Latin Represented Speech (LaReS): Linking LiLa and the DICES database

Francesco Mambrini
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123 Milan
francesco.mambrini@unicatt.it

Abstract

This paper presents LaReS (Latin Represented Speech), a Linked Open Data resource designed to model represented speech in Latin literature and to align the DICES database of direct speeches in Greek and Latin epic with the LiLa Knowledge Base. While DICES provides a rich collection of metadata on direct speech in epic poetry, its operational approach and its relatively shallow conceptual modeling limit its interoperability and extensibility. The modeling strategy implemented in LaReS is based on the separation of the textual level from the narratological dimension. CIDOC CRM and DOLCE+DnS are used to conceptualize the basic notions in the two modules. LaReS now includes 341 speech instances in the *Aeneid*, linking 36,782 tokens in LiLa to speech units derived from DICES.

Keywords: Latin, Narratology, Represented Speech

1. Introduction

Speech representation in literary texts is among the most fascinating and debated topics in criticism (McHale, 2011). Within the history of the Digital Humanities (DH), John F. Burrows's (1987) study of linguistic characterization in Jane Austen, conducted through a stylometric analysis of direct speeches, remains a landmark example of how traditional literary questions can be effectively addressed through computational approaches.

Burrows's research relied on the granular annotation of Austen's novels: instances of direct and (free) indirect speech were systematically marked in digitized texts and distinguished from narrative passages. In addition, speech segments were attributed to individual characters, so that distributional statistics of lexical patterns across speakers could be computed.[1]

More recently in the field of Classical Studies, a group of researchers from the universities of Mount Allison, Amsterdam and Rostock launched the "Digital Initiative for Classics: Epic Speeches" (DICES). The goal of the initiative was twofold. Firstly, it aimed to construct an open database of passages with direct speech in Greek and Latin epic poems (Forstall et al., 2022). Secondly, the project team assembled a network of scholars interested in using the data to promote innovative research on the subject of speech representation and characterization; the outcome was published in Forstall and Verhelst (2025a).

The fundamental architectural choice that was adopted to build the DICES database (DB) was to collect metadata about the passages, instead of their text. With a design choice that is inspired by the Linked Open Data (LOD) paradigm, the records about the speech instances are connected via persistent identifiers to the texts, but also to a wider network of information about historical data. DICES relies on the Canonical Text Service protocol and its system of flexible identifiers, the CTS-URNs (Smith, 2009; Tiepmar and Heyer, 2019), to reference the relevant loci and allow users to retrieve them from digital libraries. Furthermore, the DB incorporates URIs from authority datasets such as Wikidata or MANTO, the digital resource for Greek myth (Hawes and Smith, 2021), to align the literary characters participating in instances of direct speech with the historical and mythological figures to which the characters refer, or from which they are derived.

Extraction of linguistic information or even NLP tasks such a Named Entity Recognition and Sentiment Analysis were explicitly envisaged as a use case for the collected data (Forstall and Verhelst, 2025b, 30). However, no connection other than that to the full text via CTS-URN is implemented. Considering that DICES now includes metadata on 1,915 passages from 20 Latin works, one connection that appears particularly fruitful is that to the LiLa Knowledge Base (KB) of Latin linguistic resources (Passarotti et al., 2020). Linking the two dataset would enable researchers to pursue lexicon-based sentiment analysis of the speeches (Sprugnoli et al., 2023), or to answer sophisticated questions, such as what the distribution of derivational morphemes is in the speeches of female vs male speakers. The metadata about the gender of the speaker is indeed recorded in DICES, while information about derivational morphology is stored

[1] Burrows' annotation of the six canonical novels is now available via the Oxford Text Archive; see, for example, Austen (1988).

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 50–59
12 May 2026.

in LiLa (Pellegrini et al., 2022).

This paper describes an initiative that originated from the idea of connecting DICES to the tokens in the corpora linked to LiLa. Considering the potential for expansion and reuse of this dataset, in light of the multiple genres and possible different approaches to direct speech and related phenomena, however, we decided to broaden our perspective. We built a new textual resource dedicated to represented speech, called Latin Represented Speech (LaReS), which is linked to and (for the present) entirely based on the DICES data, as well as to the tokens in LiLa's corpora. The creation of LaReS entailed a considerable amount of conceptual work on modeling for our domain.

The paper is organized as follows. Section 2 describes DICES and the LiLa KB (2.1) and lists some relevant previous works (2.3). The alignment process with LiLa contributed to highlight some limitations inherent in DICES, which are discussed in 2.2. In order to keep the useful connection to the DICES data but also to overcome some of those limits, we decided to: 1. sketch a preliminary draft of a conceptual model for represented speech in literature, and 2. use it to model the data about the passages attesting represented speech available in both DICES and LiLa. The main principles behind the conceptual model are discussed in Section 3, while the results are presented in Section 4. Section 5 discusses the open problems and future directions.

