Journal of Open Humanities Data

Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM)

DATA PAPER

REUBEN J. PITTS

ubiquity press

ABSTRACT

The *Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE* (CEIPoM) is a linguistic database which covers the Oscan, Umbrian, Old Sabellic, Messapic and Venetic languages, as well as epigraphic Latin up to 100 BCE. The database is hosted on GitHub and Zenodo, and provides manually annotated linguistic information on all levels of language structure, ranging from phonology to syntax. In providing a high-resolution digital dataset for language varieties that have until now been largely restricted to printed reference works, this corpus opens up new avenues for research into this unique ancient linguistic area.

CORRESPONDING AUTHOR:
Reuben J. Pitts
Faculty of Arts, KU Leuven, Leuven, BE
reuben.pitts@kuleuven.be

KEYWORDS:
corpus linguistics; language contact; linguistic area; Italic; epigraphy

TO CITE THIS ARTICLE:
Pitts, R. J. (2022). Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM). *Journal of Open Humanities Data*, 8: 1, pp. 1–4. DOI: *https://doi.org/10.5334/johd.65*

(1) OVERVIEW

REPOSITORY LOCATION

https://reubenjpitts.github.io/Corpus-of-the-Epigraphy-of-the-Italian-Peninsula-in-the-1st-Millennium-BCE/

Current version (1.2): *https://zenodo.org/record/5602978#.YXkw8Z5Bw2w*

DOI: *https://doi.org/10.5281/zenodo.4759134*

CONTEXT

This database was created in the context of a PhD project on language contact in Ancient Italy, entitled *The interplay between language contact and language change in a fragmentary linguistic area: the Italic peninsula in the first millennium BCE*.

https://www.kuleuven.be/onderzoek/portaal/#/projecten/3H190594?lang=en&hl=en

(2) METHODOLOGY

STEPS

Most of the data was entered manually by the author, based on standard reference works for the languages in question. In some cases, basic forms of automation were used to create an initial dataset which was then corrected. For instance, an initial morphological analysis for Venetic was created by linking the attested tokens to a digitised version of Lejeune's (1974: 315–341) Venetic word list, and the result was then systematically checked and corrected by the author. The method used for any given field is described in the accompanying documentation on GitHub.

A few fields were generated automatically using Python modules. These include, for instance, the field "Token_clean", which uses the *unidecode* package to generate a version of the token stripped of special characters, intended for ease of searching. Once again, the documentation on GitHub describes in detail which fields are automatic and how they are generated.

SAMPLING STRATEGY

The aim of the database is to include all texts in Oscan, Umbrian, Old Sabellic, Messapic and Venetic, as well as epigraphic Latin texts before 100 BCE. The corpus does not include Etruscan, due to the additional complexities of incorporating a non-Indo-European language into the structure of the database. Within the languages encompassed by the database, however, the primary aim is exhaustivity, and the corpus currently contains over 36,000 tokens.

QUALITY CONTROL

Data was entered manually and checked multiple times by the author.

(3) DATASET DESCRIPTION

OBJECT NAME

Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM)

FORMAT NAMES AND VERSIONS

CSV

CREATION DATES

2017–2021

DATASET CREATORS

Reuben J. Pitts

LANGUAGE

Metadata are provided in English.

REPOSITORY NAME

A continually updated version of the corpus is hosted on GitHub. Each old version of the corpus is permanently stored at Zenodo. In traditional publications CEIPoM should be cited as this paper, where relevant also specifying the version of the corpus used to achieve any given research result.

PUBLICATION DATE

2021–05–13

(4) REUSE POTENTIAL

This database has a wide range of applications in linguistic research on the languages of ancient Italy. Currently, such research is hampered by the absence of searchable digital information, as the description of these languages is mostly spread over disparate written reference works (e.g. Bakkum, 2009; Lejeune, 1974; Santoro, 1982; Untermann, 2000; Wachter, 1987). This database aims to address that research need head-on.

The salience of digital and corpus-based approaches to ancient languages has increased in recent years (e.g. Adamik, 2016; Eckhoff et al., 2018; Mambrini et al., 2020; Qiu et al., 2018), and these methods have proven their effectiveness even in relatively poorly attested languages. It goes without saying that a digital dataset is more easily and more efficiently queried than a written corpus, facilitating research results that would otherwise be difficult or impossible to achieve. Moreover, the use of a digital dataset means any research results thus obtained can be replicated by other researchers, conferring a key advantage in terms of academic transparency. These advantages hold true in fragmentary languages such as Venetic or Messapic as much as in large corpus languages such as Classical Latin or Greek.

Since annotation is provided on multiple levels of description, this corpus can serve as a tool for linguistic research of various kinds, including research on the syntax, word order, morphology, lexicon, semantics, phonology and orthography of the ancient languages in question. To give an example of a simple linguistic query in CEIPoM, if one is researching the usage of syntactic objects in these languages, one can simply use spreadsheet software to search for instances of *OBJ* in the field *Relation*, and thus obtain a list of all tokens in the corpus with a syntactic analysis containing this value. The GitHub documentation offers considerable detail on how each of these features are annotated, and how the different levels of linguistic description can be related to one another to formulate more complex queries.

