Cross-linguistic Representation for Multilingual Access to Archaeological Data

Giulia Speranza, Innovative Industrial PhD student [gsperanza@unior.it]
Supervisor: Prof. PhD **Johanna Monti** [jmonti@unior.it]
UNIOR NLP Research Group, University of Naples "L'Orientale"
Department of Literary, Linguistic and Comparative Studies, Via Duomo 219 - 80138, Naples, Italy

1. Introduction

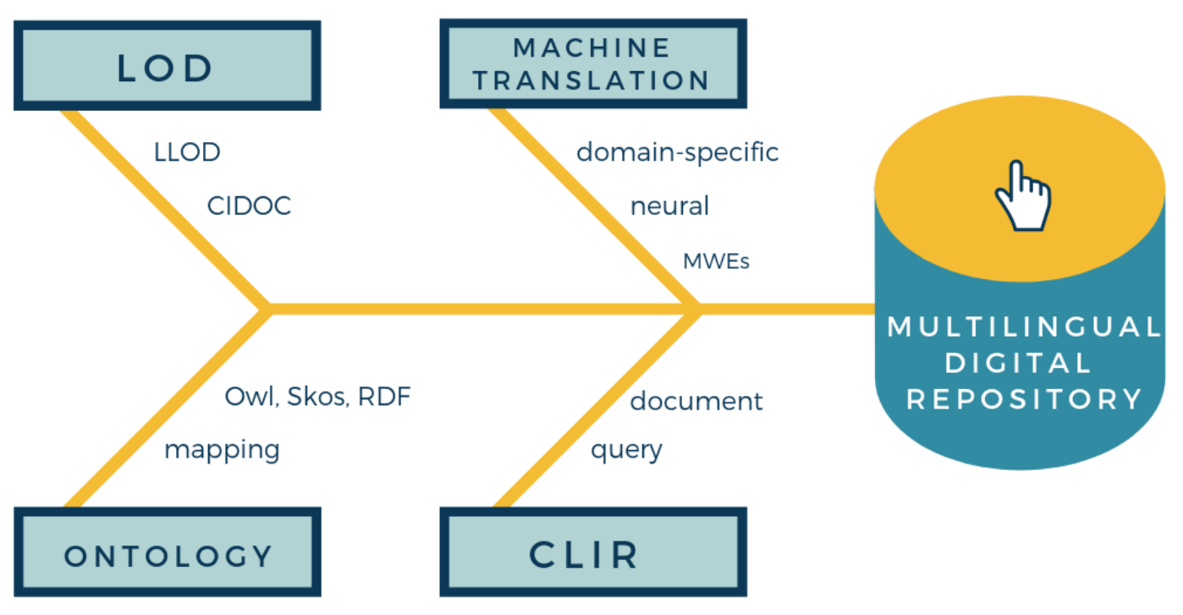
The project can be framed within the more general process of digital innovation of Cultural Heritage (CH), in particular, the internationalization of archaeological data in a multilingual perspective with the final aim of enhancing the interaction and communication between cultural institutions and the public, thus contributing to make the cultural experience in museums or archaeological sites more interactive and inclusive, overcoming linguistic and cultural barriers. The project is in line with the objectives set by the European Institutions for the Year of Cultural Heritage 2018, which promote the application of digital technologies in the field of humanities (Digital Humanities). The scientific community stresses the need for multilingualism in the field of CH, which, compared to other specialized field such as medicine or law, relies more and more on the use of the most common *linguae francae* to solve the issue of communicating globally [1]. The actual contribution that disciplines such as linguistics and machine translation (MT) may offer to cultural institutions, constitutes a valid method to provide international visitors with an inclusive cultural experience, bridging the linguistic gap.

2. State-of-the-art

The starting point of this research is to analyse the multilingual resources available online as open data such as dictionaries, thesauri, term banks (e.g. Getty and Wikidata). Another key-point is the investigation about previous research and applications in the field of computational linguistics and machine translation technologies applied to CLIR [2]. The research project also focuses on examining the recent experiments on Neural Machine Translation systems, with a special focus on domain-specific translation environments. Part of the project will concern an in-depth study about the possibilities offered by the semantic web technologies [3] and the application of LOD principles as well as CH standards (e.g. CIDOC, EDM) with particular attention to terminology and lexicography models for the representation and creation or conversion of annotated multilingual resources as LOD.

