Towards a linked open data resource for direct speech acts in Greek and Latin epic

Forstall, C.W.; Finkmann, S.; Verhelst, B.

DOI
10.1093/llc/fqac006

Publication date
2022

Document Version
Final published version

Published in
Digital Scholarship in the Humanities

Towards a linked open data resource for direct speech acts in Greek and Latin epic

Christopher W. Forstall
Department of Classics, Mount Allison University, Sackville, New Brunswick, Canada

Simone Finkmann
Heinrich Schliemann-Institut für Altertumswissenschaften, Universität Rostock, Germany

Berenice Verhelst
Amsterdam School of Historical Studies, University of Amsterdam, Amsterdam, the Netherlands

Abstract

The Digital Initiative for Classics: Epic Speeches (DICES) research group reports here on preliminary work to integrate research on Greek and Latin epic speeches into the larger ecosystem of linked open data (LOD) for classical scholarship. The ability to collate speech data from different researchers and to leverage external repositories of texts and characters opens up new possibilities for interrogation of the epic corpus. We briefly survey the current state of scholarship on epic speeches and of the digital infrastructure on which we propose to build. We outline a model for harmonizing speech data across studies and aligning it with existing LOD standards. Finally, we discuss some early proof-of-concept results and the larger implications of this approach for the field. The long-term aim of the DICES project is to build a database of metadata on direct speech in Greek and Latin epic, not only covering canonical texts such as Homer and Virgil, but also including the less-studied texts of the late antique period, which will benefit greatly both from the increased accessibility and also from the diachronic perspective afforded by a corpus-based approach. The envisioned database also has the potential to include diachronic data from additional genres and languages at a later stage.

Correspondence:
Christopher W. Forstall, Mount Allison University, Sackville, New Brunswick, Canada.
E-mail:
cforstall@mta.ca

1 Background

Direct speech is a significant component of most narrative genres. In the field of Classics, scholars have taken particular interest in the speeches of Greek and Latin epic poetry. Whether characters meet in battle or in the bedroom, they have a lot to say to each other and distinctive ways of expressing themselves. Yet, despite the established importance of speeches for ancient epic, previous studies have been based on scholars' personal datasets, whose details are often unavailable to the public. Among quantitative studies, Elderkin (1906) and Lipscomb (1909) remain the most comprehensive with respect to the corpus considered, but lack transparency with regard to their calculation methods and classification

https://doi.org/10.1093/llc/fqac006 Advance Access published on 16 March 2022

system. Highet (1972), Dominik (1994), and especially the online databases by Beck (2012) (https://homeric-speech-beck.la.utexas.edu/) and Verhelst (2016) (https://www.dsgep.ugent.be/) are notable exceptions. However, even these are presented in proprietary, standalone formats, rendering the data difficult for users to collate or compare—an essential consideration for new digital projects. The present work aims to address this vacancy by establishing new standards for the study of direct speech representation in ancient and late antique epic with an open, aggregated dataset and an interface designed to facilitate reuse.

The growing ecosystem of linked open data (LOD) initiatives in Digital Classics provides additional incentives to standardize and expand access to existing speech data. The use of LOD is one part of a broader shift in the discipline towards interoperability of resources and collaborative modes of research, with emergent and sometimes unforeseen benefits (on the importance of collaborative research efforts for the field of Latin Computational Linguistics see McGillivray, 2014, esp. 10–11). Alignment of the envisioned database with this paradigm grants our users access to an extensive existing infrastructure while also making a significant contribution to the corpus of LOD. Particularly important for our purposes is the canonical text services (CTS) protocol (Blackwell and Smith, 2014), which describes a system of uniform resource names (URNs) as standard identifiers for ancient texts. Combined with the Text Encoding Initiative (TEI) guidelines for representation of texts (TEI Consortium, 2020), CTS supplies a standardized framework for core classical corpora like the Perseus Digital Library (https://scaife.perseus.org). Standards for the representation of entities referring to people (Bodard *et al.*, 2017; http://snapdrgn.net), places (Simon *et al.*, 2015; https://pleiades.stoa.org), and dates (https://godot.date), as well as for representing the formal relationships between these data, allow cross-referencing and exploration across a broad range of information about the ancient world. Yet, while standards already exist for historical persons in ancient prose texts, the recorded details may not be as pertinent (or accurate) for fictionalized versions of these characters in historical epics (e.g. Caesar in Lucan's *Bellum Civile* or Hannibal in Silius Italicus' *Punica*). Even more challenging are the mythological characters of epic, which to this point have not been robustly covered by databases of historical persons. General-purpose semantic web resources such as WikiData (https://www.wikidata.org) include many, but by no means all, of these characters and lack a domain-specific ontology of the attributes and relations peculiar to them. This need is now being addressed by MANTO (Hawes and Smith, 2021; https://www.manto-myth.org), whose efforts to map Greek myth include a prosopographic database still under development, but with preliminary data from both poetic and mythographic sources.