In addition to the strictly linguistic annotation, chronological and geographical information (including longitude and latitude) is integrated into the data throughout, allowing the evolution and distribution of these linguistic features to be tracked through time and space. Although the focus of the corpus does not lie on epigraphical metadata, the texts in the corpus are linked to their ID in the Trismegistos database (Depauw & Gheldof, 2014), which means they can easily be linked to further metadata and bibliography, as well as to other epigraphic databases (such as EDR or EDCS). In addition to its linguistic uses, therefore, the database also holds promise for related fields such as history, epigraphy and onomastics.

The corpus focuses strongly on ensuring that the information provided for the languages of ancient Italy is *intercomparable*. This makes it particularly well adapted for the study of convergence, language contact and other cross-linguistic typological trends in ancient Italy. This region has sometimes been described as a linguistic area (Zair, 2016: 311–312), a geographic region where prolonged language contact is responsible for grammatical similarities across distantly related languages (Friedman & Joseph, 2017: 55). Since the data is (with a few clearly signalled exceptions) annotated in the same way for all six languages currently in the corpus, this makes it possible to track the evolving differences and similarities between these languages, and to test hypotheses on contact-based change in this region.

The main current limitation of the database lies in the fact that, inevitably, its data is not fully complete. In particular, the emphasis until now has been on providing a single plausible

linguistic analysis for each token, even when the scholarly literature offers multiple possible interpretations. Since this is frequently the case in disputed fragmentary texts, this may cause queries to miss potentially relevant and interesting forms. However, since the state of the data in each field is described in detail in the documentation on GitHub, researchers can take these limitations into account and adjust their use of this research tool in line with their research aims. Future updates to the corpus will continue to improve and fine-tune the quality of the data offered, as well as expanding the coverage of alternative analyses for individual tokens.

ACKNOWLEDGEMENTS

I would like to thank Toon Van Hal, Freek Van de Velde, Mark Depauw and Tom Gheldof for their help and advice in making this corpus.

FUNDING INFORMATION

This research was carried out with a grant from the *Fonds Wetenschappelijk Onderzoek (FWO) – Vlaanderen* (Research Foundation – Flanders) (grant no. 1150720N).

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Reuben J. Pitts *orcid.org/0000-0002-3960-1490*
Faculty of Arts, KU Leuven, Leuven, BE

REFERENCES

Adamik, B. (2016). Computerized Historical Linguistic Database of the Latin Inscriptions of the Imperial Age: Search and Charting Modules. In Á. Szabó (Ed.), *From Polites to Magos: Studia György Németh Sexagenario Dedicata* (pp. 13–27). Budapest: Debrecen.

Bakkum, G. C. L. M. (2009). *The Latin Dialect of the Ager Faliscus: 150 Years of Scholarship*. Amsterdam: Amsterdam University Press.

Depauw, M., & **Gheldof, T.** (2014). Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. In P. Goodale & N. Houssos (Eds.), *Theory and Practice of Digital Libraries—TPDL 2013 Selected Workshops* (pp. 40–52). Cham: Springer.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E., & **Jøhndal, M.** (2018). The PROIEL Treebank Family: A Standard for Early Attestations of Indo-European Languages. *Language Resources and Evaluation*, *52*(1), 29–65. DOI: *https://doi.org/10.1007/s10579-017-9388-5*

Friedman, V. A., & **Joseph, B. D.** (2017). Reassessing Sprachbunds: A View from the Balkans. In R. Hickey (Ed.), *The Cambridge Handbook of Areal Linguistics* (pp. 55–87). Cambridge, UK: Cambridge University Press. DOI: *https://doi.org/10.1017/9781107279872.005*

Lejeune, M. (1974). *Manuel de la langue vénète*. Heidelberg: Carl Winter Universitätsverlag.

Mambrini, F., Cecchini, F. M., Franzini, G., Litta, E., Passarotti, M. C., & **Ruffolo, P.** (2020). LiLa: Linking Latin: Risorse linguistiche per il latino nel Semantic Web. *Umanistica Digitale*, *8*, 63–78. DOI: *https://doi.org/10.6092/issn.2532-8816/9975*

Qiu, F., Stifter, D., Bauer, B., Lash, E., & **Ji, T.** (2018). Chronologicon Hibernicum: A Probabilistic Chronological Framework for Dating Early Irish Language Developments and Literature. In M. Ioannides, E. Fink, R. Brumana, P. Patias, A. Doulamis, J. Martins, & M. Wallace (Eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection* (pp. 731–740). Cham: Springer International Publishing. DOI: *https://doi.org/10.1007/978-3-030-01762-0_65*

Santoro, C. (1982). *Nuovi studi messapici*. Galatina: Congedo editore.

Untermann, J. (2000). *Wörterbuch des Oskisch-Umbrischen*. Heidelberg: Winter.

Wachter, R. (1987). *Altlateinische Inschriften: Sprachliche und epigraphische Untersuchungen zu den Dokumenten bis etwa 150 v. Chr*. Lausanne: Lang.

Zair, N. (2016). Vowel Weakening in the Sabellic Languages as Language Contact. *Indogermanische Forschungen*, *121*(1), 295–315. DOI: *https://doi.org/10.1515/if-2016-0016*

TO CITE THIS ARTICLE:
Pitts, R. J. (2022). Corpus of the Epigraphy of the Italian Peninsula in the 1st Millennium BCE (CEIPoM). *Journal of Open Humanities Data*, 8: 1, pp. 1–4. DOI: *https://doi.org/10.5334/johd.65*

Published: 03 January 2022