2. Methodology

2.1. The data: DICES and LiLa

The DICES database is a structured collection of passages displaying direct speech in Greek and Latin epic poetry. While the genre is constrained to epos, the promoters attempted to expand the canon to include a variety of texts from the earliest surviving Greek literary works (the Homeric poems) to Late Antiquity (Forstall and Verhelst, 2025b). The core idea behind the DB is to create one record for each instance of a direct speech, attaching minimal information to it, such as the starting and ending line number and who is speaking to whom.

The database is publicly available online.[2] The data can be programmatically accessed via a dedicated API that, although not accompanied by formal documentation, adopts the standard structure of Django REST Framework.[3] Figure 1 represents a simplified illustration of the DB structure, its tables

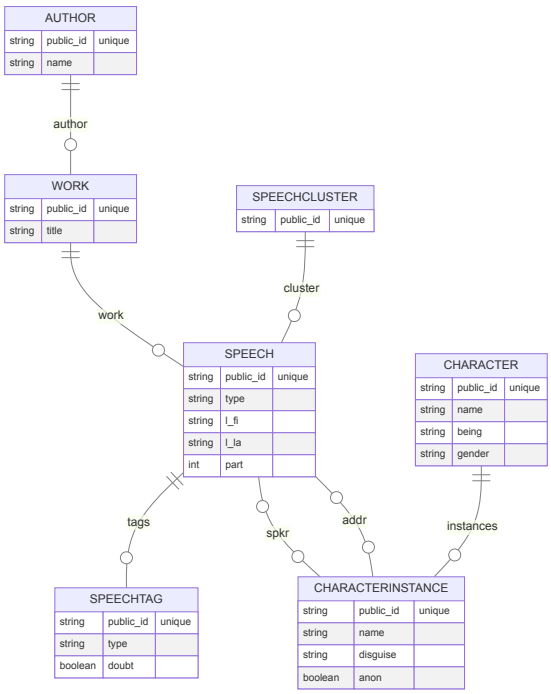

Figure 1: Simplified structure of the DICES database.

and the properties for entries in each of them; some properties in the tables are omitted for brevity. As visible, the central element in the architecture is the "Speech" table, which collects the passages in the Greek and Latin works that are labeled as instances of direct speech.[4]

At the moment, DICES collects 4,689 instances of direct speech, from 27 authors and 52 works (32 in Greek, 20 in Latin). Metadata that is needed to generate a CTS-URN, like initial and final line number, is stored in the DB. Some of the speech passages are grouped in "speech clusters" (2,965 records), which represent larger conversational units made of several turns, such as questions and answers, that are explicitly numbered.

For each speech, the identity of the speaker and addressee is marked. The attribution references an entry in the "character instance" table (2,236 entries, visible in the bottom-right corner of Fig. 1), i.e. a figure identified locally within each work. Figures that are culturally perceived as avatars of the same character are linked to entries in a "character" table (997 entries). The articulation between instances and characters allows users to retrieve all speeches by a particular figure (e.g. the Greek goddess Hera and the Roman Juno, who were identified in Latin

[2] https://db.dices.mta.ca/. An archival copy of the DB, exported in CSV format, is available in Forstall et al. (2025).

[3] https://www.django-rest-framework.org/.

[4] The diagram was reconstructed from the relevant Django modules published by DICES's main developer: https://github.com/cwf2/dices/blob/main/speechdb/models.py.

literature) across all works.[5]

The LiLa KB is a network of textual and lexical resources in Latin modeled as Linked Open Data (Passarotti et al., 2020). The LiLa Lemma Bank (Mambrini and Passarotti, 2023) is a collection of more than 230,000 canonical forms that are used as lemmas to index lexical entries and to lemmatize texts (Moretti et al., 2023); it functions as the central element that keeps all other resources connected. LiLa leverages a series of widely used ontologies for Linguistic Linked data to model language resources as RDF. In particular, LiLa relies on Ontolex-Lemon (McCrae et al., 2017) to represent lexical entries and lemmas, and on POWLA (Chiarcos, 2012) for annotated corpora, while also implementing a minimalist ontology to express aspects that are relevant for lemmatizaion, such as the relation between corpus tokens and lemmas in the LiLa Lemma Bank.[6]

Although LiLa links 544 works covering a broad range of genres and periods of Latin literature, including those published in LASLA's extensive *Opera Latina* of classical texts (Fantoli et al., 2022; Fantoli et al., 2023), the overlap between the two collections is minimal. Only two poems among those represented in DICES, Virgil's *Aeneid* and Lucan's *Pharasalia* (Iurescia et al., 2023), are available in LiLa. The two poems amount to 442 speech passages and 154 character instances.