Significant advances have been made in the integration of primary textual resources both with analytical tools including parsers, indices, and commentaries, as well as with published scholarship. For example, the Perseids Project (https://www.perseids.org/) combines reading and research tools with digital texts in an open infrastructure, while independent research tools such as Tesserae (https://tesserae.caset.buffalo.edu) are adopting service-based models facilitating integration into digital libraries (Coffee, 2018). Matteo Romanello's work on identifying and linking canonical citations in print (e.g. Romanello, 2015) has been used not only to connect new e-publications such as Reitz and Finkmann (2019) back to the primary sources they cite (http://epibau.ub.uni-rostock.de/app/, Romanello et al., 2020), but also to link digital primary texts forward to existing scholarship (http://aeneid.citedloci.org/).

In what follows, we describe an outline for and example data from a new project that integrates the study of epic speeches into the larger context of Digital Classics, utilizing and expanding upon the already extant LOD resources. This involves unifying existing speech data in a new, open-access database with a public Application Programming Interface (API), designing a scholar-oriented code library to provide connectivity between the speech data and related LOD resources, and collating additional data in order to cover as much as possible of the extant corpus of Greek and Latin epic, with particular emphasis on less-studied late antique texts.

2 Project Structure

The first stage of this project comprises the collation and classification of all relevant metadata on direct

speech in the most prominent Greek and Latin epics (Homer, *Iliad* and *Odyssey*; Apollonius, *Argonautica*; Virgil, *Aeneid*; Ovid, *Metamorphoses*; Lucan, *Bellum Civile*; Valerius Flaccus, *Argonautica*; Statius, *Thebaid*; Silius Italicus, *Punica*; Quintus Smyrnaeus, *Posthomerica*; Nonnus, *Dionysiaca*). These texts contain approximately 4,000 instances of direct speech (embedded speeches, where one character directly quotes another, are counted separately). For the purpose of our study, we define the epic genre in the broadest sense as all complete or almost complete narrative hexameter poems on any topic and of any size, including biblical epics and shorter poems such as *epyllia* and hymns, but not fragmentary epics. The targeted corpus of ancient and late antique epic is estimated to include approximately fifty works. An authoritative, shared dataset of this size, particularly when paired with an open API and standard, machine-readable references, will increase the connectivity between researchers and support complex research questions at new scales, as illustrated by the example workflow discussed below.

We have demonstrated the viability of this model with a proof-of-concept prototype and preliminary speech data for Homer, Apollonius, Virgil, and Ovid (https://github.com/cwf2/dices). Data are served by a PostgreSQL backend via the Python web framework Django, in conjunction with the Django REST framework. This provides a standard, web-based API, upon which we can layer specific functionality for interaction with the database, such as parameterized search for speeches or characters. While Django also provides a human-oriented web interface for browsing and faceted search, our eventual goal is a more full-featured site built on top of the API and designed for classical scholars and students, some of whom may be only modestly familiar with the use of databases and digital tools.

A primary goal of the prototype is to show how external LOD can be leveraged to augment the speech data and enable new lines of questioning (Fig. 1). The resources of interest here are digital libraries, which provide textual services and bibliographic metadata, and indices of mythological characters—the speakers and addressees of our speech records. We use Perseus for text services via the CTS server at https://cts.perseids.org and the *MyCapytain* library for Python (https://mycapytain.readthedocs.io). For character data, we currently use WikiData via the *qwikidata* library (https://qwikidata.readthedocs.io). At the same time, we are working with MANTO to develop a similar interface for their character data, as their models for relationships among mythical persons are more finely tuned to the epic context and their data are more closely controlled by domain experts. In addition to these text and character repositories, our prototype pipeline also makes use of *CLTK* (Johnson *et al.*, 2014), a library that provides natural language processing (NLP) tools for ancient languages.

3 Speeches and Their Contexts

In our data model (Fig. 2), the entity of primary interest is the speech. Other entities are the works in which the speeches occur, their authors, the characters involved as speakers and addressees, as well as speech clusters. We also employ a pseudo-entity, the character instance, representing a character's ephemeral attributes specific to that time, in order to account for transtextual characters, i.e. 'characters that exist in more than one text' (Richardson, 2010, p. 527), and the generally dynamic nature of epic character concepts.

Each speech consists of a sequence of contiguous lines in a given poem, and represents words spoken by one character to another. Core textual attributes recorded for each speech are the author and work, and the first and last lines included in the speech. Although full bibliographic information for the work is encoded in the CTS URNs, authors and works are recorded as separate entities to simplify common queries. Additional attributes describe the larger context in which each speech is situated. Speeches are grouped into speech clusters, entities representing larger conversations in our model. These are classified by type (soliloquy, monologue, dialogue, general interlocution) and the position of each speech within the cluster is recorded as an integer. A second integer records embeddedness, the degree to which the speech is nested within enclosing speeches.

