AIUCD 2021
DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale
10° congresso annuale PISA 19-22 gennaio
DIGITAL PUBLIC HUMANITIES
OPEN CULTURE
RETI SOCIALI
TECH ECONOMY
E-PARTICIPATION
TECNOLOGIE ASSISTIVE

## Versione PROVVISORIA del contributo presentato al Convegno Annuale

DISCLAIMER

Questa versione dell'abstract non è da considerarsi definitiva e viene pubblicata esclusivamente per facilitare la partecipazione del pubblico al convegno AIUCD 2021

Il Book of Abstract contenente le versioni definitive e dotato di ISBN sarà disponibile liberamente a partire dal 19 gennaio sul sito del convegno sotto licenza creative commons.

# Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models

A. Marinetti[1], F. Murano[2], V. Quochi[3], M. Ballerini[2], F. Boschetti[3], A.M. Del Grosso[3], S. Piccini[3] , L. Rigobianco[1] , P. Solinas[1]

[1] Università "Ca'Foscari", Venezia, Italy <linda;solinas;luca.rigobianco>(«»)unive.it

[2] Università di Firenze, Italy name.surname(«»)unifi.it

[3] Istituto di Linguistica Computazionale, CNR, Italy name.surname(«»)ilc.cnr.it

## ABSTRACT

The poster presents a newly started project about languages and cultures of Ancient Italy, which brings together competences from Historical Linguistics, Computational Lexicography and Digital Humanities. The main objective of the project is to investigate the cultures of ancient Italy on the basis of their linguistic documentation (7th - 1stc. B.C.) by means of digital tools specifically tailored for their peculiarities.

## KEYWORDS

Restsprachen, Ancient Italy, Epigraphic Text Corpora, Computational Lexica, Semantic Web

## 1. INTRODUCTION AND BACKGROUND

During the period dating from the appearance of writing (8th c. B.C.) until complete linguistic Romanisation (1st c. B.C. to 1st c. A.D.), ancient Italy is characterised by the presence of numerous linguistic varieties, which are both Indo-European and non-Indo-European. These varieties are documented by *corpora* that are not homogenous in their quantity, quality, geographical distribution, and chronology. Apart from this lack of homogeneity, with the exception of the Latin of Rome they can be qualified as *Restsprachen*, i.e. fragmentarily attested languages, inasmuch as their linguistic documentation consists almost exclusively of epigraphic texts, mostly short and repetitive. For this reason, *Restsprachen* often present specific problems concerning the reading of the inscriptions, the segmentation into words, their linguistic analysis, and their interpretation. The linguistic framework as *Restsprachen* of the languages involved in this project – Venetic, Oscan, Faliscan, and Celtic – thus requires the adoption of specific methods and approaches for the analysis and edition of their texts[10][17][20]. To achieve this challenging goal, we will combine the traditional method proper to Historical Linguistics and its products (study, edition, and the linguistic, historical, and cultural commentary on the texts) with the setting up of digital technologies that will facilitate scholars in their activities.

## 2. OBJECTIVES

The main goal of the project is to investigate the cultures of ancient Italy on the basis of the relevant linguistic documentation in order to show the forms of linguistic variability in Italy before Romanisation. The study and valorisation of this cultural heritage is to be achieved through interdisciplinary work that brings together methods and practices from the "traditional" study of such materials, computational lexicology and lexicography, semantic web and various digital humanities technology, such as digital archiving, textual corpora, contextualisation and visualisation methods. In particular, we aim at digitising key materials and deploying electronic tools and resources that will permit their easy consultation and revision, which in turn will both facilitate scholars' work and allow greater disclosure of the acquired scientific knowledge. Given the experimental nature of the project, we will focus on the Venetic, Oscan, Faliscan, and Celtic *corpora*. Although ambitious and challenging, given the main goal of the project, i.e. to develop a new methodology that leverages specifically tailored digital technologies to facilitate scholars' work and collaboration while at the same time ensuring long term preservation of the acquired knowledge, the risks are calculated. To minimise them we build upon our own direct experiences in developing tools for the digital humanities and in modeling language data, as well as upon other similar national and European endeavours. At the end of the project, it is expected that both the methodology as well as the digital toolkit and technology set up will be ready to be extended to the other *Restsprachen*.