2.2. The limits of the DICES model

As the essays in Forstall and Verhelst (2025a) attest, DICES represents a huge leap forward in a research field that was often built on personal datasets recorded in manually annotated printed editions or spreadsheets. Nevertheless, while we address the question of integrating the resource in a wider context like the network of the Linguistic Linked Data Cloud, it is important to assess its limitations and the potential for further extensions of the model.

The first limitation is the narrow scope of the resource. DICES was deliberately tailored to work with one phenomenon (direct speech) in a specific genre (epic poetry). It remains to be tested whether the model is solid enough to be extended beyond its original use case.

Perhaps the most evident limit, however, is the somewhat shallow definition of the concepts and terms supporting the DB structure. The notion of "direct speech" itself is specified only loosely and in operational terms as "a sequence of contiguous lines in a given poem, [which] represents words spoken by one character to another" (Forstall et al., 2022, 974). While such a definition is (mostly) sufficient for a highly stylized genre like ancient epos, this may not be true for different kinds of works. Also, DICES's approach introduces some tension between the view of "speeches as portion of texts" and that of "speeches as narrative elements". What the nature of the relation between the two domains, the textual and the narratological, is is never discussed or clarified.

Problems of linguistic and literary theories also come to the forefront if any categorization of the speeches is attempted. The DB itself includes a table for tags, named "Speechtag" and visible in the bottom-left corner of Fig. 1, that lists labels such as "question", "taunt", "challenge" etc.; speech instances are also classified per type into "soliloquy", "monologue", "dialogue" and "general". This kind of categorization is clearly more controversial, and open to methodological debate. A controlled vocabulary of tags or types, although useful for retrieval, is not sufficient to give justice to such complex matters.

On account of those limits and the potential for expansion beyond the confines of epic poetry, we think that the right approach to foster interoperability between DICES and LiLa is to move in two directions. Firstly, to work on the conceptual model to describe direct and potentially other forms or represented speech; this model should provide a list of classes and properties to work with the data, as well as an adequate framework to express the concepts operative in a broad range of theoretical approaches. The second step is to create a textual resource which implements those concepts and is aligned with both DICES and LiLa, while remaining potentially open to include data from other genres of Latin literature.

2.3. Related works

The idea of creating a LOD textual resource with information on represented speech touches upon two different areas of research on the Semantic Web and the Digital Humanities. The first is the representation of textual annotation. DICES's operative definition of "direct speech" as a portion of text spanning across a poem's lines already implies the operations of segmenting a digital edition and predicating metadata about some of the sections. Research on Semantic Web models and language resources has witnessed spectacular advancements in the last years (Khan et al., 2022), but the degree of consolidation achieved across different resource types is far from uniform. While the modeling of lexical resources has converged around OntoLex-Lemon and its extensions, the sit-

[5] The DB thus records that Juno in e.g. Vergil's *Aeneid* and Hera in the Homeric poems are two separate instances, connected to the same character labeled Hera: https://db.dices.mta.ca/app/character/6743/.

[6] http://lila-erc.eu/ontologies/lila/.

uation is considerably less mature with respect to textual resources and linguistic annotation. Cimiano et al. (2020) discuss three solution that can be used to represent corpora (and corpus annotation) as Linked Data: the aforementioned POWLA, the NLP Interchange Format (NIF, Hellmann et al., 2013) and Web Annotation (Sanderson et al., 2017). While these models vary in their expressivity for linguistic concepts, they all support the most important use cases of linguistic annotation, including text segmentation at different levels of granularity, from long spans to single tokens.

Another area that is relevant in this context is that of narratology. A series of initiatives have attempted to create formal ontologies of story elements, like characters, plot events and narrative sequences, some working on specific domains and from a singular theoretical perspective, some broader in scope and in support for different theories.[7] Most recently, Pianzola et al. (2025) created GOLEM with the ambition not only of providing a general coverage for concepts used across the domain, but also of supporting statements about provenance and and alignment with other foundational and high-level ontologies used in the DH, like DOLCE (Borgo et al., 2022) and the CIDOC-CRM (Doerr, 2003). The GOLEM ontology is structured in 6 modules dedicated to characters, social relationships, events, settings, narrative (i.e. narrative material or *Erzählstoffe*), and inference (i.e. documentation of interpretation and provenance).[8] These ontologies (including GOLEM) typically aim to account for the structure of narrative works, but do not consider the anchoring of the narratological elements to specific portions of the texts. In other words, they do not support annotating the texts with the narratological concepts.