3. Methodology

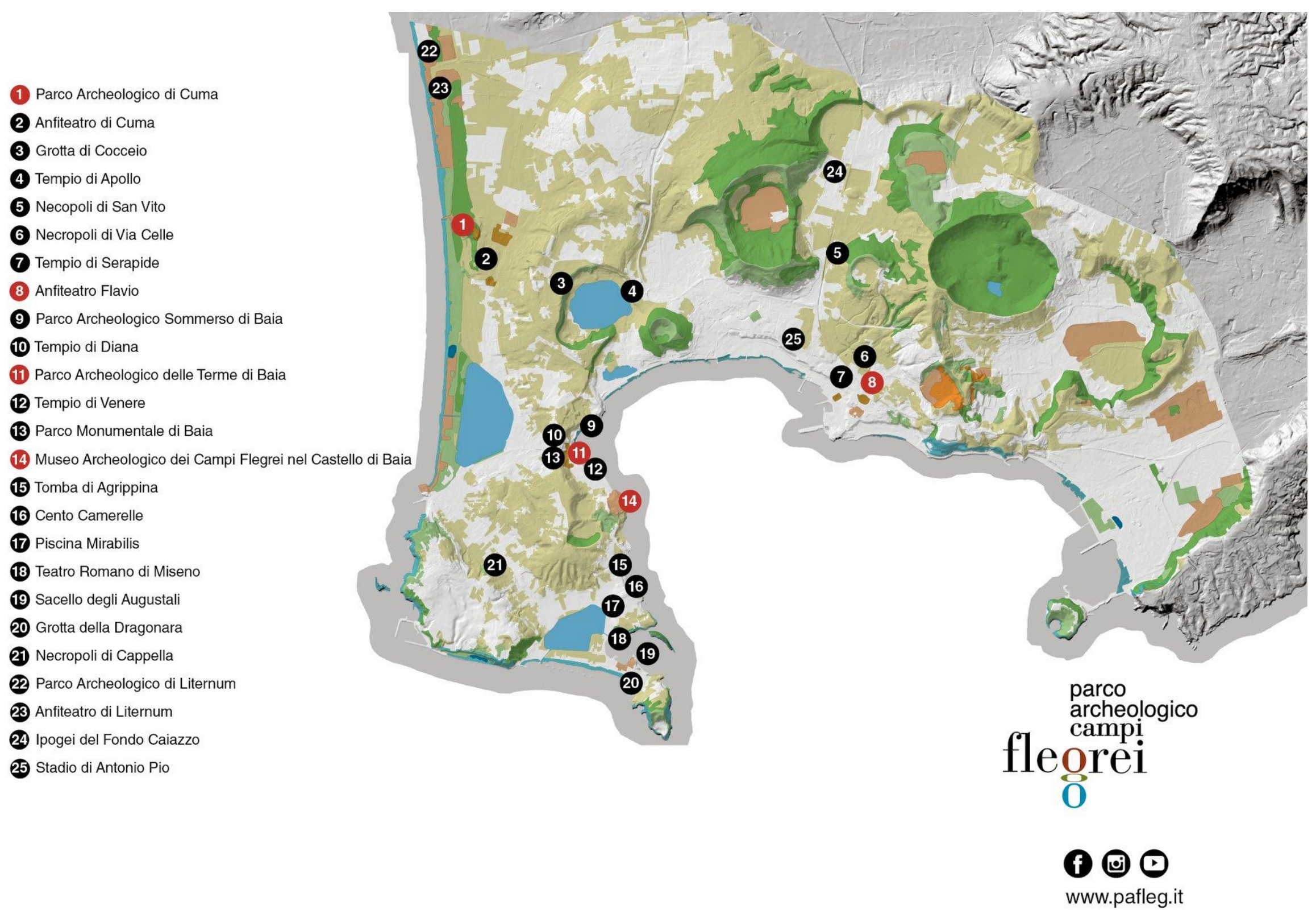
The focus is on a multidisciplinary approach, which ranges from computational linguistics (CL) and Natural Language Processing (NLP) to the use of new technologies for machine translation (MT) in the field of Cross-Language Information Retrieval (CLIR), as well as the application of Linguistic Linked Open Data (LLOD) principles. An in-depth study of ontologies and the CIDOC Conceptual Reference Model (CRM) for describing the relationships used in cultural heritage documentation, is a crucial aspect of the research project.