## 3. METHODOLOGIES AND EXPECTED OUTCOMES

The project plans to create a digital set of interrelated resources: a *corpus* of epigraphic texts, a computational lexicon for the languages involved, a dataset of bibliographic references, and an experimental semantic dataset of interpretations. The *corpus* of the texts will be managed and exploited in a digital archive containing the formal representation of the texts leveraging the TEI/EpiDoc encoding schema [9]; it may be necessary to create an *ad-hoc* schema for the peculiarities presented by languages of fragmentary attestation, the EpiDoc model having been used so far only for the creation of annotated 'major' corpora, especially of Greek and Latin inscriptions. For example, *Restsprachen* may require the identification of standards suitable for the encoding of unusual and odd writing ductus, punctuation marks with particular values (e.g. syllabic punctuation), characters whose linguistic nature is uncertain, multiple hypotheses of word segmentation, etc. The application of the EpiDoc model to *Restsprachen* is, thus, a complete novelty in the field. Each text in the archive will be enriched with shared and standard metadata allowing for an accurate description, both as a linguistic object (text: language, alphabet, date, etc.) and as a material object (support: chronology, data of discovery, material, etc.). Accordingly, such a description will facilitate the retrieval of relevant information. The project additionally plans to experiment with the use of CRMtex[11][12] and CRMinf[4] extensions of CIDOC CRM, the *de facto* standard ontology in the Digital Humanities for the representation of the texts and their scientific interpretation in a semantic format[5]. As regards the lexicon, the project will investigate the specific requirements for the design of an efficient computational lexical model specifically dedicated to languages of fragmentary attestation in order to produce a multilingual (Venetic, Oscan, Faliscan, and Celtic) computational lexicon. We will adopt Semantic Web standards and vocabularies for providing a structured and formal representation of the lexical items and their related information as well as for allowing for a sophisticated semantic access to the *corpus* of epigraphies [1][16]. The challenges to be faced in lexical modeling are numerous, since we are dealing with Restsprachen, and they range from lemmatisation issues, as in many cases the relationship between words and lemmas is not certain for various reasons (different graphic standards, difficult linguistic analysis, incomplete paradigms, etc.), to sense representation, given that meanings are often only partially and hypothetically reconstructible, and smart reference or linking to attestations. Fortunately, recent efforts within modeling and standardisation initiatives are tackling many of these critical issues, and they will provide us with inspiration and guidance. We refer in particular to the works on the extension of existing representation models such as TEI and Ontolex-lemon for the account of attestations[3] and etymologies[2][14][13]. Models and methods for interlinking the datasets will also be designed as the backbone for a digital toolkit that will allow for both the creation/revision of some of the materials and their online fruition, see [15][8][19][18]. Specifically, the heart of the toolkit will be a lexicon editing and consultation service specifically tailored to Restsprachen, based on inhouse previous experiences[1]; this will be enriched for allowing the interlinking among the different datasets, thus acting as a sort of hub that will primarily integrate lexicon and epigraphy transcriptions, together with contextual metadata, bibliography, and experimentally the hermeneutic positions. It is also planned to experiment with Domain-Specific Languages to deploy a system that can assist scholars in the creation of the textual digital resources and ensure compatibility with the aforementioned standards[19]. The tools and resources produced and developed within the project will finally be made available through relevant European-wide Research Infrastructures such as CLARIN[6] and DARIAH[7], the two currently most relevant infrastructures for the e-Humanities and (immaterial) Cultural Heritage. This will ensure both a long-term preservation of the resources produced and a high valorisation of this heritage.

## 4. ACKNOWLEDGEMENTS

The project is supported by the Italian Ministry of Research [PRIN  2017XJLE8J]

# REFERENCES

[1] Bellandi, Andrea, Emiliano Giovannetti, Silvia Piccini and Anja Weingart. "Developing LexO: a Collaborative Editor of Multilingual Lexica and Termino-ontological Resources in the Humanities". In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*, Montpellier, France, 19 September 2017. Edited by Francesca Frontini, Larisa Grčić Simeunović, Špela Vintar, Anas Fahad Khan, Artemis Parvisi, 2017. https://www.aclweb.org/anthology/W17-7010.pdf.

[2] Bowers, Jack and Laurent Romary "Deep encoding of etymological information in TEI". Journal of the Text Encoding Initiative, Issue 10, 2016. https://doi.org/10.4000/jtei.1643

[3] Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Anas Fahad Khan, Sander Stolk, Thierry Declerck, John Philip McCrae. "Modelling Frequency and Attestations for OntoLex-Lemon". In: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*. Edited by: Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi and Besim Kabashi. LREC, May 2020. pp.1--9. ELRA. European Language Resources Association, 2020.

