

Conference information

JADT 2026

18th International Conference on Statistical Analysis of Textual Data

Palermo, Italy
8–10 July 2026

<https://jadt2026.unipa.it>

These proceedings collect the contributions presented at JADT 2026, the 18th International Conference on Statistical Analysis of Textual Data, held in Palermo, Italy, from 8 to 10 July 2026.

The conference brings together scholars working on statistical analysis of textual data, text mining, natural language processing, corpus linguistics, digital humanities, social sciences, economics, official statistics, health-related text analytics, and related methodological fields.

The proceedings are organized in two volumes according to the conference programme:

- **Volume I:** sessions of 8 July 2026.
- **Volume II:** sessions of 9–10 July 2026.

Printed ISBN: 978-88-5509-882-3

Online ISBN: 978-88-5509-883-0

Scientific Committee

- Ramón Álvarez Esteban, Universidad de León, Spain
- Valérie Beaudouin, EHESS / Telecom Paris, France
- Marie Chandelier, Université Côte d'Azur, France
- Francesca Della Ratta, Istituto nazionale per l'analisi delle politiche pubbliche, Italy
- Anne Dister, UCLouvain Saint-Louis Bruxelles, Belgium
- Cédric Fairon, UCLouvain, Belgium
- Bénédicte Garnier, Institut national d'études démographiques, France
- Francesca Greco, University of Udine, Italy
- Serge Heiden, ENS de Lyon / CNRS IHRIM, France
- Domenica Fioredistella Iezzi, University of Rome Tor Vergata, Italy
- Margareta Kastberg, Université de Franche-Comté, France
- Ludovic Lebart, CNRS / Telecom Paris, France
- Jean-Marc Leblanc, Université Paris-Est Créteil, France
- Alain Lelu, Université de Franche-Comté, France
- Dominique Longrée, Université de Liège, Belgium
- Véronique Magri, Université Côte d'Azur / CNRS BCL, France
- Pascal Marchand, Université de Toulouse, France
- Damon Mayaffre, Université Côte d'Azur / CNRS BCL, France
- Michelangelo Misuraca, University of Salerno, Italy
- Bénédicte Pincemin, ENS de Lyon / CNRS IHRIM, France
- Céline Poudat, Université Côte d'Azur / CNRS BCL, France
- Pierre Ratinaud, Université de Toulouse, France
- Maria Spano, University of Naples Federico II, Italy
- Arjuna Tuzzi, University of Padua, Italy
- Mathieu Valette, Institut national des langues et civilisations orientales, France
- Laurent Vanni, Université Côte d'Azur / CNRS BCL, France

Organizing Committee

University of Palermo, Italy

- Antonella Plaia, Chair
- Mariangela Sciandra, Chair
- Antonio Abbruzzo
- Giada Adelfio
- Alessandro Albano
- Luigi Augugliaro
- Giovanni Boscaino
- Marcello Chiodi
- Nicoletta D'Angelo
- Chiara Di Maria
- Vincenzo Giuseppe Genova
- Angelo Mineo
- Vito Michele Rosario Muggeo
- Roberto Pirrone
- Andrea Simonetti
- Irene Siragusa
- Gianluca Sottile
- Michele Tumminello
- Erasmo Vassallo

Contents

Conference information	ii
Scientific Committee	iii
Organizing Committee	iv
Introduction	v
Acknowledgements	vi
Programme at a Glance	viii
S04 – Specialized Session	
Advanced Text Analytics for Official Statistics: Methods, Models, and Applications	1
From textual agreements to structured data: using Large Language Models to enhance Official Statistics on anti-violence networks — <i>Elena Catanese, Mauro Bruno, Paolo Pizzo, Giuseppina Muratore and Claudia Villante</i>	2
Using AI techniques to support cause of death coding for Official Statistics — <i>Angela Pappagallo, Francesco Pugliese, Chiara Orsi, Tania Bracci, Francesco Grippo and Simone Navarra</i>	12
Semantic Search Approach to Occupation Coding in Official Statistics — <i>Giulio Massacci and Mauro Bruno</i>	22
Exploring User–Chatbot Communication: Methods, Insights, and Channel Differences — <i>Samanta Pietropaoli, Salvatore Agrillo, Paola Bosso, Barbara D’Amen and Gabriella Fazzi</i>	31
A Methodological Framework for Knowledge Graph Design Using Large Language Models — <i>Adriana Fidanza, Raffaella Maria Aracri, Alessio Buonocore and Roberta Radini</i>	40
C08 – Topic modeling, classification and interpretability	50
Editorial Transitions and Thematic Change in the American Journal of Sociology: A Structural Topic Modeling Analysis — <i>Stefano Sbalchiero, Valentina Rizzoli and Arjuna Tuzzi</i>	51
Titles, Abstracts, or Full-Texts? A Comparative Study of LDA and BERTopic Performance across Text Levels — <i>Francis Lareau and Christophe Malaterre</i>	60
A Multi-Method Validation Framework for Large-Scale Multilingual Text Analytics — <i>Stefano Blando and Domenica Fioredistella Iezzi</i>	70
Transforming UN Internal Audit Reports into Oversight Knowledge: Lexicometric Perspective — <i>Li Chen and Domenica Fioredistella Iezzi</i>	80

Comparing textometric approaches: Reinert-style Descending Hierarchical Classification versus probabilistic and neural topic modeling on a web-sourced corpus of French rap lyrics — <i>Robin Quillivic, Jules Dilé-Toustou and Pierre Ratinaud</i> . .	90
C17 – Translation, stylistics and literary variation	100
Demonstratives in human and machine translation between English and French — <i>Daniel Henkel</i>	101
Beyond Questions: A Corpus-Based Analysis of Interrogatives and Their Responses in Spoken French — <i>Yeo Jun Yun, Iris Eshkol-Taravella and Marianne Desmets</i> . .	111
The French Drama Revolution: Political Economy and Literary Production (1700-1900) — <i>Thiago Dumont Oliveira</i>	121
From Lyrics to Values: A Computational-Sociological Analysis of Cultural Change in Italian Popular Music — <i>Ciro Clemente De Falco and Emilia Romeo</i>	131
A Computational Authorship Attribution in the Children’s Tales of Oscar and Constance Wilde: The Case of “The Selfish Giant” — <i>Liviu P. Dinu, Alina Iacob and Cosmin Ciotloș</i> [not included in proceedings]	
S05 – Specialized Session	
Quantitative Text Analysis Methods for Scientific Literature	141
Mapping the Conceptual Landscape of Funded Computer Science Research in Spain — <i>Manuel Jesus Cobo Martín, Verónica Duarte Martínez, Luis M. Castillo-Chávez and Jose A. Moral-Munoz</i>	142
Modeling Emotion Constructs in Consumer Reviews: A Word2Vec-ELIta Hybrid Method Combined with PLS-SEM — <i>Cosima Valentino</i>	152
Mapping Communicative Purposes in Academic Discourse: designing an analytic linguistic pattern — <i>Alessandra Belfiore, Walter Giordano, Massimo Aria and Corrado Cuccurullo</i>	159
Enhancing leader-based community detection in scientific collaboration using publications metadata — <i>Geremia Sara and Domenico De Stefano</i>	168
Assessing BERTopic stability on a specialized scientific corpus: seed sensitivity, outlier reduction, and temporal semantic drift — <i>Carolina Vasconcelos, Sandro Mendonça and Bruno Damasio</i>	177
C16 – Andrea Camilleri and Sellerio corpora	185
Camilleri’s Sicilian: From Stopwords To Topic Modelling — <i>Vincenzo Giuseppe Genova and Andrea Sciandra</i>	186
Assunti teorico-metodologici su linguaggio letterario esclusivo e dati testuali — <i>Pietro Mazzarisi</i>	194
Classification of Andrea Camilleri’s works using bottom-up and top-down approaches. — <i>Andrea Sciandra and Michele A. Cortelazzo</i>	204
Distinctive Lexical Traits in Andrea Camilleri: A Keyness Measures Analysis of a Corpus of Sellerio Crime Novels — <i>Michele A. Cortelazzo, Matilde Trevisani and Arjuna Tuzzi</i>	213
A New Investigation for Sellerio — <i>Arjuna Tuzzi</i>	223
C11 – Science, technology and knowledge networks	233
Language as a Predictor of Scholarly Citation in European Policy Documents — <i>Francis Lareau, Jérémie Dion, Louis Renaud-Desjardins, Malvina Barra, Luisa Ripoll-Alberola and François Claveau</i>	234
Mapping the languages of science: a comparative cross-media textual analysis of scientific expertise. — <i>Gabriella Punziano, Rosanna Cataldo, Barbara Saracino and Ferdinando Iazzetta</i>	244