4 Character Identities and Instances

The addresses and speakers are modelled as character 'instances,' examples of a particular character in a

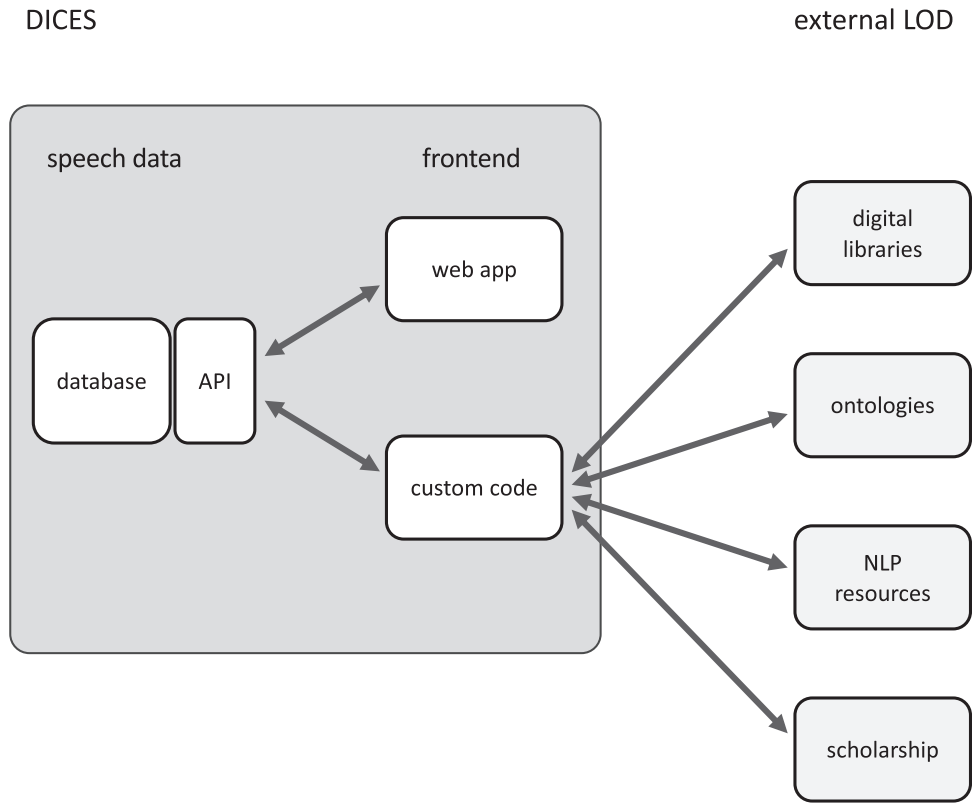


Fig. 1 Schematic representation of the epic speech database DICES within a larger ecosystem of external LOD resources for ancient Greek and Latin

particular context. Every identifiable speaker and addressee in the corpus is first represented as a 'character' entity, reflecting their abstract or essential nature: attributes recorded in this table are core characteristics such as entity type (e.g. god, mortal, personification, animal), number (singular, tandem, collective), and gender, along with external LOD identifiers. The number of characters in the full dataset is not yet known; our best estimate is that it will exceed 1,000.

But characters' traits often change. Therefore, a separate set of character instance entities capture a given speaker's or addressee's relevant attributes in the context of a particular speech. For example, Achilles is alive when he speaks to Odysseus at Hom. Il. 9.308, but dead when he speaks to him at Hom. Od. 11.473. Almost any feature of a character can change at some point in our corpus: e.g. the character who speaks as a woman named Caenis at Ov. Met. 12.201 speaks as a man named Caeneus at 12.490. Character instance attributes may override attributes specified at the character level, like gender or entity type, or may specify attributes that have no clear meaning at the character level, such as age or health. An important example here is the 'disguise' attribute, which will allow us to specify a false (or presumed) as well as a true character ID.

Particular challenges for this system are anonymous characters, collective identities, and especially passages in which the identity of the speaker is ambiguous or uncertain. In these cases, the character instance may not be linkable to a well-defined character and thus will have no character ID; the database will provide supplementary descriptive attributes, such as an 'anonymous' flag, where possible. Incomplete speech acts, the occurrence of a lacuna, and other problems in a direct speech, as well as important variant readings will be treated similarly.

