-based Analysis in English Tourism. *Corpora, 14*(3).

Francesconi, S. (2011). Images and writing in tourist brochures. *Journal of Tourism and Cultural Change, 9*(4), 341-356.

Francesconi, S. (2014). *Reading tourism texts: a multimodal analysis* (Vol. 36). Bristol: Channel view publications.

Geerts, W., Popova, N., Bremner, C., & Nelson, P. (2017). *Top 100 City Destinations Ranking: WTM London 2017 Edition*. Retrieved from http://www.internationalsupermarketnews.com/top-100-city-destinations-ranking-wtm-london-2017-edition/.

Gillen, J. (2011). *Three diverse projects on multimodality - is it possible to bring CHAT together with the social semiotic approach?* Paper presented at the International Society for Cultural and Activity Research Congress, Rome, Italy.

Hiippala, T. (2015). *The Structure of Multimodal Documents: An Empirical Approach* . London and New York: Routledge.

Ip, J. Y. L. (2008). Analyzing tourism discourse: A case study of a Hong Kong travel brochure. *LCom Papers, 1*, 1-19.

Jaworska, S. (2013). The quest for the “local” and “authentic”: corpus-based explorations into the discursive constructions of tourist destinations in British and German commercial travel

advertising. In D. H öhmann (Ed.), *Tourismuskommunikation: Im Spannungsfeld von Sprach- und Kulturkontakt* (pp. 75-100). Frankfurt am Main: Peter Lang.

Jaworska, S. (2016). A comparative corpus-assisted discourse study of the representations of hosts in promotional tourism discourse. *Corpora, 11*(1), 83-111. doi:https://doi.org/10.3366/cor.2016.0086

Jaworska, S. (2017). Metaphors We Travel by: A Corpus-Assisted Study of Metaphors in Promotional Tourism Discourse. *Metaphor and Symbol, 32*(3), 161-177. doi:10.1080/10926488.2017.1338018

Jewitt, C., Bessemer, J., & O'Halloran, K. (2016). *Introducing multimodality*. London and New York: Routledge.

Jiansheng, C. (2012). *Construing Experience in Tourism Discourse: A Corpus-based Study of Transitivity System.* (PhD thesis), The Hong Kong Polytechnic University, Hong Kong.

Kang, N., & Yu, Q. (2011). Corpus-based Stylistic Analysis of Tourism English. *Journal of Language Teaching and Research, 2*(1), 129-136.

Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 487-500). London and New York: Routledge.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 66-79). London and New York: Routledge.

Kress, G. (2010). *Multimodality: a social semiotic approach to contemporary communication*. London and New York: Routledge.

Lemke, J. L. (2002). Travels in hypermodality. *Visual Communication, 1*(3), 299-325. doi:10.1177/147035720200100303

Manca, E. (2008a). From phraseology to culture: Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics 13*(3), 368-385.

Manca, E. (2008b). Immerse yourself in the traditions of the simply way of life': analysing English translations of Italian agriturismi websites. *Rivista Internazionale di Tecnica della Traduzione, 10*, 33-45.

Manca, E. (2013). Describing Through the Five Senses: A contrastive socio-cultural and linguistic analysis of Italian and British tourist websites In *Tourism and tourist promotion around the world: a linguistic and socio-cultural perspective* (pp. 109-124). Università del Salento.

Manca, E. (2016). *Persuasion in Tourism Discourse: Methodologies and Models*. Newcastle: Cambridge Scholars Publishing.

Manca, E. (2018). Verbal Techniques of the Language of Tourism Across Cultures: An Analysis of Five Official Tourist Websites. In M. Bielenia-Grajewska & E. Cortes de los Rios (Eds.), *Innovative Perspectives on Tourism Discourse* (pp. 91-110). Hershey PA: IGI Global.

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

Pagano, N. (2014). Tourism Destination Image and Irish Websites. *International Journal of Business and Social Science, 5*(8), 178-188.

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

Reppen, R. (2010). Building a corpus: What are the key considerations? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 31-37). Abingdon, New York: Routledge.

Scollon, R., & Scollon, S. W. (2003). *Discourses in place: Language in the material world*: Routledge.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails... In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 43-57). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Stoian, C. E. (2015). *The Discourse of Tourism and National Heritage: A Contrastive Study from a Cultural Perspective*. Newcastle: Cambridge Scholars Publishing.

Tang, W. M. (2013). AntConc Basics. Retrieved from https://corpora.files.wordpress.com/2008/01/antconc-basics5.pdf

TripAdvisor. (2017). About Tripadvisor. Retrieved from https://tripadvisor.mediaroom.com/us-about-us

Van Leeuwen, T. (2015). Multimodality. In D. Tannen, H. Hamilton, & S. Deborah (Eds.), *The Handbook of Discourse Analysis* (2nd ed., pp. 447-465). Malden, MA: John Wiley & Sons, Inc.

World Travel & Tourism Council. (2017). *World Travel and Tourism Council: Economic Impact 2017 - March 2017*. Retrieved from https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2017/world2017.pdf

Xiao, Z., & McEnery, A. (2005). Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics, 33*(1), 62-82.