Discourse structure across technological maturity levels: a textometric approach — <i>Mélusine Caillard and Gaël Lejeune</i>	254
Structuring sectorial knowledge: automated detection of lexical relations — <i>Claudia Brunini, Eugenio Concetti and Giampiero D'Alessandro</i>	264
A hybrid rule-based and data-driven pipeline for author name normalization in IRIS institutional repositories — <i>Pasquale Pavone, Simone Righi and Sara Colombini</i>	274
Identifying Knowledge Gaps in Textual Networks: A Structural Holes Approach to Literature-Based Discovery — <i>Andrej Kastrin, Bojan Cestnik, Alja Nike Kastrin and Nada Lavrač</i>	284
S06 – Specialized Session	
Text Analysis in Economics	294
From Fed Speeches to Signals Embedding Central Bank Communication in Macroeconomic Forecasting — <i>Angelo Fama</i> ’	295
AI Meets Fiscal Policy: Mapping Government Spending Actions Across 64 Countries — <i>Davide Furceri</i> [not included in proceedings]	
Global and local effects of Federal Reserve communication: uncertainty and real activity — <i>Antonio Pietro Maria Morreale</i> [not included in proceedings]	
BILATERAL CONFLICT RISK AND TRADE: MILITARY WARS, TRADE WARS, AND DIPLOMATIC NOISE — <i>Jamel Saadaoui</i> [not included in proceedings]	
C03 – Public, political and media discourse	305
Détection des isotopies du discours présidentiel d’Emmanuel Macron : de la cooccurrence à la self-attention — <i>Damon Mayaffre and Laurent Vanni</i>	306
Les formes du futur en planification urbaine. Analyse textométrique des temps verbaux en planification régionale (Île-de-France, 1919-2024) — <i>Isabelle Chesneau</i>	316
Investigation de l’évolution des discours médiatiques à propos des réseaux socio-numériques à partir de la distance de Labbé et de sa version améliorée — <i>Pierre Ratinaud and Brigitte Sebbah</i>	326
Caractériser les commentaires en ligne : les émojis dans la loi « Duplomb » — <i>Pascal Marchand and Pierre Ratinaud</i>	336
Étudier les vidéos TikTok : comparaison lexicométrique des plus-values et des limites des retranscriptions automatiques versus les descriptions des vidéos. — <i>Panos Tsimpoukis, Lucie Loubere, Nikos Smyrnaïos, Pierre Ratinaud and Charlis Papaeangelou</i>	346
C02 – French NLP, lexicon and contrastive neology	354
Vers une correction automatique des formes inconnues : application du Masked Language Modeling de CamemBERT au corpus des Cahiers citoyens — <i>Sami Guembour, Solomiia Korol and Catherine Dominguès</i>	355
Séjour à l’étranger et développement des marqueurs discursifs en FLE : étude sur corpus chez les apprenants japonophones — <i>Kaori Sugiyama</i>	365
Le potentiel lexicographique des néologismes des TIC dans une perspective franco-slovaque : analyse par le biais de Sketch Engine — <i>Lucia Rackova</i>	375
Vers un chatbot basé sur la recherche sémantique et le raisonnement à base de règles pour l’optimisation des opérations de la Sécurité Sociale française — <i>Wenxiang Zhou, Touria Aït El Mekki, Frédéric Lardeux, Marc Legeay and Jean-Olivier Boesmier</i>	385
L’extraction semi-automatique de termes métaphoriques grâce à la spécificité sémantique : étude appliquée à un corpus français-italien de presse spécialisée dans le domaine de la bijouterie-joaillerie — <i>Marius François and Maria Grazia Massimo</i>	395

Utilisation des emprunts à l'anglais de type V-ING en espagnol, français et italien : une étude en diachronie courte basée sur les corpus du web du Wortschatz de Leipzig — <i>François Maniez, Maria Belén Villar Diaz and Sidonie Larato</i>	405
C01 – Textometry, corpus methods and visualization	415
Articuler méthodes quantitatives et épistémologies critiques en linguistique : enjeux, tensions et perspectives méthodologiques — <i>Tiago Joseph</i>	416
A la croisée des disciplines informatique et psychanalytique : une application sémio- tique de la méthode Reinert — <i>Claire Tissot</i>	426
De la mesure des paragraphes à la classification des visualisations : portée heuristique et démonstrative — <i>Claude Patricia Tardif</i>	436
De l'analyse de données en cartes : Visualisations géographiques et chronologiques de corpus textuels — <i>Jean-Marc Leblanc, Anne-Lise Humain-Lamoure and Marie Pérès-Leblanc</i>	446
Etude des cooccurrences du verbe Dire dans un corpus d'entretiens à propos d'un événement traumatique — <i>Anna Colli, Delphine Battistelli and Marie Chagnoux</i>	457
Europarser: exploiter de larges corpus de presse — <i>Marceau Hernandez, Gaël Lejeune, Félix Alié, Émile Rebours and Julien Bezançon</i>	467
C05 – Italian institutional and cultural corpora	477
La sfida del lavoro da remoto. Analisi di focus group a responsabili delle risorse umane — <i>Francesca della Ratta</i>	478
Un'analisi testuale dei discorsi parlamentari italiani — <i>Matteo Farnè and Marco Nannetti</i>	488
A Multilevel Analysis of Italy's Three-Year School Planning Documents Using a Lexical and Distributional-Semantic Framework — <i>Pasquale Pavone, Maria Teresa Sagri, Elettra Morini and Eniko Tolvay</i>	496
The Lexical Construction of Women's 'Image': Party Specificity and Gender Asym- metry in Italy's Constituent Assembly — <i>Francesca Aulino, Rosario D'Agata, Domenico De Stefano and Amin Gino Fabbrucci Barbagli</i>	505
Perceptions of Emilia-Romagna: a cross-linguistic analysis of keywords in English and French travel literature — <i>Rita Gramellini and Daniel Henkel</i>	515
C13 – Health, clinical and accessibility text analytics	525
Living with osteoporosis across contexts: a lexicometric comparison of Italian and Spanish narratives — <i>Chiara Tedesco, Ercole Vellone, Rosaria Alvaro, Gianluca Pucciarelli, Vicente Bernalte-Martí and Mariachiara Figura</i>	526
Self-care behaviours in people with osteoporosis: a multidimensional analysis of textual data — <i>Chiara Tedesco, Ercole Vellone, Rosaria Alvaro, Gianluca Pucciarelli, Vicente Bernalte-Martí and Mariachiara Figura</i>	536
Text Mining Applications in the Evaluation of AIFA Drug Innovativeness Reports in Italy — <i>Daniele Pirotta, Andrea Marletta, Carlotta Galeone and Paolo Mariani</i>	546
Characterization of texts written by deaf people through syntactic markers — <i>Mireille Esther Gettler Summa, Caroline Bogliotti and Anne Lacheret Dujour</i>	554
S07 – Specialized Session	
Finance and Textual Analysis	564
Interpretable Natural Language Concept Bottleneck Models for Default Risk estima- tion — <i>Andrea Conti and Giacomo Morelli</i> [not included in proceedings]	
Web Automatic Investigator (WAI): an Agentic Framework for Regulatory Multimodal Analysis of Web Contents — <i>Andrea Coletta, Miriam Corrado, Gabriele De Capoa, Oliver Giudice, Alessandro Maggi, Marco Mori, Anna Russo Russo and Valerio Signore</i>	565