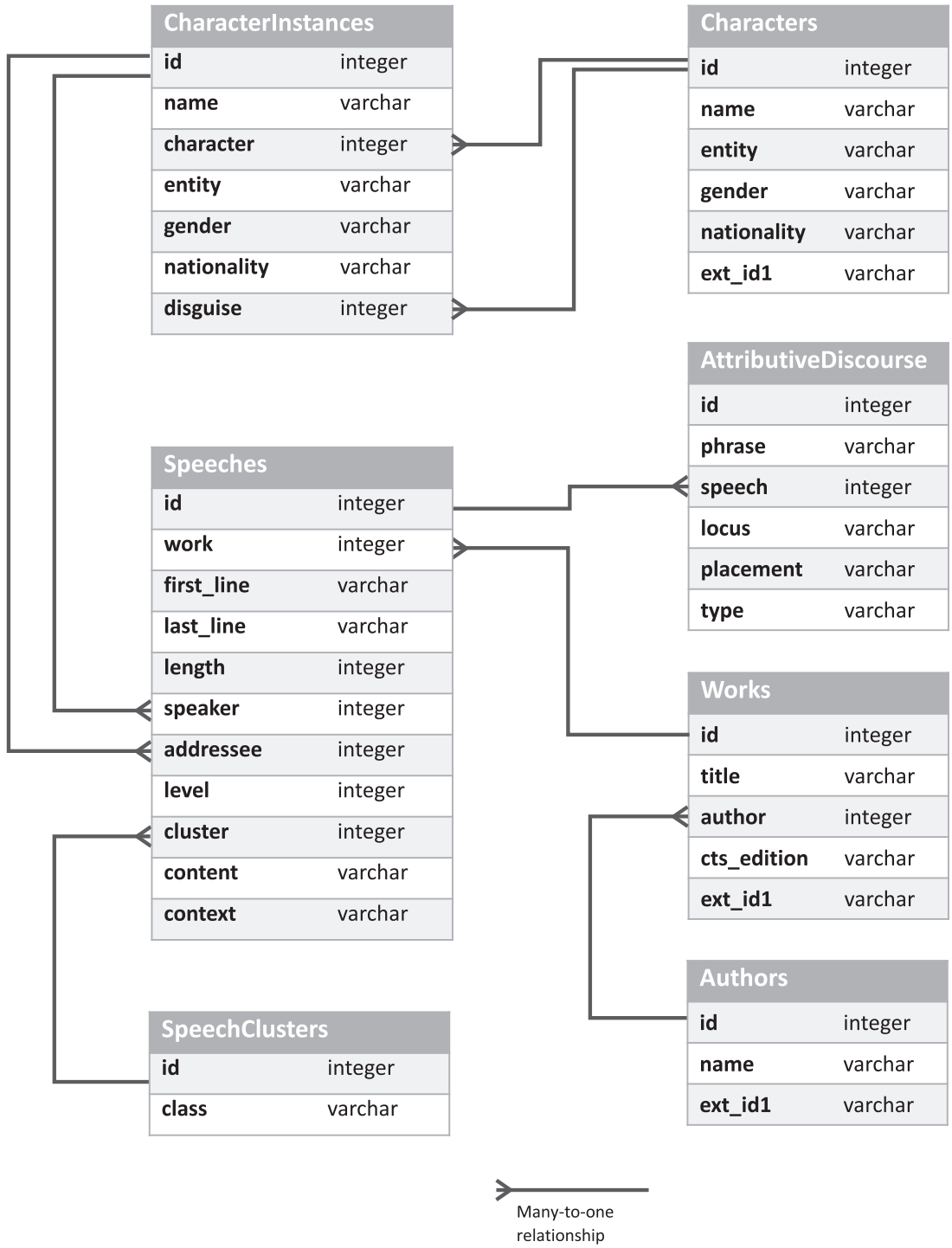


Fig. 2 Entity–relationship model for the prototype database. Lines show foreign key relationships between the tables, with 'crow's foot' connectors indicating directionality of many-to-one relationships

5 Example Workflow

The remainder of this report outlines one example of the database's use. Complete code for this and other sample experiments is available on GitHub (https://github.com/cwf2/dices-examples), and we are hopeful that future users will contribute their own examples to the repository.

Early in Homer's *Odyssey*, Penelope asks that the court singer Phemius refrain from singing of the Trojan War, which only makes her feel her husband's absence more keenly (Hom. Od. 1.337–344). To her surprise, her son Telemachus rebukes her (1.346–359), insisting that it is proper to recall his father's deeds and demonstrating his own growing self-confidence and authority. For a scholar studying this interchange (e.g. Clark, 2001), the affordances provided by our model significantly expand the scope of computational analysis over current digital resources (e.g. Beck, 2012; Verhelst, 2016). For example, Beck's proprietary search interface returns a paged list of speeches for manual inspection (Fig. 3); Verhelst's provides tabular results but without the text of the speeches (Fig. 4). In either case, applying NLP tools to the speeches themselves would require significant manual intervention.

