Publishing data papers in the humanities: my experience from the Journal of Open Humanities Data

Barbara McGillivray

SADiLaR DH Colloquium
17 May 2023
The Journal of Open Humanities Data

- Launched in 2015
- It is part of the ‘metajournals’ family in Ubiquity Press, publishing papers about:
  - research data
  - software
  - hardware

Credits: Paola Marongiu
Vision

Be a **key** part of a **thriving community** of scholars **sharing** humanities data
Open Humanities Data
The humanities have produced and collected data for centuries
What is different now?
Many words, many objects, all data
What’s special in humanities data?

- Manual curation
- Subjective/personal/individual interpretation, following standards/consensus
- Importance of context

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

**Author:** Reuben J. Pitts

**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 Sabellian, 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.

**Keywords:** corpus linguistics, language contact, linguistic area, Italic, epigraphy

**How to Cite:** 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. DOI: http://doi.org/10.5334/fohd.65
Open Humanities Data
Open Humanities and open science

“Open science commonly refers to efforts to make the output of publicly funded research more widely accessible in digital format to the scientific community, the business sector, or society more generally.”

Contextualisation is critical

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

Author: Reuben J. Pitts

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 Sabellian, 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.

Keywords: corpus linguistics, language contact, linguistic area, Italic, epigraphy

An example from linguistics


10 The Dative Alternation Revisited

Fresh Insights From Contemporary British Spoken Data

Gard B. Jenset, Barbara McGillivray and Michael Rundell

10.1 Motivation

A well-known feature of English grammar is the dative alternation, whereby a verb may be used in a V-NP-NP construction (Give me the money) or with a prepositional phrase in the pattern V-NP-PP, typically with the preposition to (Give the money to me). In this study, we use data from the early access Sample of the Spoken British National Corpus 2014 (the Spoken BNC2014S) to investigate the behaviour of six high-frequency verbs whose argument structure preferences include the dative alternation. Given that speakers have both patterns available to them, our goal is to discover whether the choice of pattern is motivated rather than arbitrary—and if so, which factors influence that choice.

Although the dative alternation is a well-researched topic, most published work draws either on introspection or on data from written sources. Using contemporary unscripted spoken text from face-to-face conversations takes us into new territory, especially as the linguistic data in the Spoken BNC2014S are complemented by rich sociolinguistic information on participating speakers. By “sociolinguistic information” we mean the social phenomena that co-occur with linguistic variables (Bayley, 2002, p. 118). The corpus represents a powerful new research resource, and in this chapter we show how it yields new insights into the use of the dative alternation.

<table>
<thead>
<tr>
<th>2</th>
<th>give</th>
<th>A9</th>
<th>VNN</th>
<th>you</th>
<th>3</th>
<th>you</th>
<th>TRUE</th>
<th>Z8</th>
<th>TRUE</th>
<th>a list</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>you</td>
<td>3</td>
<td>you</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>a list of special offers</td>
</tr>
<tr>
<td>4</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>you</td>
<td>3</td>
<td>you</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>a list of different types</td>
</tr>
<tr>
<td>5</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>you</td>
<td>3</td>
<td>you</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>a few like oh these</td>
</tr>
<tr>
<td>6</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>them</td>
<td>4</td>
<td>them</td>
<td>TRUE</td>
<td>Z8</td>
<td>FALSE</td>
<td>exactly what you want</td>
</tr>
<tr>
<td>7</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>them</td>
<td>4</td>
<td>them</td>
<td>TRUE</td>
<td>Z8</td>
<td>FALSE</td>
<td>the makeup</td>
</tr>
<tr>
<td>8</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>him</td>
<td>3</td>
<td>him</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>the choice of which to experience</td>
</tr>
<tr>
<td>9</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>me</td>
<td>2</td>
<td>me</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>half an hour of being a</td>
</tr>
<tr>
<td>10</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>you</td>
<td>3</td>
<td>you</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>all sorts of crap</td>
</tr>
<tr>
<td>11</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>me</td>
<td>2</td>
<td>me</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>all sorts of crap</td>
</tr>
<tr>
<td>12</td>
<td>give</td>
<td>A9</td>
<td>VNN</td>
<td>me</td>
<td>2</td>
<td>me</td>
<td>TRUE</td>
<td>Z8</td>
<td>TRUE</td>
<td>the agency number</td>
</tr>
</tbody>
</table>
(4) Reuse potential