Word2Prices: embedding central bank communications for inflation prediction — <i>Douglas Araujo, Nikola Bokan, Fabio Alberto Comazzi and Michele Lenza</i> [not included in proceedings]	
The cyber risk of non-financial firms — <i>Francesco Columba, Manuel Cugliari, Marco Orlandi and Federica Vassalli</i> [not included in proceedings]	
Processing instability and the value of information: from individual noise to common risk — <i>Rudraksh Kansal and Taejin Park</i> [not included in proceedings]	
C06 – Digital tools, annotation and corpus infrastructures	575
A Text analysis application: a library-led introductory course open to all University users — <i>Deborah Grbac and Valentina Schiariti</i>	576
JADT Anthology: decades of textual data analysis in one corpus. — <i>Laurent Vanni</i>	586
ActiveTigger: An open source collaborative text annotation software for computational social sciences — <i>Emilien Schultz, Julien Boelaert, Axel Morin, Emma Bonutti D’Agostini, Annina Claesson, Étienne Ollion and Arnault Chatelain</i>	596
ParaTaxe: a Suite of Tools for Processing and Publishing Multi-parallel Corpora — <i>Olivier Kraif, Cyrille François, Elnaz Jalilian, Natacha Rimasson Fertin and Pascale Roux</i>	606
Analyse de textualités multiples avec TXM : du texte brut aux éditions TEI, transcriptions synchronisées, traductions alignées et micro-textes — <i>Serge Heiden</i>	614
C07 – Statistical methods and pattern extraction	624
The UMCC: Usage and Measurement of Connectors in Context. A methodology and interface for measuring the use of connectors in different genres of corpora — <i>Corinne Rossari, Claudia Ricci, Linda Sanvido, Jonathan Schaber and Gerold Schneider</i>	625
The Motif Work Bench Pipeline: From Sequential Pattern Mining to Multidimensional Statistics — <i>Hugo Dumoulin and Timothée Premat</i>	635
The Moving Border Model: Applying Chronological Specificities to the Tuning of a Time Boundary — <i>Bénédicte Pincemin, Céline Guillot-Barbance and Alexei Lavrentiev</i>	645
Validation of Simultaneous Additive Trees in Textual Data Visualization — <i>Ludovic Lebart</i>	655
Pulling strings: scalable symbolic methods for Multi-Word Units detection in text corpora — <i>Antonio Balvet</i>	665
Estimating the typical-source distribution in imbalanced corpora — <i>Aris Xanthos</i>	675

C05

Italian institutional and cultural corpora

09-lug
17:00 - 18:30
Aula 12

Perceptions of Emilia-Romagna: a cross-linguistic analysis of keywords in English and French travel literature¹

Rita Gramellini¹, Daniel Henkel²

¹University of Bologna – rita.gramellini2@unibo.it

²University of Paris 8 Vincennes-Saint-Denis – daniel.henkel@univ-paris8.fr

Abstract

This paper reports a cross-linguistic keyword analysis of travel literature in English and French describing the Italian region of Emilia-Romagna. Using a multi-corpus design, this study identifies the lexical items that distinguish representations of Emilia-Romagna from those of other Italian regions and examines convergences and divergences across languages. Keywords were extracted using the ratio of relative frequencies, log-likelihood and a dispersion criterion, and were ranked by frequency to identify the most specific and commonly used items. Qualitative categorization was then undertaken based on the analysis of their distributional profiles. Results show a broadly convergent cross-linguistic representation of the region defined by its artistic heritage, religious traditions, and historical figures, with some differences in lexical salience. Comparison with travel writing on other Italian regions confirms Emilia-Romagna's distinctive profile and cultural influence, particularly with respect to its pictorial tradition.

Keywords: keyword analysis, corpus linguistics, comparable corpora, travel literature, Italy.

Résumé

Cet article présente une analyse contrastive de mots-clés dans les récits de voyage en anglais et en français consacrés à la région italienne d'Émilie-Romagne. S'appuyant sur un modèle multi-corpus, cette étude identifie les unités lexicales qui distinguent les représentations de l'Émilie-Romagne de celles d'autres régions italiennes et examine les convergences et divergences entre les deux langues. Les mots-clés ont été extraits sur la base du ratio des fréquences relatives, du calcul log-likelihood (fonction log-vraisemblance), et d'un critère de dispersion, et ont été classés par fréquence afin d'identifier les mots-clés les plus spécifiques et les plus utilisés. Une catégorisation qualitative a ensuite été effectuée à travers l'analyse de leurs profils distributionnels. Les résultats révèlent une représentation interlinguistique largement convergente de la région, centrée sur son patrimoine artistique, ses traditions religieuses et ses figures historiques, avec quelques variations de saillance lexicale. La comparaison avec les récits de voyage consacrés à d'autres régions italiennes confirme le profil distinctif de l'Émilie-Romagne et son influence culturelle, notamment en ce qui concerne sa tradition picturale.

Mots clés : analyse de mots-clés, linguistique de corpus, corpus comparable, littérature de voyage, Italie.

1. Introduction

Since the mid-1990s, keyword analysis has become a central method in corpus linguistics for identifying thematic content and stylistic distinctiveness. In statistical terms, keywords are defined as unusually frequent lexical items in a target text or corpus by comparison with a reference corpus (Scott, 1997; Baker, 2004; Scott and Tribble, 2006 *inter alia*), typically

¹ Rita Gramellini was responsible for writing the manuscript. Daniel Henkel contributed to data collection and analysis and revised the manuscript.

extracted using significance tests such as log-likelihood (LL) (Dunning, 1993). The resulting statistic, associated with a p -value, indicates the reliability of differences in word relative frequencies, though it does not convey information about its magnitude (Gabrielatos, 2018). Effect-size measures like the ratio of relative frequencies (RRF) (Damerau, 1993) are therefore recommended alongside significance tests to provide a more meaningful assessment of the degree to which a lexical item is associated with a given text or corpus (Gries, 2010; Gabrielatos, 2018; Pojanapunya and Watson Todd, 2018).