The Digital Initiative for Classics: Epic Speeches (DICES) API, in contrast makes it straightforward not only to search the database for a particular speaker–addressee pair, such as Telemachus and Penelope, but also to bulk-download the results as JSON records. Since the records use CTS, retrieval of the passages themselves from a remote library like Perseus is easily automated. These steps are further simplified by the client library, which returns search results as ready-to-use Python objects and provides a

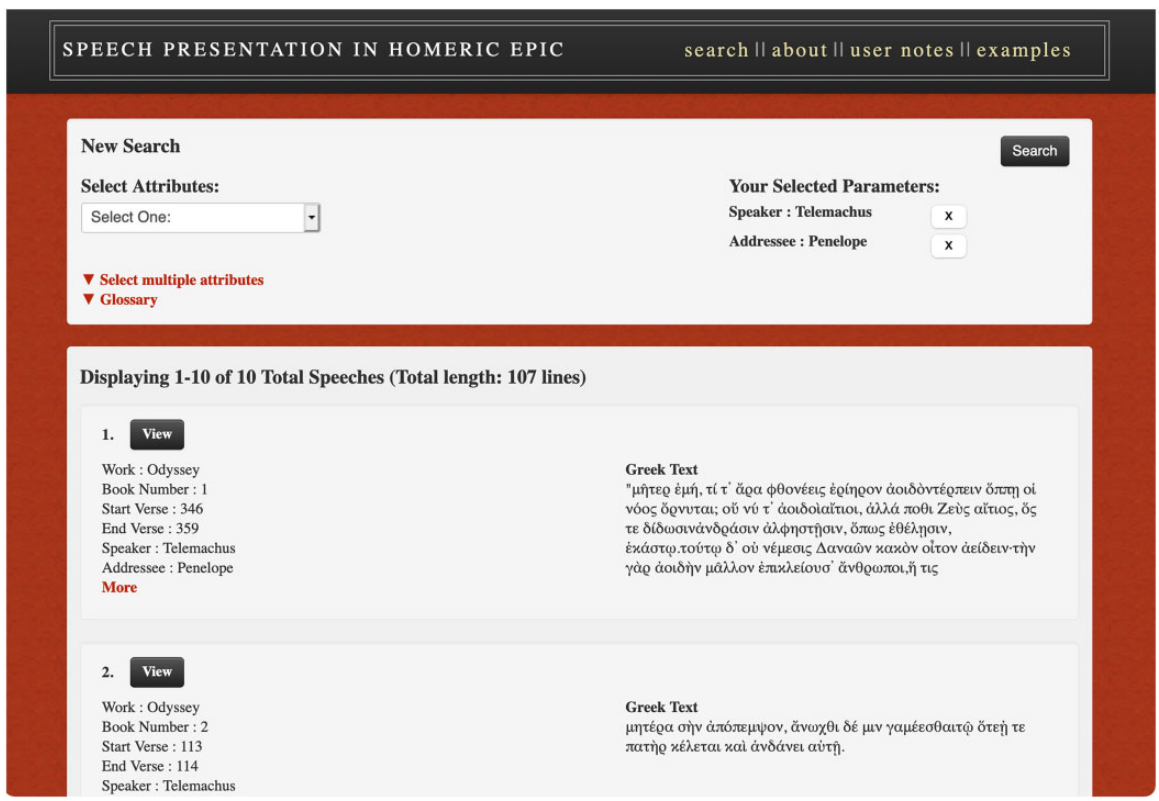
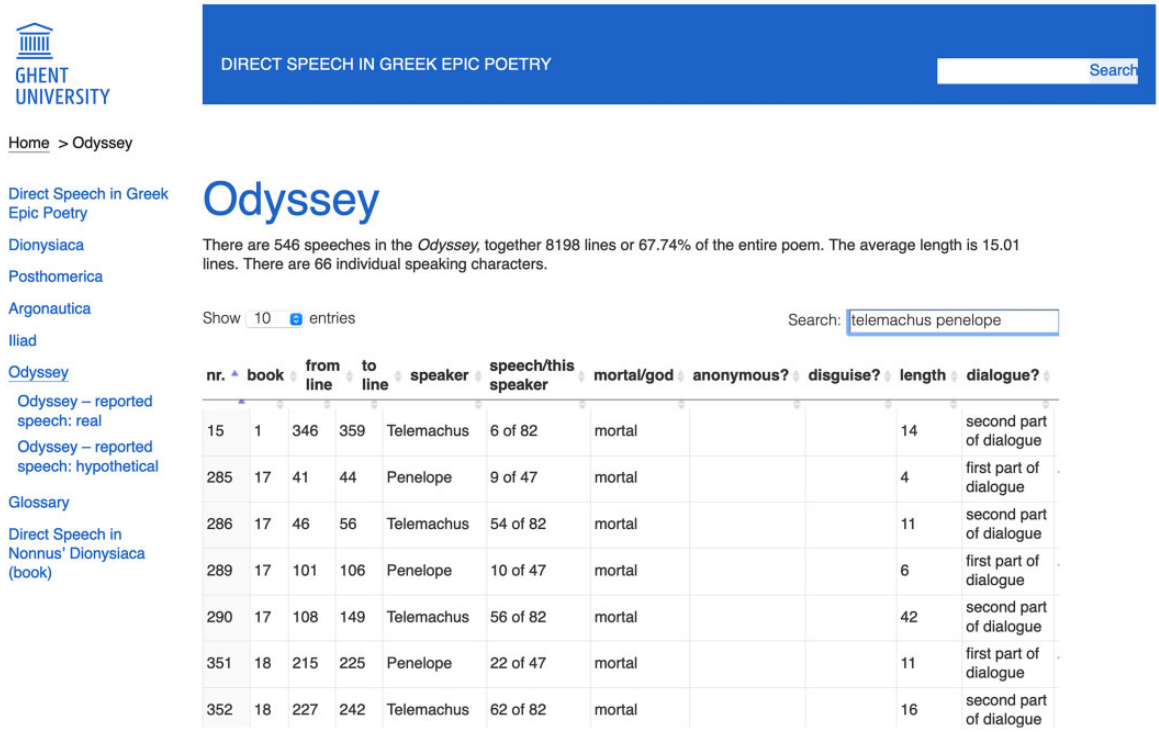