GOLEM provides an effective framework to model character instances, persisting characters (called "Character-Stoff" in GOLEM)[9] and character traits. It would be possible to rely on its definitions to cast speeches as narrative units and events, which in GOLEM are aligned respectively to DOLCE's perdurants and descriptions.[10] There are however some limits in this course of action. For our domain of represented speech in literary texts, GOLEM is both over- and under-specified. As said, the module aims to cover all major concepts in narratology and express all kinds of narrative sequences and plot elements, whereas our interest is focused on a specific type of communicative acts. At the same time, GOLEM's use of the concept of roles from DOLCE seems to be oriented towards general macro-roles played by actants in a story (similar to those defined by Propp, 1968); it seems less useful to convey a more granular classification of micro-roles played by participants in communication that are repeatedly exchanged as one speaker becomes the addressee in the space of one conversational turn.

The DOLCE+DnS expansion (Gangemi and Mika, 2003), which is also reused by GOLEM, introduces the concepts of descriptions and situations and exemplifies them with a theory of communication based on Jackobson's model. DOLCE+DnS (Descriptions and Situations) is an extension of the foundational ontology DOLCE aimed at modeling contextualization and intentional structures. It introduces *descriptions*, as reified theoretical constructs, and *situations*, as structured configurations of entities that satisfy such constructs. In particular, a description is defined as 'an entity that partly represents a (possibly formalized) theory T (or one of its elements) that can be "conceived" by an agent: either human, collective, social, or artificial' (Gangemi and Mika, 2003, 694). Descriptions contain functional components such as roles, parameters, and courses which classify entities within a context. Gangemi and Mika (2003) exemplify the architecture by discussing a model of communication theory: drawing on Jakobson's schema, the six elements of communication (addresser, addressee, message, code, context, and channel) are represented as functional roles within a description, while individual communicative acts are situations satisfying that description.

DOLCE+DnS offers several advantages for the conceptualization of represented speech in literary texts. Firstly, the ontology is clearly suited to represent a theory of communication with its roles. Secondly, since DnS is integrated into the broader DOLCE framework, communicative situations can be seamlessly connected to other ontological categories, such as narrative events modeled as perdurants or dialogic sequences conceptualized as courses. Finally, the explicit separation between the structuring conceptual schema (Description) and the concrete configuration it organizes (Situation) enables the coexistence of multiple interpretative frameworks.

3. Towards an ontology for representing speech in literature

This section discusses the design principles that were followed to create a basic ontological skeleton to represent DICES's speeches in LaReS. As stated, the aim is both to improve the theoretical

[7] See the review of 10 projects in Pianzola et al. (2025, 3-8), with the useful synoptic table at p. 4.

[8] https://ontology.golemlab.eu/.

[9] https://w3id.org/golem/ontology#G0_Character-Stoff.

[10] See https://w3id.org/golem/ontology#G5_Narrative_Event, and https://w3id.org/golem/ontology#G9_Narrative_Unit.

solidity of the dataset, and to support future extensions to other phenomena (like the indirect or free-indirect speech), multiple theoretical perspective, genres and even different literary traditions, outside the domain of Classics.

In view of these goals, we start by defining our domain broadly, so as to encompass the general phenomenon of represented speech. For present purposes, we understand the concept as covering all cases in which a text represents, quotes or re-enacts an enunciation that is attributed to a speaker/encoder and is recognizable, within the general framing context where it is embedded, as an autonomous enunciative unit.[11]

Our proposal for LaReS is based on the separation between the two analytical levels that are operatively conflated in DICES: (i) the textual level, corresponding to the precise segments where the quoted speech is found, or where the linguistic clues signaling a represented speech are identified by the critics; (ii) the narratological level where: (ii.a) the communicative dynamics evoked by the text, and (ii.b) the events taking place in the narrative words (e.g. two characters having a conversation) are reconstructed. This distinction constitutes the core architectural principle of the model and underlies all subsequent decisions.

In LaReS we adopt the CIDOC Conceptual Reference Model (CIDOC CRM, Doerr, 2003),[12] a higher-level ontology for cultural heritage that is widely used in DH and was recently adopted in LiLa for documenting historical aspects of Latin linguistics (Pellegrini et al., 2025), to provide a general conceptualization of these two levels. We align the textual units to `crm:E33_Linguistic_Object`, whereas the speeches as units in a narrative perspective are primarily seen as propositional content and conceptualized as `crm:E89_Propositional_Object`. The connection between the two levels is established through `crm:P67_referes_to`, which links the textual passage to the narratological entity.