Frequency-based keyword lists, however, can be skewed by items that are statistically key yet appear in only a small number of texts, making them unrepresentative of the corpus as a whole. To address this, dispersion criteria are commonly applied to ensure that selected keywords occur across multiple texts rather than being concentrated in a limited subset (or single text), so that findings reflect generalizable trends rather than idiosyncratic authorial choices (Baker, 2004; Scott and Tribble, 2006; Egbert and Biber, 2019; Gries, 2021).

Building on this methodological framework, the present study investigates the lexical representation of the Italian region of Emilia-Romagna in travel literature. While keyword-based approaches have proven effective for analyzing the lexical and stylistic features of this genre (see Gerbig, 2010), few corpus-based studies have systematically examined how Italy has been represented across languages. Even less is known about individual regions, despite the strongly localized nature of Italy's cultural, historical and artistic identities. To address this gap, this paper reports a cross-linguistic keyword analysis of English and French travel literature describing Emilia-Romagna. A multi-corpus design is adopted to distinguish region-specific lexical features from broader genre patterns. For each language, a target corpus on Emilia-Romagna is compared with a contrastive corpus covering other Italian regions, both of which are evaluated with respect to a reference corpus of general travel literature. The analysis combines quantitative keyword extraction with a qualitative examination of collocations and concordances to identify shared patterns across languages and corpora.

The study is guided by the following research questions:

- Which lexical items characterize representations of Emilia-Romagna in English and French travel literature, as determined by statistically significant disparities in frequency?
- To what extent do these keywords converge across the two languages?
- How does the lexical representation of Emilia-Romagna differ from that of other Italian regions?

By focusing on a single region and integrating a cross-linguistic perspective, multi-corpus design and combined quantitative-qualitative approach, the study offers an innovative contribution to corpus-based travel literature research and sheds light on the interaction between language, place, and cultural representation in historical travel discourse.

The remainder of the paper is structured as follows: Section 2 outlines the methodology, including the corpus design, keyword extraction procedures and qualitative categorization. Section 3 presents the results, and Section 4 discusses the findings.

2. Methodology

2.1. Corpus design

The analysis is based on comparable corpora of travel literature in English and French. For each language, three corpora were compiled: a corpus of interest focusing on Emilia-Romagna

(§2.1.1.), a corpus describing other Italian regions for purposes of comparison (§2.1.2.), and a reference corpus representative of the overall genre (§2.1.3.).

To ensure reliable results, all texts underwent lemmatization and part-of-speech tagging using *TreeTagger* (Schmid, 1994). Moreover, spelling variations and archaic forms typical of historical corpora were normalized prior to analysis.

2.1.1. BER-TL corpus

The Bologna-Emilia-Romagna-Travel-Literature (BER-TL) corpus includes travel writings on the Italian region of Emilia-Romagna spanning the 17th to the 20th centuries, with one 21st-century French text. It builds on the multilingual UniVOCItà corpus² compiled for heritage tourism purposes by the LBC-CeSLiC unit at the University of Bologna, comprising English, French, Spanish, and Russian literary testimonies on the area of Bologna (Zotti, 2026). The English and French subcorpora were subsequently restricted to travel writing and expanded to cover the broader Emilia-Romagna region.

2.1.2. ITA-TL corpus

The Italy-Travel-Literature (ITA-TL) corpus comprises travel writings on Italy outside Emilia-Romagna. The English subcorpus draws on a freely available historical corpus of 30 non-fictional travel narratives³ (Sprugnoli, 2018). Since no pre-existing equivalent French subcorpus was available, a comparable dataset was compiled collecting text from open-access repositories such as the Project Gutenberg⁴. To avoid overlap with the BER-TL corpus, all chapters or sections describing Emilia-Romagna were identified and removed.

2.1.3. REF-TL corpus

The Reference-Travel-Literature (REF-TL) corpus is representative of the travel writing discourse type and serves as a common baseline for both BER-TL and ITA-TL in keyword extraction. It consists of English and French travel writings describing destinations worldwide, compiled using the *Project Gutenberg Corpus Builder* (Moncomble, 2025) through searches constrained by genre (“Travel Writing”) and language (“English” or “French”). The resulting texts were reduced to a random sample of 100 documents per language, ensuring that all were originally written in the target language and published between the 17th and 21st centuries.

Table 1 provides an overview of all comparable corpora described above, reporting the number of files, token count, Type/Token Ratio (TTR), Standardized Type/Token Ratio (STTR) and time frame⁵.

Table 1. Metadata of the comparable corpora

Corpus	Lang.	Files	Tokens	TTR	STTR	Time frame (median)
BER-TL	English	175	1,354,787	0.034	0.468	17 th -20 th c. (1847)
	French	147	1,226,309	0.043	0.472	17 th -21 st c. (1857)
ITA-TL	English	30	2,395,623	0.023	0.464	19 th -20 th c. (1909)

² <https://univocitta.github.io/UniVOCItta/#/corpus>

³ <https://sites.google.com/view/travelwritingsonitaly/corpus>

⁴ <https://www.gutenberg.org/>

⁵ Tokens are defined as sequences of alphanumeric characters. STTR = mean TTR over 1,000-token windows.

REF-TL	French	27	1,709,854	0.036	0.500	18 th -20 th c. (1866)
	English	100	8,047,253	0.013	0.458	18 th -20 th c. (1892)
	French	100	5,712,416	0.019	0.472	18 th -20 th c. (1887)

2.2. Keyword extraction and categorization

Keywords were extracted from the BER-TL and ITA-TL corpora using *AntConc* 3.5.9 (Anthony, 2020), with the REF-TL corpus serving as the reference corpus to filter out lexical units typical of travel writing and obtain keyword lists primarily capturing regional distinctiveness rather than genre-specific lexical patterns⁶.

The extraction procedure combined LL ($p < 0.001$), as a significance filter retaining only items whose frequency differences are unlikely to be due to chance, and RRF (≥ 10), which provides a more intuitive and interpretable measure of the overrepresentation magnitude, ensuring that extracted items are statistically significant and substantially overrepresented. To improve the results' generalizability, a dispersion criterion was applied. Since no consensus exists in the literature regarding the optimal dispersion cut-off, a minimum of 5% of the texts was adopted here as a practical, albeit necessarily arbitrary, threshold, while remaining consistent with dispersion criteria commonly used in corpus-based studies (Pojanapunya and Watson Todd, 2021). Finally, as these measures do not in themselves capture the relative prominence of a keyword within the corpus, insofar as certain technical terms may be highly specific yet relatively uncommon, frequency was used as a final criterion for prioritizing the “core vocabulary” (Henkel, 2020), namely keywords that are both highly specific and widely used in the discourse type under investigation.

As Table 2 shows, the $RRF \geq 10$ threshold eliminated between 41 and 51% of the initial keywords, while the subsequent dispersion criterion further reduced them by approximately 20-34%, suggesting that a non-negligible proportion of statistically significant items were concentrated in a small number of texts and thus unevenly distributed across the corpus.