There is a growing trend in linguistics for quantitative research, a trend which is not proceeding at the same pace in all branches of linguistics [6]. A natural corollary of this increasing quantitative research is a focus on replicable and reproducible research [7].

True replicability is difficult to achieve in many field-based disciplines and social sciences [7]. A more achievable goal is reproducibility. Reproducibility is clearly important for increasing scientific transparency and accountability. A move towards greater reliance on usage-based theory development can drive convergence in linguistic theory generally [8] as well as in specific sub-fields [6]. Despite some notable exceptions (such as second language acquisition), most linguistic sub-fields do not have a strong tradition for making research data available [7]. Publishing not only corpora and raw data, but also the annotated research datasets means that data can be compared quantitatively across
JOHD’s team

- 1 editor-in-chief
- 3 associated editors
- 2 social media editors
- 2 editorial assistants
- 6 copy editors
- 10 members in the editorial board

➡️ We are researchers in linguistics, classics, law and more
JOHD’s publications

Two types of publication, both data-focused:

1) data papers (1000 words)
2) research papers (3000-5000 words)

We have published 83 papers so far
JOHD’s publications
Sharing your data and publishing your data paper

The dataset must be stored in an open access repository:
- JOHD Dataverse
- Zenodo
- Figshare
- SND
- DANS

Image created by Amanda Sorensen
The special collections

- **New perspectives and practices in Digital Humanities Data**: 3 articles
- **Humanities data in the time of COVID-19**: 5 articles
- **Language documentation collections: Assessment and Recognition**: 2 articles
- **Computational Humanities Research Data**: 9 articles

JOHD accepts submissions to the general issue on a rolling basis
Disciplinary scope

Credits: Eleonora Zordan
Conferences and events

- LIBER conference (session #3 “Working with Software and Data”)
- Round table of experts about data citation, organised by SSHOC (Social Sciences & Humanities Open Cloud)
- #dhnord2021 about the publication of data papers in the humanities
- DHASA conference 2021 (Digital Humanities Association of Southern Africa)
- SciDataCon-IDW Seoul 2022 about the impact of data papers and datasets
- Humanities and Data Science Discussion Group at the Alan Turing Institute
- RIOT science club
- Oxford Open Humanities Seminar Series
The #showmeyourdata campaign

What is it?

- We ask our authors to post an image of their data and describe it. It can be the screenshot of (any type of) dataset, a visualisation etc.

What for?

- Our aim is to promote open research and data sharing values: we would like to encourage the conversation on this topic and demystify it.
Other activities

**Open Humanities Data Forums**: promoting values of data sharing and reusability in the humanities

Data papers and impact

Credits: Nilo Pedrazzini
Data papers and impact

Credits: Nilo Pedrazzini
Data papers and datasets

Credits: Nilo Pedrazzini and Marton Ribary
Keep in touch...

@up_johd
#johdpapers and #johdCfP: newly published papers and call for papers
#johdnews: news about JOHD (activities and events)
#johdagenda: external events about open data, data sharing and related matters
#showmeyourdata: share an image of your data and describe it

Newsletter:
Every 6 months, an update on all our activities, publications, new members of our team

Recordings of the events organised by JOHD (e.g. Open Humanities Data Forums)
Recordings of presentations to conferences and events
Short videos on key topics for JOHD e.g. how to define humanities data?
Thank you!

Barbara.mcgillivray@kcl.ac.uk