[11] The term "reported speech" is often used to describe situations in which an enunciation act E_1, attributed to a speaker L, is embedded within a framing enunciation E, produced either by L or by another author/encoder (Mortara Garavelli, 1985, 21). Although this notion is more narrowly circumscribed, the present work adopts the broader expression "represented speech" in order to remain neutral with respect to specific structural configurations. This choice anticipates possible extensions to genres such as drama, where the notion of a general framing enunciation, represented by the work itself, is more controversial and difficult to grasp, yet the notion of autonomous represented enunciation acts by the characters remains operative.

[12] The latest stable version available at the moment of the CIDOC-CRM is 7.3.1: `https://cidoc-crm.org/Version/version-7.1.3`.

Further specifications beyond this general modeling is left to the two modules. The following sessions discuss some further requirements and solutions adopted for each of them.

3.1. Representing textual units

This component of the ontology is responsible to handle the task of isolating and identifying the textual portions that provide evidence for the analysis. DICES relied on a model where the quoted speech attributed to a character was identified with start and ending line. This model must be made more robust to support the reference system adopted for works where line is not a valid textual unit, and to be capable of being more granular. Already in DICES, in fact, line is not always a perfect unit for segmentation. The first speech in the *Aeneid*, for instance, is set to start at line 1.37, but the first words of that verse are a tagging phrase by the narrator: *haec secum:...*, "this [Juno] said to herself." The granularity level of (spans of) tokens is also needed to comment upon linguistic clues of represented speech (such as the use of deixis, or verbal tenses) that are localized in single words.

At the ontology level, we decided to avoid committing to any specific theory of linguistic annotation: different implementations of tokenization or text segmentation can be delegated to the single projects. The represented-speech passage is modeled as an instance of `crm:E33_Linguistic_Object`, capturing the identity of the passage as a symbolic and intentional linguistic entity, independently of any particular material support or edition. The linguistic object proper is distinguished from its annotatable textual representation. To capture this, we introduce the custom class `AnnotatableTextRepresentation`. The class denotes textual objects in which linguistic content becomes addressable. Rather than identifying this interface between linguistic content and annotation with specific frameworks, we treat it as a notion that can be implemented in different ways, such as instances of `powla:Document` or `nif:Context`.

For our specific use case, LaReS relies on POWLA, the ontology for textual annotation adopted in LiLa. We thus model the `AnnotatableTextRepresentation` of each speech instance as a document, which is defined in POWLA as "a(n annotated piece of) primary data". Within the annotable text, non-terminal nodes are created to isolate the phrases that we want to comment.

3.2. Representing the communicative situation

Represented speeches evoke a situation where communication involving at least one actor takes

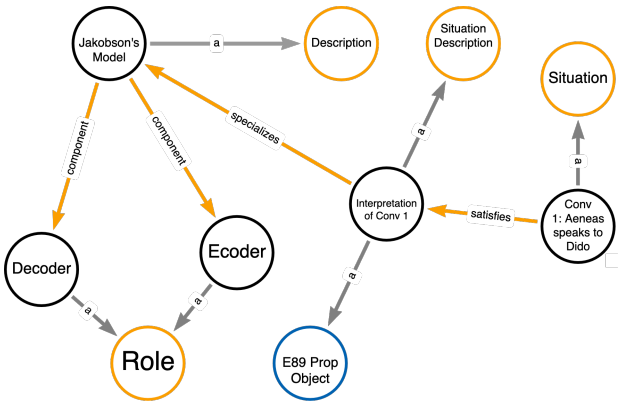

Figure 2: Model of a communication theory (simplified Jakobson model) as a Description and an application to a fictitious scene where Aeneas speaks to Dido (Conv 1), modeled as a Situation-Description. Yellow is used for classes and properties from DOLCE+DnS, blue for CIDOC CRM.

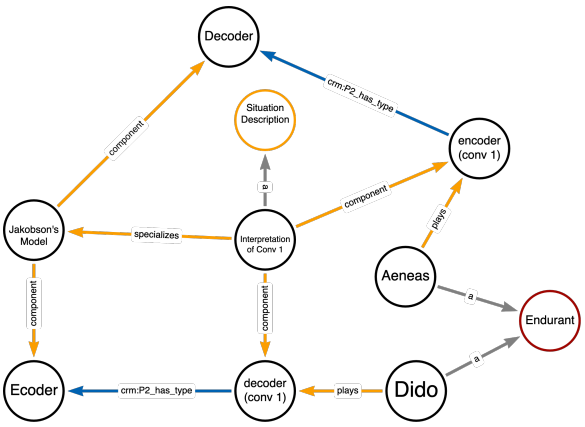

Figure 3: Character instances (endurant), local and general roles in the interpretative frame of the (fictitious) speech scene between Aeneas and Dido (same as in Fig. 2). Red is used for the core DOLCE classes.

place. This situation can be explained with the help of a theory of communication. Adopting a model from cognitive literay studies (e.g. Stockwell and Mahlberg, 2015), we can also see the text as evoking an event that is parsed by the readers with the help of the same mental model or frame that is used for real-word communicative acts.