Table 2. Number of keywords retained at each extraction step for each subcorpus

Subcorpus	LL	LL + RRF ≥ 10	LL+RRF ≥ 10 +dispersion $\geq 5\%$
English BER-TL	1,488	871	614
French BER-TL	1,176	621	410
English ITA-TL	1,780	959	771
French ITA-TL	1,203	592	418

Following quantitative extraction, the top 50 keywords were qualitatively categorized through a close examination of their collocational profiles and concordance patterns. Collocates were identified within a span of ± 3 words left and right, with a minimum frequency of 3 occurrences (range ≥ 2 texts), using LL ($p < 0.05$) and mutual information (MI) as association measures. Thematic categories were assigned based on concordance analysis, in line with the Firthian principle that meaning is determined by context (Firth, 1957: 11). This applies also to items that may function as components of larger n-grams rather than as independent units, such as Italian function words (e.g. *di*, *della*). In cases where a keyword demonstrated multiple distinct uses across the corpus⁷, the item was assigned to all relevant categories.

⁶ Candidate keywords include all lemmas attested in either corpus, with zero-frequency items not excluded a priori.

⁷ For instance, the French keyword *parmesan* refers both to the painter Parmigianino and to the cheese.

3. Results

Table 3 presents the top 50 keywords extracted from each BER-TL and ITA-TL subcorpus against the REF-TL corpus, ranked by normalized frequency/10k words (f), to prioritize widely used highly specific items, and followed by their RRF values (r) indicating the degree of overrepresentation. While only the top 50 keyword are reported here for reasons of space, the discussion draws on the full ranked lists for a more comprehensive interpretation of the results.

Table 3. Top 50 keywords for each subcorpus

	English BER-TL	French BER-TL	English ITA-TL	French ITA-TL
1	<i>bologna</i> [f:17.9, r:349]	<i>bologne</i> [f:19.7, r:257.5]	<i>rome</i> [f:10.4, r:11]	<i>rome</i> [f:8.2, r:12.3]
2	<i>italy</i> [f:13, r:20.7]	<i>tableau</i> [f:14.4, r:10.9]	<i>italian</i> [f:9.2, r:11.7]	<i>italie</i> [f:8.2, r:11.3]
3	<i>ravenna</i> [f:9, r:225.7]	<i>italie</i> [f:10, r:17.9]	<i>italy</i> [f:8.1, r:12.9]	<i>corinne</i> [f:7.2, r:4044.5]
4	<i>painting</i> [f:7.5, r:15.1]	<i>pape</i> [f:8.3, r:15.4]	<i>venice</i> [f:6.5, r:23.6]	<i>di</i> [f:7.1, r:22.2]
5	<i>ferrara</i> [f:6.2, r:709.3]	<i>rome</i> [f:6.9, r:11.3]	<i>florence</i> [f:5.1, r:16.2]	<i>lord</i> [f:3.3, r:34.6]
6	<i>pope</i> [f:6.1, r:23.8]	<i>parme</i> [f:6.8, r:240]	<i>di</i> [f:4.3, r:19.6]	<i>che</i> [f:3.1, r:61.8]
7	<i>virgin</i> [f:5.9, r:14.9]	<i>ferrare</i> [f:6.5, r:394.9]	<i>piazza</i> [f:4.1, r:21.7]	<i>dante</i> [f:2.7, r:22.5]
8	<i>parma</i> [f:5.8, r:354.6]	<i>carrache</i> [f:6.4, r:875.9]	<i>venetian</i> [f:4, r:28.7]	<i>questo</i> [f:1.9, r:156.7]
9	<i>caracci</i> [f:5.3, r:2124.9]	<i>ravenne</i> [f:5.9, r:697.1]	<i>naples</i> [f:3.7, r:10.2]	<i>per</i> [f:1.8, r:25.8]
10	<i>venice</i> [f:5.3, r:19.4]	<i>autel</i> [f:5.8, r:10.2]	<i>pope</i> [f:3.5, r:13.3]	<i>napolitain</i> [f:1.6, r:13.1]
11	<i>madonna</i> [f:5.2, r:59.4]	<i>peintre</i> [f:5.7, r:11.2]	<i>madonna</i> [f:3, r:34.7]	<i>come</i> [f:1.5, r: 46.3]
12	<i>di</i> [f:5.2, r:23.6]	<i>modène</i> [f:5.7, r:5050.2]	<i>fresco</i> [f:2.6, r:18.6]	<i>pise</i> [f:1.4, r:21.2]
13	<i>guido</i> [f:4.9, r:216.4]	<i>bolonais</i> [f:5.4, r:2155.8]	<i>via</i> [f:2.5, r:12.3]	<i>una</i> [f:1.4, r:44.9]
14	<i>florence</i> [f:4.7, r:14.8]	<i>cardinal</i> [f:5.2, r:13.7]	<i>etruscan</i> [f:2.2, r:61.4]	<i>al</i> [f:1.4, r:14.7]
15	<i>fresco</i> [f:4.6, r:33]	<i>venise</i> [f:5, r:11.2]	<i>palazzo</i> [f:2.2, r:17.5]	<i>montaigne</i> [f:1.2, r:39.1]
16	<i>rimini</i> [f:4.2, r:481.3]	<i>di</i> [f:4.6, r:11.7]	<i>della</i> [f:2.1, r:35.5]	<i>sicilien</i> [f:1.2, r:11.1]
17	<i>modena</i> [f:3.8, r:340.8]	<i>guerchin</i> [f:4.2, r:612.4]	<i>maria</i> [f:2.1, r:12.1]	<i>essere</i> [f:1.1, r:42.7]
18	<i>correggio</i> [f:3.7, r:600.5]	<i>rimini</i> [f:4.1, r:812.5]	<i>giovanni</i> [f:2, r:31.5]	<i>più</i> [f:1, r:65.4]
19	<i>angel</i> [f:3.7, r:10.3]	<i>corrège</i> [f:4.1, r:173.9]	<i>medici</i> [f:2, r:98]	<i>altro</i> [f:1, r:80.2]
20	<i>mosaic</i> [f:3.3, r:21.3]	<i>plaisance</i> [f:3.7, r:22.3]	<i>lorenzo</i> [f:2, r:59.9]	<i>vésuve</i> [f:1, r:15]
21	<i>milan</i> [f:3.3, r:30]	<i>madone</i> [f:3.6, r:15.5]	<i>pisa</i> [f:1.8, r:23]	<i>messin</i> [f:1, r:12.9]
22	<i>raphael</i> [f:3.1, r:38.7]	<i>raphaël</i> [f:3.4, r:16.9]	<i>cardinal</i> [f:1.7, r:11.4]	<i>catane</i> [f:0.9, r:20.5]
23	<i>piazza</i> [f:2.9, r:15.7]	<i>catherine</i> [f:3.2, r:16.4]	<i>genoa</i> [f:1.6, r:17.9]	<i>con</i> [f:0.9, r:24.2]
24	<i>po</i> [f:2.9, r:96.3]	<i>pétrone</i> [f:3.2, r:2539]	<i>cosimo</i> [f:1.4, r:369.7]	<i>syracuse</i> [f:0.9, r:11.8]
25	<i>cardinal</i> [f:2.8, r:19.3]	<i>arioste</i> [f:3.1, r:141.1]	<i>capri</i> [f:1.3, r:45.6]	<i>lady</i> [f:0.8, r:21.6]
26	<i>piacenza</i> [f:2.8, r:742.8]	<i>césar</i> [f:3.1, r:10.1]	<i>façade</i> [f:1.3, r:19.4]	<i>padoue</i> [f:0.8, r:16.3]
27	<i>guercino</i> [f:2.8, r:554.1]	<i>pó</i> [f:3.1, r:34.7]	<i>duomo</i> [f:1.3, r:46.6]	<i>arno</i> [f:0.7, r:18.4]
28	<i>palazzo</i> [f:2.7, r:21.8]	<i>apennin</i> [f:3, r:42.7]	<i>doge</i> [f:1.3, r:23.6]	<i>léonard</i> [f:0.7, r:11.6]
29	<i>bolognese</i> [f:2.7, r:715.2]	<i>dante</i> [f:3, r:17.6]	<i>florentine</i> [f:1.3, r:34.5]	<i>sienne</i> [f:0.7, r:11.1]
30	<i>da</i> [f:2.5, r:16.9]	<i>jérôme</i> [f:2.8, r:23.5]	<i>donatello</i> [f:1.2, r:497.9]	<i>tutto</i> [f:0.7, r:30.8]
31	<i>dante</i> [f:2.3, r:22]	<i>académie</i> [f:2.5, r:10.2]	<i>renaissance</i> [f:1.2, r:15.1]	<i>quale</i> [f:0.6, r:183.4]