Fig. 3 The digital appendix to Beck (2012), here demonstrating a search for speeches in which Telemachus addresses Penelope. The platform is reader-friendly, prioritizing faceted search and manual inspection of the text, but does not easily support automated queries or additional processing and analysis of the returned text


GHENT UNIVERSITY

DIRECT SPEECH IN GREEK EPIC POETRY

Search

Home > Odyssey

Direct Speech in Greek Epic Poetry
Dionysiaca
Posthomerica
Argonautica
Iliad
Odyssey
Odyssey – reported speech: real
Odyssey – reported speech: hypothetical
Glossary
Direct Speech in Nonnus' Dionysiaca (book)

Odyssey

There are 546 speeches in the *Odyssey*, together 8198 lines or 67.74% of the entire poem. The average length is 15.01 lines. There are 66 individual speaking characters.

Show 10 entries

Search: telemachus penelope

nr.	book	from line	to line	speaker	speech/this speaker	mortal/god	anonymous?	disguise?	length	dialogue?
15	1	346	359	Telemachus	6 of 82	mortal			14	second part of dialogue
285	17	41	44	Penelope	9 of 47	mortal			4	first part of dialogue
286	17	46	56	Telemachus	54 of 82	mortal			11	second part of dialogue
289	17	101	106	Penelope	10 of 47	mortal			6	first part of dialogue
290	17	108	149	Telemachus	56 of 82	mortal			42	second part of dialogue
351	18	215	225	Penelope	22 of 47	mortal			11	first part of dialogue
352	18	227	242	Telemachus	62 of 82	mortal			16	second part of dialogue

Fig. 4 The digital appendix to Verhelst (2016), demonstrating a similar search to that in Fig. 3. These results are slightly more amenable to bulk treatment, for example copy–pasting into a spreadsheet, but the text of the speeches is not provided

built-in method for retrieving CTS passages. CTS can also be used to traverse the text outside the bounds of speeches recorded by DICES, so scholars interested in introductory or concluding formulae could, for example, use the DICES API to search for speeches of interest and then use CTS to download the text directly following each speech. Once collected, the Greek and Latin text can be piped to NLP tasks like tokenization and lemmatization using CTLK; the DICES client, being written in Python, is easily inserted into existing CLTK workflows. The ability to move seamlessly from search criteria to collated text ready for analysis (whether by custom code or close reading) makes literary studies easier and more replicable by cutting out much of the potentially opaque manual data manipulation and tallying traditionally required (e.g. phrases such as 'according to my count. . .,' Clark, 2001) and will generate valuable research data for scholars working in affiliated fields such as narratology, linguistics, and conversation analysis.

But the real power of the model is its ability to generalize questions at larger scales. How do Telemachus' words to Penelope compare to other mother-and-child exchanges? Here, we can leverage the power of external ontologies to filter the speech data based on the relationship between speaker and addressee. Using ancillary data on the characters collected by WikiData or MANTO, we can programmatically ask whether a given addressee is the mother of a given speaker. Instead of searching DICES for a given speaker–addressee pair, we can search for works or authors to download a large corpus: for example, all speeches in Homer, Apollonius, and Virgil. Filtering the results based on a WikiData query for the relation 'mother' produces an exhaustive set of child–mother speeches in the selected corpus. Although this set is much larger than the six Telemachus–Penelope speeches, the same pipeline can be used to download and parse the text with no additional effort (Fig. 5). The list of lemmata used by Telemachus with his mother can now be compared with those in all