This framework is ideally suited to be captured by the classes and properties of DOLCE+DnS. In DOLCE+DnS, we can explicitly model the theory of communication that we rely on for our interpretation. For the DICES data, a simplified version of the Jakobsonian model discussed by Borgo et al. (2022) with just two roles (that we label "encoder" and "decoder") is sufficient to be the general Description underlying all analyses. If other datasets require more complex theoretical frameworks, all that is needed is to model the framework as a Description and to make the adopted model available as part of the dataset.

The simplified encoder/decoder description represents the conceptual model adopted in LaReS for the DICES data, but a specific Situation-Description (itself a subclass of Description) is needed as a contextualized or applied version of the theory that mediates between the abstract schema and the concrete configuration of entities in a situation. We propose to conceptualize our narratological interpretation of reported speech as Situation-Descriptions, i.e. as the analytical unit (produced as the issue of general model of communication) that generates a structured representation of a situation. This conceptualization is compatible with the previous definition of the speech unit as a `crm:E89_Propositional_Object` and is similar to the propositional interpretation of narrative units given by GOLEM.[13] However, whereas GOLEM's narrative units are defined generically as descriptions, our proposal is to conceptualize the speech sections as anchored to a situation (thus, Situation-Descriptions); in our opinion, this choice is more appropriate to the status of the analytical unit, which mediates between a general theory and a particular situation. Note that, while a Situation-Description conceptually requires the existence of a situation that satisfies it, it may not be relevant to actually instantiate it in a dataset, and this hasn't been done in LaReS so far. Figure 2 illustrates the relation between the general theory (Description) and its application to a (fictitious) example where Aeneas speaks to Dido (Situation-Description).

Communicative roles are required both by the general theory and by the Situation-Description applied to the singular case. In our dataset a character may switch multiple times between the role of speaker/encoder and that of addressee/decoder; in the *Aeneid*, for instance, the main character Aeneas is the speaker of 70 speeches, and the addressee for 80. In our model, it is important to keep a precise inventory of the role played by Aeneas in each of these. Therefore, the Situation-Description defines local roles for each communicative scene. Those roles are then mapped to the abstract role required by the general theory. Characters, once again defined locally for each work like the "character instances" in DICES, can be linked to the local role via the DOLCE+DnS property `dns:play`, which connects a DOLCE's endurant with a role. Character instances are thus defined as endurants. For them, a mapping to the class of `golem:G1_Character` would certainly

[13] See `https://w3id.org/golem/ontology#G9_Narrative_Unit`.

be possible and is under consideration. This option would open the possibility to use the "character-Stoff" class (golem:G0_Character-Stoff) to express the continuity between figures across cultures, works and media (DICES's relation between character instances and characters); it would also allow us to leverage GOLEM's class of character features (golem:G17_Character_Feature) to model relevant traits of the fictional persona (such as gender, age or geographical origin). At the moment, a character (instance) is simply defined as an endurant and a propositional object (crm:E89_Propositional_Object).[14]

Figure 3 illustrates the relation between the character instances Aeneas and Dido in the *Aeneid*, their role in the interpretative frame of a fictitious scene where Aeneas is speaker, and the relation between those local roles and the general ones defined in the theory (the simplified Jakobsonian model). Note that the general theoretical framework (the simplified Jakobson's model) ant the general roles of Encoder and Decoder are the same as in Figure 2 and are repeated to show how the two diagrams are interconnected.

4. LaReS – Latin Represented Speech

The model discussed in Sec. 3 was applied to the 341 speeches in Virgil's *Aeneid*. For the *Pharsalia*, the version linked to LiLa does not list line numbers or otherwise connect the tokens to canonical citation units: automatic alignment with DICES is therefore impossible at the moment.

Figure 4 illustrates the two views on the represented speech, as a textual object and as a narrative unit in LaReS, using the first speech found in the poem (*Aeneid* 1.37-49) as an example.[15] The two representations are embodied by the main nodes in the middle of the figure, on the left and center, colored in blue (the textual element, labelled "Speech text (Aen. 1,37-49)"), and purple (the speech as narrative unit, labeled "Juno'speech (Aen. 1,37-49)").