32	<i>apennine</i> [f:2.3, r:55.2]	<i>dominique</i> [f:2.5, r:17.8]	<i>micelangelo</i> [f:1.2, r:922.5]	<i>gondole</i> [f:0.6, r:10.8]
33	<i>ariosto</i> [f:2.2, r:889.6]	<i>annibal</i> [f:2.5, r:84.5]	<i>gondola</i> [f:1.2, r:12.6]	<i>bianca</i> [f:0.6, r:17.2]
34	<i>tasso</i> [f:2.2, r:123.3]	<i>antoine</i> [f:2.4, r:10.5]	<i>andrea</i> [f:1.1, r:48.2]	<i>pisan</i> [f:0.6, r:22.3]
35	<i>il</i> [f:2.1, r:13.1]	<i>albane</i> [f:2.3, r:99.2]	<i>titian</i> [f:1.1, r:28.3]	<i>home</i> [f:0.6, r:10.5]
36	<i>padua</i> [f:2.1, r:65.2]	<i>faenza</i> [f:2.3, r:195.7]	<i>neapolitan</i> [f:1.1, r:13.2]	<i>agata</i> [f:0.6, r:667.5]
37	<i>arcade</i> [f:2.1, r:18.5]	<i>forli</i> [f:2.3, r:927.9]	<i>porta</i> [f:1.1, r:14.3]	<i>milanais</i> [f:0.6, r:55.6]
38	<i>francesco</i> [f:2, r:108.8]	<i>légat</i> [f:2.3, r:228.5]	<i>forum</i> [f:1, r:13]	<i>tibre</i> [f:0.6, r:20.7]
39	<i>della</i> [f:2, r:33.7]	<i>alphonse</i> [f:2.2, r:30.8]	<i>giotto</i> [f:1, r:82.3]	<i>véronne</i> [f:0.6, r:19.4]
40	<i>giovanni</i> [f:2, r:30.7]	<i>imola</i> [f:2.2, r:1698.8]	<i>campagna</i> [f:1, r:20.1]	<i>vatican</i> [f:0.6, r:10.7]
41	<i>venetian</i> [f:1.9, r:13.8]	<i>augustin</i> [f:2.1, r:15.7]	<i>siren</i> [f:1, r:136.5]	<i>era</i> [f:0.6, r:161.9]
42	<i>maria</i> [f:1.9, r:11.3]	<i>parmesan</i> [f:2.1, r:273.9]	<i>siena</i> [f:1, r:54]	<i>francesco</i> [f:0.6, r:13.5]
43	<i>adriatic</i> [f:1.8, r:43.9]	<i>farnèse</i> [f:2, r:136.2]	<i>gondolier</i> [f:1, r:30.5]	<i>beaux</i> [f:0.6, r:10.8]
44	<i>petronio</i> [f:1.8, r:1442.2]	<i>reggio</i> [f:2, r:260.1]	<i>vatican</i> [f:1, r:29.3]	<i>arts</i> [f:0.6, r:11.4]
45	<i>theodoric</i> [f:1.8, r:75.6]	<i>ancône</i> [f:2, r:110.1]	<i>fra</i> [f:0.9, r:48]	<i>antonia</i> [f:0.6, r:26.4]
46	<i>domenico</i> [f:1.8, r:109.1]	<i>benoît</i> [f:2, r:27.5]	<i>angela</i> [f:0.9, r:716]	<i>matteo</i> [f:0.5, r:38.4]
47	<i>este</i> [f:1.8, r:176.6]	<i>basilique</i> [f:1.9, r:15.7]	<i>perugia</i> [f:0.9, r:39.8]	<i>dell</i> [f:0.5, r:23.6]
48	<i>saviour</i> [f:1.7, r:11.5]	<i>padoue</i> [f:1.9, r:27.7]	<i>pompeii</i> [f:0.9, r:14.3]	<i>corso</i> [f:0.5, r:20.5]
49	<i>papal</i> [f:1.6, r:32.4]	<i>che</i> [f:1.8, r:26.7]	<i>francesco</i> [f:0.9, r:47.5]	<i>io</i> [f:0.5, r:50.7]
50	<i>cupola</i> [f:1.6, r:19.5]	<i>romagne</i> [f:1.8, r:1431]	<i>campanile</i> [f:0.9, r:24.2]	<i>vasari</i> [f:0.5, r:11.6]

4. Discussion

Keywords of the English and French BER-TL subcorpora show a marked degree of cross-linguistic convergence in the lexical representation of Emilia-Romagna. Many items have direct correspondences or close semantic equivalents in the other language, whether among the top-ranking keywords or lower in the lists. Moreover, non-shared keywords tend cluster around the same domains. While the discussion below focuses primarily on the top 50 items, reference to the full ranked lists is necessary to capture the subcorpora's lexical specificity.

Geographical references form the most salient shared domain. The highest-ranking keywords include major cities such as *Bologna* (f:17.9, r:349)/*Bologne* (f:19.7, r:257.5), *Ravenna* (f:9, r:225.7)/*Ravenne* (f:5.9, r:697.1), and *Ferrara* (f:6.2, r:709.3)/*Ferrare* (f:6.5, r:394.9), alongside natural features, namely the *Po* (f:2.9, r:96.3)/*Pô* (f:3.1, r:34.7) and the *Apennine* (f:2.3, r:55.2)/*Apennin* (f:3, r:42.7). Together, these items suggest that Emilia-Romagna is represented in both traditions as a region defined by historically prominent urban centers embedded within a recognizable landscape.