4. Objectives

The final aim of this research is to create a digital repository containing information about cultural objects of archaeology in Italian, English and German, with the possibility of expanding it with other languages.
Based on the gold terminology collected, a multilingual web portal or application for pc, smartphone and tablet will be developed.
In conclusion, the resource to be implemented is intended to engage with a multicultural public making use of multimodal channels.

5. Case Study


1 Parco Archeologico di Cuma
2 Anfiteatro di Cuma
3 Grotta di Cocceio
4 Tempio di Apollo
5 Necopoli di San Vito
6 Necropoli di Via Celle
7 Tempio di Serapide
8 Anfiteatro Flavio
9 Parco Archeologico Sommerso di Baia
10 Tempio di Diana
11 Parco Archeologico delle Terme di Baia
12 Tempio di Venere
13 Parco Monumentale di Baia
14 Museo Archeologico dei Campi Flegrei nel Castello di Baia
15 Tomba di Agrippina
16 Cento Camerelle
17 Piscina Mirabilis
18 Teatro Romano di Miseno
19 Sacello degli Augustali
20 Grotta della Dragonara
21 Necropoli di Cappella
22 Parco Archeologico di Liternum
23 Anfiteatro di Liternum
24 Ipogei del Fondo Caiazzo
25 Stadio di Antonio Pio
parco archeologico campi flegrei
www.pafleg.it

The case study will be on the specialized terminology used to describe archaeological finds coming from the archaeological area of Campi Flegrei in the Campania region (Italy). The source data-set is composed of highly technical Italian terms but also words from everyday language, multi-word expressions (MWEs), Greek and Latin words, compounds, loan-words, cultural-bound elements and ambiguous words. This variety becomes even more compelling from a translation perspective. Multi-word expressions, in particular, are one of the trickiest issues when dealing with MT and NLP tasks [4].

6. Partners

- **Smart Apps Soc. Coop.:** provides its digital repository containing terminological dictionaries related to 27 types of CH in the Italian language (180.000 entries) and application programming skills
- **Berlin School of Library and Information Science (Humboldt University of Berlin):** offers insights for investigation on CLIR, Computer Science and Digital Cataloguing, Europeana Data Model
- **University of Naples "L'Orientale":** supplies ways for deepening knowledge on Linguistics, Computational Linguistics, Neural Machine Translation, Terminology management and Specialized Lexicon representation

7. References

[1] Cecilia Garibay, Steven Yalowitz, and Guest Editors. Redefining multilingualism in museums: A case for broadening our thinking, 2015.

[2] Cristina España-Bonet, Juliane Stiller, Roland Ramthun, Josef van Genabith, and Vivien Petras. Query translation for cross-lingual search in the academic search engine pubpsych. In *Proceedings of the 12th International Conference on Metadata and Semantics Research. Metadata and Semantics Research Conference (MTSR-2018), October 23-26, Limassol, Cyprus*. Springer, 10 2018.

[3] Konstantinos N Vavliakis, Georgios Th Karagiannis, and Pericles A Mitkas. Semantic Web in cultural heritage after 2020. In *Proceedings of the 11th International Semantic Web Conference (ISWC), Boston, MA, USA*, pages 11–15, 2012.

[4] Johanna Monti, Ruslan Mitkov, Violeta Seretan, and Gloria Corpas Pastor. Multiword units in machine translation and translation technology. In Ruslan Mitkov, Johanna Monti, Violeta Seretan, and Gloria Corpas Pastor, editors, *Multiword units in machine translation and translation technology*, pages 1–38. John Benjamins Publishing Company, 2018.