The left-side of the figure shows how the annotation is treated. The speech (an instance of crm:E33_Linguistic_Object, as visible from the dark-green node connected to the speech element) is linked to an "Annotable text representation" (lighter-green node at the left). This, in turn, is typed as a powla:Document and linked to a document layer and, through it, to a non-terminal node (labeled "Speech annotation unit" and displayed as a yellow-green node). The non-terminal connects all the tokens that are assigned to the direct speech, thus allowing for an effective retrieval. In Fig. 4 (dark green nodes at the bottom of the image) only the first and last LiLa tokens are displayed for brevity; those tokens are *me*,[16] and *honorem*.[17] The reuse of LiLa's tokens ensure the connection between LaReS and the LiLa KB.

The right section of the image, around the purple node labeled "Juno'speech (Aen. 1,37-49)", systematizes the interpretation of the speech as an element of the narrative, with the help of the selected theory (which is represented here in the red node at the center-right side of the figure). The role played by the character instance Juno defined locally for this speech (the blue node labeled "Encoder role (SU00000001)") is explicitly connected to the general role 'Encoder' required by the theory. Note that, although both roles envisaged in the simplified version of Jakobson's model adopted here are reported in the image (orange nodes in the right side), only one, the Encoder, is actually realized as local role, since this particular speech is in fact a soliloquy.

At present, the alignment with DICES is ensured by signaling that the DICES dataset is used as source, via the prov:wasDerivedFrom; this property links the speech unit to the whole DB and to a specific record (the latter is not shown in the figure).[18] At the same time, the human-readable web paged of both the speech and the character instance are linked (via rdfs:seeAlso) to the speech unit and the endurant Juno respectively. Another important source of alignment is the use of a CTS URN to identify the speech textual representation, which is not visible in the image. The process of creating CTS URNs for all LiLa's textual elements is ongoing and should be an important upgrade for the whole collection of textual resources.

Currently, 36,782 tokens from the *Aeneid* in LiLa

[14] Note that the CIDOC's FRBRoo extension included a class F38 Character and a property P57 is based on. The former was defined as a subclass crm:E89_Propositional_Object, but did not differentiate between an instance in a given work and the general figure: "Harry Potter", for instance, was listed in the documentation as covering both the main character of the book series and the films. The property was described as a shortcut for the path from a Conceptual Object (E28) through a Creation process (E65) motivated (P17) by a CRM Entity (E1) restricted to E39 Actor. Class and properties were however deprecated in version 0.6 of FRBRoo's successor, the Library Reference Model (LRMoo).

[15] See https://db.dices.mta.ca/app/speech/A201/.

[16] http://lila-erc.eu/data/corpora/Lasla/id/corpus/VergiliusAeneis/Vergilius_Aeneis_VerAen01.BPN_t_0000279.

[17] http://lila-erc.eu/data/corpora/Lasla/id/corpus/VergiliusAeneis/Vergilius_Aeneis_VerAen01.BPN_t_0000364.

[18] The record is retrievable via DICES's API: https://db.dices.mta.ca/api/speeches/1528/. This API address is used for the RDF triple.