Moreover, the occurrence of extra-regional cities such as *Venice* (f:5.3, r:19.4)/*Venise* (f:5, r:11.2) and *Padua* (f:2.1, r:65.2)/*Padoue* (f:1.9, r:27.7) indicate that the region is frequently framed within the context of broader Italian itineraries and through inter-city comparisons, a pattern further supported by collocation and concordance evidence⁸. By the same logic, it is not surprising that the ITA-TL subcorpora, dominated by Italy's cultural capitals like *Rome* (En:

⁸ “The railroad from Venice to Bologna passes through Ferrara”; “[In] Modena (...) the houses are built of brick, and most of them have portici, as at Padua and Bologna”.

f:10.4, r:11; Fr: f:8.2, r:12.3), *Pisa* (f:1.8, r:23)/*Pise* (f:1.4, r:21.2), *Venice* (En: f:6.5, r:23.6), identity markers (e.g. *florentine*, *venetian*, *neapolitan/napolitain*) and distinctive local features (e.g. *gondola*, *gondolier*, and *doge* in relation to Venice), still contain references to some Emilian and Romagnol cities, albeit lower in the ranking and with considerably reduced salience. Examples include *Ferrara* (f:0.4, r:46.2)/*Ferrare* (f:0.3, r:15.8), *Rimini* (En f:0.1, r:16.7, Fr: f:0.2, r:43), and *Modène* (f:0.2, r:178.4).

The French BER-TL subcorpus seems to lexicalize the region at a finer level of geographical coverage. In fact, in addition to the major shared toponyms, the French top 50 includes *Faenza* (f:2.3, r:195.7), *Forlì* (f:2.3, r:927.9), *Imola* (f:2.2, r:1698.8), and *Reggio* (f:2, r:260.1). It must be noted, however, that all four also occur within the English BER-TL top 100. Particularly notable is the keyword *Romagne* (f:1.8, r:1431) in the French BER-TL top 50, whose high RRF and collocational profile (*duché*, *capitale*, *pays*) suggest an explicit awareness of Romagna as a distinct historical and political entity. English shows a comparable, though less salient, pattern for *Romagna* (f:0.7, r:1064).

Artistic heritage constitutes a second major area of convergence. In both BER-TL subcorpora, Emilia-Romagna emerges as a region strongly associated with the visual arts, and painting in particular. Artists linked to the Emilian pictorial tradition, such as *Caracci* (f:5.3, r:2124.9)/*Carrache* (f:6.4, r:875.9), *Correggio* (f:3.7, r:600.5)/*Corrège* (f:4.1, r:173.9), and *Guercino* (f:2.8, r:554.1)/*Guerchin* (f:4.2, r:612.4), are highly salient in both languages, alongside more canonical extra-regional references such as *Raphael* (f:3.1, r:38.7)/*Raphaël* (f:3.4, r:16.9), whose works are housed in collections across the region. In addition, both lists clearly highlight the Bolognese school of painting: *bolognese* (f:2.7, r:715.2)/*bolonais* (f:5.4, r:2155.8) show strong collocational attraction with *school* (MI:8.8) and *école* (MI:8.3) respectively, suggesting that Bologna is framed as a center of a recognizable artistic tradition.

The two subcorpora differ, however, in how salience is distributed within this domain. The English top 50 includes *Caracci* as a collective label, alongside *Guido* (f:4.9, r:216.4), a keyword with multiple possible referents but strongly associated with Guido Reni (MI:10.9). By contrast, the French top 50 gives prominence to *Carrache*, *Annibal* (f:2.5, r:84.5) and *Augustin* (f:2.1, r:15.7), and also includes additional artist names such as *Albane* (f:2.3, r:99.2)⁹.

Unlike the BER-TL corpus, both ITA-TL subcorpora associate the visual arts with canonical artists, including *Donatello* (En: f:1.2, r:497.9), *Veronese* (f:0.4, r:29.6; *Véronese* (Fr: f:0.4, r:19), *Tintoretto* (f:0.8, r:60.9)/*Tintoret*; (f:0.4, r:12.9), *Leonardo* (f:0.7, r:33.3)/*Léonard* (f:0.7, r:11.6), the latter strongly associated with *Vinci* (En: MI:13.8 ; Fr: MI:13.2). However, some of the Emilian painters mentioned in the BER-TL do appear among the ITA-TL keywords: in the French list, *Corrège* (f:0.3; r:17.3), *Carrache* (f:0.1; r:20.1), and *Guerchin* (f:0.1; r:23.1); in the English one, *Hannibal* (f:0.3; r:15.1) and *Guido* (f:0.3; r:11.3), which occur in collocation with *Caracci* (MI:13) and *Reni* (MI:15.2) respectively. Concordance analysis shows that these items typically appear in contexts where the artists are cited for artworks located in other Italian cities¹⁰, discussed in broader reflections on Italian art¹¹, or mentioned in comparison with other

⁹ These keywords also occur in the English BER-TL list at lower levels of ranking.

¹⁰ “On your right is the Church of S. Ambrogio [of Genoa], built by Pallavicini, with three pictures, a Guido Reni, the Assumption of the Virgin, and two Rubens”

¹¹ “For the last three hundred years, ever since the Carracci opened their academy at Bologna, there has been no lack of art education in Italy.”

artists elsewhere in Italy¹². The comparison with ITA-TL thus highlights the cultural significance of Emilia-Romagna's artistic tradition more broadly in Italy.

A further contrast concerns the lexical realization of artistic discourse. The English BER-TL subcorpus emphasizes media and techniques through items like *painting* (f:7.5, r:15.1), *fresco* (f:4.6, r:33), and *mosaic* (f:3.3, r:21.3), whereas the French BER-TL top 50 includes *tableau* (f:14.4, r:10.9), *peintre* (f:5.7, r:11.2), and *autel* (f:5.8, r:10.2), the latter frequently occurring in collocation with *tableau* (MI:6.1). The difference between the two languages appears again to be one of relative salience. In fact, art-related vocabulary is present further down the ranked lists in both BER-TL subcorpora, suggesting that pictorial heritage is a distinctive feature that characterizes framing of Emilia-Romagna. This regional specificity is further reinforced by the comparison with the ITA-TL corpus, where none of these terms reach the $RRF \geq 10$ threshold in the respective languages. The only exception is *fresco* (f:2.6, r: 18.6), which is however considerably less salient than in the English BER-TL.

Literary and historical figures emphasize the region's cultural profile. *Ariosto* (f:2.2, r:889.6)/*Arioste* (f:3.1, r:141.1) and *Dante* (En: f:2.2, r:22; Fr: f:3, r:17.6) appear in connection with specific regional cities, most notably Ferrara and Ravenna, anchoring literary prestige to local urban memory. Unlike *Ariosto*, whose salience is concentrated in the BER-TL corpus, *Dante* also appears prominently in the French ITA-TL top 50 (f:2.7, r:22.5), which is unsurprising given his broader association with Italian literary culture and cities such as Florence. This salience in the French ITA-TL subcorpus, however, should be interpreted with caution: although this lexical item is mentioned in 20 out of 27 of the texts, roughly 58% of its occurrences are concentrated in a single work, pointing to an uneven distribution that was not eliminated by the dispersion criterion, and which nonetheless distorts the statistical results.

