“Two days we have passed with the ancients...”: a Digital Resource of Historical Travel Writings on Italy

Rachele Sprugnoli autore

Digital Humanities Group, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento - sprugnoli@fbk.eu

1. INTRODUCTION

Non-fictional travel writings are powerful sources of information for many research areas, such as art history, ethnography, geography and cultural history (Burke, 1997). By collecting several books on the same place, it is possible to study how material and cultural aspects change over time (Schulz, 2012). Moreover, travel writings can give insight not only into the places and people visited by the traveller but also on his/her culture and worldviews (Quintana, 2011). Despite the importance of this type of texts, digital collections of travel writings are not easy to retrieve. Texts are often not available in a format that can be straightforwardly processed by a computer, or there is no direct download of the collection (see for example, “A Vision of Britain through Time”¹), or the documents are scattered in vast digital archives.

In order to overcome these limitations, we release the first version of a corpus of more than 2 millions words of historical English travel writings about Italy which we have retrieved from freely available sources (Project Gutenberg² and Project Gutenberg Australia) and we make them available in a cleaned text format and in TEI-XML through the following website: <https://sites.google.com/view/travelwritingsonitaly>.

2. CORPUS DESCRIPTION

The current version of the corpus is made by 30 texts including both travel narratives (reports, diaries, collections of letters) and guidebooks, for a total of 2,413,838 tokens. The authors, of both genders, belong to different nationalities (UK, US, Australia). Some books are focused on specific cities or regions, others cover different parts of Italy. All the books were published between the end of the XIX Century and the beginning of the XX Century. We choose this period because at that time the tradition of the Grand Tour decayed and leisure-oriented tourists travels emerged (Ellerbee, 2010). This radical transformation was due to innovations occurred in transportation, economics and ideology: e.g., the diffusion of the railway network and of steam-powered ships, the increase in wealth for many Americans, a greater emancipation of women that led to having more female travelers (Schriber, 1995). In addition, Italian unification allowed the opening of new routes to the South of the country that were not taken into account during the Romantic era (Ouditt and Polezzi, 2012).

Chosen texts were processed with GutenTag³ so to automatically obtain their XML version based on the TEI standard⁴ and then manually checked and integrated using the Oxygen editor⁵. In this latter phase, particular effort has been devoted to enhance the description of each file adding missing biographical information on the author and adopting an unique standard for text classification. To this end, we took advantage of the Online Computer Library Center web service⁶ to obtain the Library of Congress Classification class of each book and a set of FAST subject headings to be used as keywords (Dean, 2004). As for text structure, the main additions involved the annotation of poems by using the *</g>* element and of the formulaic opening of letters with the *<opener>* element.

3. TOWARDS A GEOGRAPHICAL ANALYSIS OF TRAVEL WRITINGS

The plain text corpus has been processed with the Stanford CoreNLP (Manning et al., 2014) named entity recognizer to automatically detect place names and provide a preliminary analysis of the geographical dimension of travel writings. More than 4,600 locations were extracted and then georeferenced using Google's services. Table 1 shows the top 10 places with their frequency while Figure 1 displays the 200 most recurring locations.

¹ <http://www.visionofbritain.org.uk/travellers/>

² <http://www.gutenberg.org> - <http://gutenberg.net.au/>

³ <http://www.projectgumentag.org>

⁴ <http://www.tei-c.org/>

⁵ <https://www.oxygenxml.com/>

⁶ <http://classify.oclc.org/classify2/>

PLACE	FREQ
Rome	2287
Venice	1291
Florence	1136
Naples	769
Pisa	394
Genoa	343
Sicily	312
Siena	206
Vatican	174
Tuscany	172

Table 1. Top mentioned places.

Figure 1. Map of the 200 most mentioned locations.

In the map all regions are represented but travellers have focused their attention in particular on central Italy. As for the South, Figure 2 presents a zoom on the most frequent places in the Neapolitan area. Not much has changed with respect to the most popular contemporary tourist spots but it is worth noticing the use of ancient spelling of cities (“Puteoli”, “Misenum”, “Baiae”) related to the strong interest for the historical/archeological aspects of travels at that time. A fine-grained analysis is also possible: as shown in Figure 3, CoreNLP can detect the names of palaces, churches and squares providing a detailed view of historical tourist attractions.

Figure 2. Most recurring place names in the Neapolitan area.

- [8] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* , pp. 55-60.
- [9] Ouditt, Sharon, and Loredana Polezzi. 2012. Introduction: Italy as place and space. *Studies in Travel Writing* 16.2 (2012): 97-105.
- [10] Quintana, Ángel Gurriá. 2001. Travelling Through Discourse, Discoursing on Travel: Recent Writing on Travel Literature and British Travellers in Mexico. *Studies in Travel Writing* 5.1 (2001): 172-188.
- [11] Schriber, Mary Suzanne. 1995. Women's Place in Travel Texts. *Prospects*, (1995)20: 161-179.
- [12] Schulz, Janice. 2012. Travel Writings in the Archives and Rare Books Library. Online: <http://libapps.libraries.uc.edu/liblog/2012/04/travel-writings-in-the-archives-and-rare-books-library/>
- [13] Sprugnoli, Rachele, Tommaso Caselli, Sara Tonelli, and Giovanni Moretti. 2017a. The Content Types Dataset: a New Resource to Explore Semantic and Functional Characteristics of Texts. In *Proceedings of EACL 2017* .
- [14] Sprugnoli, Rachele , Sara Tonelli, Giovanni Moretti and Stefano Menini. 2017b. A little bit of bella pianura: Detecting Code-Mixing in Historical English Travel Writing. To appear in *Proceedings of CLiC-it 2017* .