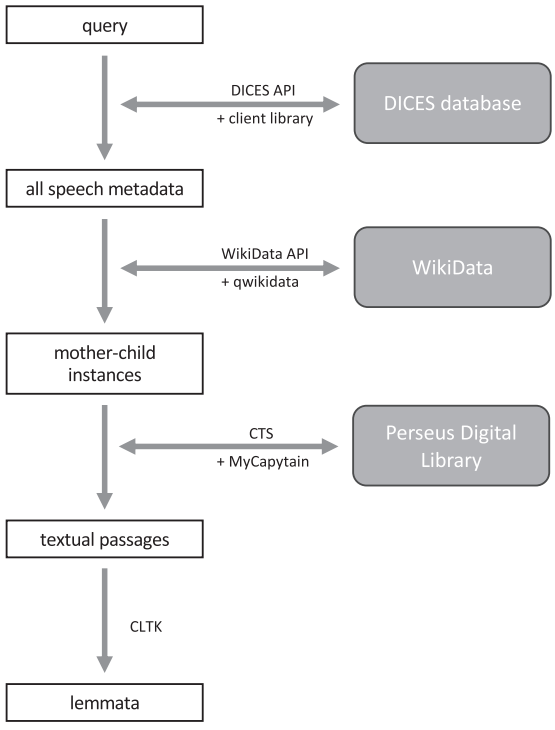


Fig. 5 A sample workflow for retrieving all speeches between mothers and their children. DICES is first queried for a large corpus of speeches using the DICES client library. Speakers and addressees are cross-referenced against WikiData to determine which speaker–addressee pairs have the correct relationship. The text of each matching speech is then retrieved from Perseus before being parsed using CLTK. Shaded elements represent external resources. Unlabelled arrows represent custom code

child–mother speeches in the epic corpus, providing a much broader baseline against which to judge this specific interaction.

As a test of the prototype's performance and interoperability with external LOD resources, the results of an automated search for all speeches between mother and child (in either direction) in Homer's *Iliad* and *Odyssey*, Apollonius Rhodius' *Argonautica*, and Virgil's *Aeneid* were compared with a hand-curated benchmark (summarized in Table 1). Of the 1,855 speeches, 61 were manually identified as between a mother and her child. The automated workflow returned 56 of these, for a recall rate of 92%, with no false positives.

Table 1. Comparison of preliminary results from an automated analysis of mother–child speeches using the prototype DICES database with a benchmark set of hand-curated data

Benchmark data	
Total speeches	1,855
Mother–child	61
Automated method	
Total results	56
True positives	56
Success rate, %	
Recall	92
Precision	100

Table 2. Mother–child interactions missed by the automated process

Work	Lines	Speaker	Addressee
Iliad	15.104–112	Hera	Gods
Iliad	15.115–118	Ares	Gods
Aeneid	9.83–92	Cybele	Jupiter
Aeneid	9.94–103	Jupiter	Cybele
Aeneid	9.481–497	Euryalus' mother	Euryalus

6 Discussion

Five speeches were missed by the mother–child filter (Table 2). Three of these instances can be attributed to deficiencies in the current data model. Two speeches in the *Iliad* between the goddess Hera and the other Olympian deities (Hom. Il. 15.104–112 and 15.115–118), among whom were some of her children, failed to register as mother–child interactions because one party was a collective. While 'the gods' is a corporate identity, it is not currently treated as including other specific characters (though both WikiData and MANTO offer limited support for identifying membership within collectives). A speech by the mother of Euryalus to her son in the *Aeneid* (Verg. Aen. 9.481–497) was missed because as an anonymous character instance she has no WikiData or MANTO IDs, which are recorded only for identifiable characters. This technical failure is illuminating, given that ironically the one thing we know about her is her relationship to Euryalus (on the mother of Euryalus, see Sharrock, 2011).

The remaining two cases, speeches in the *Aeneid* between the god Jupiter and his mother (Verg. Aen. 9.83–92 and 9.94–103), highlight difficulties in dealing with external data. In this passage Virgil identifies Rhea, the Titan mother of the Graeco-Roman gods, with the Phrygian goddess Cybele, according to a common Roman syncretism. WikiData considers the two to be separate entities from different mythologies and returns no relationship between Cybele and Jupiter. While this case might be resolved by manually altering the character record to match her Greek identity, it illustrates the larger need for a domain-specific repository of mythological characters, supported by an ontology sensitive to epic conventions. That MANTO is working to publish just such data is therefore particularly exciting.