Religious heritage is also prominent in both BER-TL subcorpora and is lexicalized through a broad and overlapping set of lexical items that includes ecclesiastical authorities, sacred figures, iconographic subjects, and religious sites. Examples include *pope* (f:6.1, r:23.8)/*pape* (f:8.3, r:15.4), *cardinal* (En: f:2.8, r:19.3; Fr: f:5.2, r:13.7), *madonna* (f:5.2, r:59.4)/*madone* (f:3.6, r:15.5) and *Petronio* (f:1.8, r:1442.2)/*Pétrone* (f:3.2, r:2539). Moreover, the English top 50 further evokes iconographic imagery by the presence of *virgin* (f:5.9, r:14.9), *angel* (f:3.7, r:10.3), and *saviour* (f:1.7, r:11.5), whereas the French list places greater emphasis on sacred sites, with items such as *autel* and *basilique* (f:1.9, r:15.7), and on saints and ecclesiastical figures (*Jérôme*, *Antoine*, *Benoît*, *Catherine*). Concordance analysis reveals that most of these items occur also in descriptions of sacred buildings and artworks (e.g. *church of St. Petronio*), functioning simultaneously as religious, artistic and cultural heritage landmarks.

More broadly, several personal names occur across multiple thematic domains, linking political history, artistic traditions, and architectural heritage. For instance, the keyword *Theodoric* (f:1.8, r:75.6) in the English BER-TL occurs in contexts referring both to historical narrative and to monuments such as the mausoleum at Ravenna. Dynastic memory is likewise salient, with the Este family evoked directly in English through *Este* (f:1.8, r:176.6) and indirectly in French through *Alphonse* (f:2.2, r:30.8), which is strongly associated with *Este*¹³ (MI:8.9). Together, these items suggest that travel writing frequently constructs an image of Emilia-Romagna through the cultural and political actors who shaped its monumental landscape.

¹² “Il y avait alors à Gênes deux peintres en vogue, appelés les frères Carlone (...) Ils prétendaient ressusciter la peinture dans leur pays, comme les Carraches à Bologne.”

¹³ *Este* itself appears at rank 55 in the French keyword list (f:1.6, r:70.2).

A further point of convergence between the BER-TL subcorpora is the presence of Italian function words such as *di*, *della*, *che*. In both subcorpora, such items typically occur in proper names of artists, monuments or places (e.g. *Arca di San Domenico*, *Monte della Guardia*), and can therefore be read as markers of local specificity and interlingual contact. More generally, the English list appears somewhat more receptive to Italianized urban and architectural vocabulary, as shown by items such as *piazza*, *palazzo* and *cupola* in the top 50, and by terms like *duomo*, *campanile*, *porta* and *torre* in the top 250. This does not imply that architecture is absent from the French corpus: *piazza* and *palazzo* appear further down the French ranked list, and *basilique*, together with items such as *dôme* and *cathédrale*, points in the opposite direction. Unlike pictorial terms, however, architecture-related lexical items also occur in ITA-TL across both languages. Their presence should therefore be understood primarily as a broader feature of Italy, rather than as a specifically Emilian-Romagnol marker.

Collocational and concordance examination also reveals a minor but consistent gastronomic thread in the specific vocabulary of Emilia-Romagna, with lexical items that are absent from the ITA-TL subcorpora. In the English BER-TL, the keyword *Bologna* functions not only as a place name but also as a gastronomic reference (*bologna sausage*). The French BER-TL subcorpus offers a parallel case in *parmesan*, which refers both to the painter Parmigianino and, more marginally, to Parmesan cheese. The region's culinary identity is further attested by the presence of *sausage* (f:0.6, r:11.4) and *parmesan* (f:0.4, r:28) in the English BER-TL, and *mortadelle* (f:0.2, r:129.3) in the French BER-TL¹⁴.

5. Conclusion

This study has examined the lexical representation of Emilia-Romagna in English and French travel literature through a multi-corpus keyword analysis. Overall, the findings reveal a broadly stable representation of the region as a culturally prestigious destination defined by pictorial traditions, religious heritage, historical figures, and culinary culture, with cross-linguistic differences concerning lexical salience rather than fundamentally different perceptions. The comparison with a corpus of travel writing on other Italian regions confirms the distinctiveness of Emilia-Romagna, particularly in its association with its painting tradition.

By focusing on a single Italian region from a cross-linguistic perspective, the study offers a novel contribution to corpus-based travel literature research, providing a more nuanced understanding on how regional identity is constructed and mediated in historical travel discourse as well as insights relevant to cultural heritage valorization and tourism promotion.

Some limitations should nonetheless be acknowledged. The qualitative categorization was largely restricted to the top 50 keywords per subcorpus due to space limitations. Furthermore, the extraction procedure focuses primarily on a single combination of keyness measures. Future work will extend the analysis to a larger portion of the ranked lists and explore alternative metrics, with the aim of gaining additional insights into the data and providing a more comprehensive picture of the lexical representation of the region.

¹⁴ The association between the region and its culinary identity is further reinforced linguistically: *bologna sausage* and *parmesan* derive metonymically from the place names Bologna and Parma.

References

- Anthony, L. (2020). *AntConc* (Version 3.5.9). Computer Software. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Baker, P. (2004). Querying Keywords: Questions of Difference, Frequency, and Sense in Keywords Analysis: Questions of Difference, Frequency, and Sense in Keywords Analysis. *Journal of English Linguistics*, 32(4): 346-359.
- Damerau, F. J. (1993). Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Information Processing & Management*, 29(4): 433-447.
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.
- Egbert J. and Bieber D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, 14(1): 77-104.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930–1955. In *Studies in Linguistic Analysis*. Blackwell.
- Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In *Corpus approaches to discourse*. Routledge.
- Gerbig, A. (2010). Key words and key phrases in a corpus of travel writing: From Early modern English literature to contemporary “blooks”. In Bondi, M. and Scott, M., editors, *Keyness in texts*. John Benjamins Publishing.
- Gries, S. T. (2010). Useful Statistics for Corpus Linguistics. In Sánchez Pérez, A. and Almela Sánchez, M., editors, *A Mosaic of Corpus Linguistics: Selected Approaches*. Peter Lang.
- Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2): 1-33.
- Henkel, D. (2020). Lexical Association as an Indicator of Conceptual Convergence and Divergence in English, French and Italian. In Aláman, A. and Zotti, V., editors, *The Language of Art and Cultural Heritage. A plurilingual and digital perspective*. Cambridge Scholar Publishing.
- Moncomble, F. (2025) 2026. *Project Gutenberg Corpus Builder*. Web application. Arras, France: Université d’Artois. <https://corpustools.prendrelangue.fr/pgcorpusbuilder/>
- Pojanapunya, P. and Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1): 133-167.
- Pojanapunya, P. and Watson Todd, R. (2021). The Influence of the Benchmark Corpus on Keyword Analysis. *Register studies*, 3(1): 88-114.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester.
- Scott, M. (1997). PC analysis of key words – And key key words. *System*, 25(2): 233-245.
- Scott, M. and Tribble, C. (2006). *Textual patterns: key words and corpus analysis in language education* (Studies in corpus linguistics 22). John Benjamins Publishing.
- Sprugnoli, R. “Two days we have passed with the ancients...”: a Digital Resource of Historical Travel Writings on Italy. In Spampinato D., editor, *The Book of Abstract of AIUCD 2018*. Bologna.
- Zotti, V. (2026), Lessico artistico plurilingue e valorizzazione del patrimonio culturale italiano: i progetti UNICittà e UniVOCittà. In Altmanova, J., Centrella, M. and Grimaldi, C., editors, *Terminologia del patrimonio culturale materiale e immateriale: analisi e approcci di studio*. Peter Lang.