[4] "CIDOC CRMinf," October 1, 2020. http://www.cidoc-crm.org/crminf/.

[5] "CIDOC Conceptual Reference Model," October 1, 2020. http://www.cidoc-crm.org

[6] "CLARIN-IT – L'Infrastruttura Comune Italiana per le Risorse e le Tecnologie Linguistiche", October 1, 2020. www.clarin-it.it.

[7] "DARIAH – Digital Research Infrastructure for the Arts and Humanities", October 1, 2020. http://stdl.cnr.it/it/dariah.

[8] Del Grosso, Angelo M., Andrea Bellandi, Emiliano Giovannetti, Simone Marchi, and Ouafae Nahili. "Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts". In *Proceedings of the IEEE 5th International Congress on Information Science and Technology (CiSt)*. Marrakech, Morocco, 21-27 October 2018, 214–219, 2018.

[9] "EpiDoc: Epigraphic Documents in TEI XML," October 1, 2020. https://epidoc.stoa.org/gl/latest/toc-it.htm.

[10] Farney, Gary D., and Guy Bradley. *The Peoples of Ancient Italy*. De Gruyter, 2017.

[11] Felicetti, Achille, and Francesca Murano. "Scripta manent: a CIDOC CRM semiotic reading of ancient texts." *International Journal on Digital Libraries* 18 (2017): 263–270.

[12] Felicetti, Achille, Francesca Murano, Paola Ronzino and Franco Niccolucci. "CIDOC CRM and Epigraphy: a Hermeneutic Challenge." In *Extending, Mapping and Focusing the CIDOC CRM. CRMEX 2015 Workshop, 19th International Conference on Theory and Practice of Digital Libraries*. Poznan, September 14-18 2015, edited by Paola Ronzino and Franco Niccolucci, 55–68, 2015. http://ceur-ws.org/Vol- 1656/paper5.pdf

[13] Khan, Fahad, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem. "Modelling Etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a Use Case". *Proceedings of the 12th Language Resources and Evaluation Conference – LREC 2020*, May 2020, Marseille, France, 2020. https://hal.inria.fr/hal-02618067/document

[14] Khan Fahad. "Towards the Representation of Etymological Data on the Semantic Web". *Information* 2018, 9(12), 304. https://doi.org/10.3390/info9120304

[15] Khan, A. Fahad, and Federico Boschetti. "Towards a Representation of Citations in Linked Data Lexical Resources". In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Context*, edited by Jaka Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek, 137–147. Ljubiana: Ljubljana University Press, 2018.

[16] Khan, A. Fahad, Andrea Bellandi, and Monica Monachini. "Tools and Instruments for Building and Querying Diachronic Computational Lexica", In *Proceedings of the Language Technology Resources and Tools for Digital Humanities (LT4DH 2016)*, Osaka, Japan, December 11-16 2016, 164–171, 2016.

[17] Klein, Jared, Brian Joseph, Matthias Fritz, and Mark Wenthe. *Handbook of Comparative and Historical Indo-European Linguistics*. Berlin ; Boston: Mouton De Gruyter, 2018.

[18] Mambrini, Francesco, Cecchini Flavio Massimo, Franzini Greta, Litta Eleonora, Passarotti Marco Carlo, Ruffolo Paolo. "LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web", *Umanistica Digitale* 8(2020), pp. 63-78. DOI: 10.6092/issn.2532-8816/9975

[19] Mugelli G., Boschetti F., Del Gratta R., Del Grosso A. M., Khan F. e Taddei A. "A user-centred design to annotate ritual facts in ancient greek tragedies", *Bulletin of the University of London. Institute of Classical Studies* vol. 59 (2016), pp. 103-120.

[20] Passarotti Marco, Mambrini Francesco, Franzini Greta, Cecchini Flavio Massimiliano, Litta Eleonora, Moretti Giovanni, Ruffolo Paolo, Sprugnoli Rachele. "Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin". In *Current Approaches in Latin Lemmatization*. Edited by Marco Passarotti . *Studi e Saggi Linguistici*, LVIII (1) 2020, pp. 177-212. DOI: 10.4454/ssl.v58i1.277

[21] Prosdocimi, Aldo Luigi, ed. *Lingue e Dialetti dell'Italia Antica*. Vol. 6. Popoli e Civiltà dell'Italia Antica. Roma; Padova: Biblioteca di Storia Patria, 1978.