We continue to develop our prototype with the aim that it will render future quantitative work in this field more transparent and replicable, and advance the study of discourse analysis and narratology by facilitating a great variety of synchronic and diachronic approaches to speech representation. We hope to introduce students of Greek and Latin epic to discourse analysis as well as digital tools and methods, and at the same time, to enable experts of speech representation to query, analyze, and export the data necessary to pursue an infinite number of quantitative research questions across Greek and Latin epics from Homer to Late Antiquity. The open API of our database and the code repositories for all experiments moreover increase the potential for future extensions of its tools and methods, e.g. for the study of direct speech in other periods, languages, and genres. The DICES group is already in talks with research groups focusing on discourse analysis in other genres of Latin and Greek literature to discuss a potential expansion of the content covered by the database.

Acknowledgements

The authors thank their student research associates, Jule Andreeßen, Dawson Fraser, and Ava Waugh for their support and the anonymous reviewers for their valuable comments.

Funding

Initial work on the prototype database was supported by a President's Research and Creative Activities grant from Mount Allison University, including funding from the J. E. A. Crake Foundation, by a research grant from the Interdisciplinary Faculty (Department WKT: Wissen-Kultur-Transformation) at the University of Rostock, by the European Social Fund (ESF) for Germany, and by the Research Foundation Flanders (FWO, grant nr. 12N6819N). Continued development of the database for direct speeches in Greek and Latin epic from Homer to Late Antiquity is supported by an Insight Development Grant (1 June 2021 to 31 May 2023, grant nr. 430-2021-00977) of the Canadian Social Sciences and Humanities Research Council (SSHRC).

References

Beck, D. (2012). *Speech Presentation in Homeric Epic*. Austin: University of Texas Press. Digital Appendix: https://homeric-speech-beck.la.utexas.edu (accessed 15 July 2021).

Blackwell, C. and Smith, N. (2014). The canonical text services URN specification, version 2.0.rc.1. http://cite-architecture.github.io/ctsurn_spec/ (accessed 15 July 2021).

Bodard, G., Cayless, H., Depauw, M., Isaksen, L., Lawrence, K. F., and Rahtz S. (2017). Standards for networking ancient person data. Digital approaches to problems in prosopographical space. *Digital Classics Online*, **3**: 28–43.

Clark, M. (2001). Was Telemachus rude to his mother? Odyssey 1.356–359. *Classical Philology*, **96**(4): 335–54.

Coffee, N. (2018). An agenda for the study of intertextuality. *Transactions of the American Philological Association*, **148**: 205–23.

Elderkin, G. W. (1906). *Aspects of the speech in the Later Greek epic*. Baltimore, MD.

Dominik, W. (1994). *Speech and Rhetoric in Statius' Thebaid*. Hildesheim: Olms-Weidmann.

Hawes, G. and Smith, S. (2021). A dataset of mythical people with stable URIs. MYTHLAB. https://www.manto-myth.org/blog/a-dataset-of-mythical-nameswith-stable-uris (accessed 15 July 2021).

Highet, G. (1972). *The Speeches in Vergil's Aeneid*. Princeton: Princeton University Press.

Johnson, K. P., Burns, P., Stewart, J., and Cook T. (2014–2021). *CLTK: The Classical Language Toolkit*. https://github.com/cltk/cltk (accessed 15 July 2021).

Lipscomb, H. C. (1909). *Aspects of the speech in the later Roman epic*. Diss. Baltimore, MD.

McGillivray, B. (2014). *Methods in Latin Computational Linguistics*. Leiden: Brill.

Reitz, C. and Finkmann, S. (eds) (2019). *Structures of Epic Poetry*, 4 vols. Berlin: De Gruyter.

Richardson, B. (2010). Transtextual characters. In Eder, J., Jannidis, F. and Schneider, R. (eds), *Characters in Fictional Worlds. Understanding Imaginary Beings in Literature, Film, and Other Media*. Berlin: De Gruyter, pp. 527–41.

Romanello, M. (2015) *From Index Locorum to Citation Network. An Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts*. Ph.D. thesis, King's College London.

Romanello, M., Finkmann, S., and Reitz, C. (2020). EpiBau: a database of ancient epic. http://epibau.ub.uni-rostock.de/app/ (accessed 15 July 2021).

Sharrock, A. (2011) Womanly wailing? The mother of Euryalus and gendered reading. *EuGeSta*, **1**: 55–77.

Simon, R., Isaksen, L., Barker, E., and de Soto Canamares, P. (2015). The Pleiades gazetteer and the Pelagios project. In Berman, M. L., Mostern, R. and Southall, H. (eds), *Placing Names: Enriching and Integrating Gazetteers*. Bloomington: Indiana University Press, pp. 97–109.

TEI Consortium (eds). (2020). Guidelines for electronic text encoding and interchange. Last updated 19 August 2020. http://www.tei-c.org/P5/ (accessed 15 July 2021).

Verhelst, B. (2016). *Direct Speech in Nonnus' Dionysiaca*. Leiden: Brill. Digital Appendix: http://www.dsgep.ugent.be (accessed 15 July 2021).