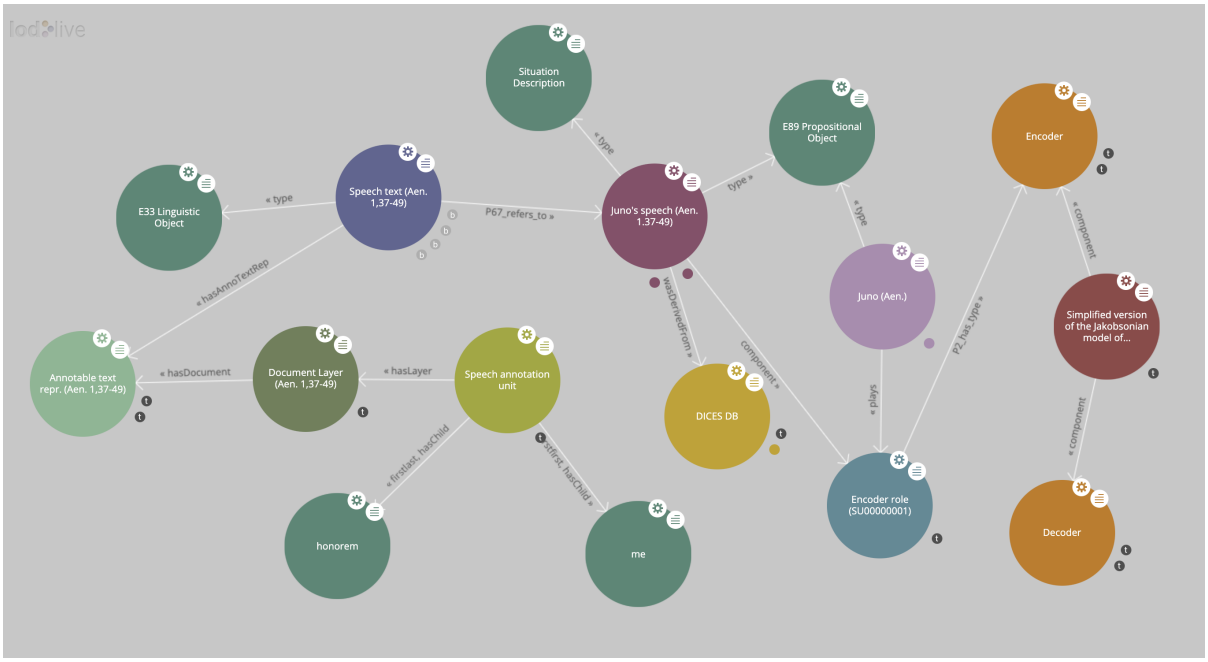

Figure 4: Overview of a speech (Juno's soliloquy in *Aeneid* 1.37-49).

have been linked to the 341 speech-text nodes created. The number is increased by the fact that represented speeches can often be embedded, resulting in tokens that are linked to multiple instances of textual sections, as in this case the same token is linked to both the framing and the quoting speech passage. In fact, the whole section from 2.3 to 3.715 is one long direct speech by Aeneas, who recounts his flight from Troy to Dido, often using embedded direct speech to report dialogues within his tale.

5. Future works and open problems

The idea of aligning DICES with LiLa has raised momentous problems that discouraged us from adopting a simple linking solution. On the contrary, in order to ensure both extensibility and support of multiple theoretical approaches at various levels of granularity, important modeling decisions had to be taken.

The next goal is to test the modeling structure adopted here and, eventually, produce a stable ontology for represented speech in literary texts. Ideally, this ontology should be applicable also beyond the scope of Latin literature, and should be expressive enough to account for the phenomena in multiple traditions. Extensive testing of this sort is ongoing with the help of specialists of literary heritage from different languages and cultures within the "Dipartimento of Scienze Linguistiche and Letterature Straniere" at the Università Cattolica del Sacro Cuore.

Several details of speech representation must be improved or defined better. Currently, speech embedding (of the sort exemplified above by *Aeneid* 2.3-3.715) is represented only in terms of structural relations between textual sections (via `crm:P106_is_composed_of` over the E33 instances). The model should, however, be made more robust, so as to allow to express the relation between a framing and an embedded speech at the speech-unit level as well.

More data should be made available in LaReS, both by making sure that more texts from DICES are present in LiLa, and especially by making the tokens from *Pharsalia* to become discoverable via canonical identifiers like line numbers.

In the scope of the LiLa project, the next stage will be to investigate how the model is portable to Seneca's tragedies. Annotated versions of the *Hercules Furens*, *Agamemnon*, and *Oedipus*, based on the LASLA texts linked to LiLa, are distributed as part of the CIRCSE UD Latin treebank. As reported by the documentation, they include the speaker metadata.[19]

Acknowledgments

This work is part of the project "Per un'ontologia dei discorsi riportati nelle opere letterarie", supported by the "Dipartimento di Studi Linguistici e Letterature Stranierie" at Università Cattolica del Sacro Cuore.

[19] See https://github.com/UniversalDependencies/UD_Latin-CIRCSE.

6. Bibliographical References

Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M. Sanfilippo, and Laure Vieu. 2022. Dolce: A descriptive ontology for linguistic and cognitive engineering. *Applied Ontology*, 17(1):45–69.

John Frederick Burrows. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Clarendon Press, Oxford.

Christian Chiarcos. 2012. POWLA: Modeling Linguistic Corpora in OWL/DL. In *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 225–239, Berlin, Heidelberg. Springer.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

Martin Doerr. 2003. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3):75–92. Number: 3.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France. European Language Resources Association.

Christopher W Forstall, Simone Finkmann, and Berenice Verhelst. 2022. Towards a linked open data resource for direct speech acts in Greek and Latin epic. *Digital Scholarship in the Humanities*, 37(4):972–981.

Christopher W Forstall and Berenice Verhelst, editors. 2025a. *Direct Speech in Greek and Latin Epic: Expanding the Methods and Canon*. Brill, Leiden.

Christopher W. Forstall and Berenice Verhelst. 2025b. Introduction. In Christopher W Forstall and Berenice Verhelst, editors, *Direct Speech in Greek and Latin Epic*, pages 1–31. Brill.

Aldo Gangemi and Peter Mika. 2003. Understanding the Semantic Web through Descriptions and Situations. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Robert Meersman, Zahir Tari, and Douglas C. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, volume 2888, pages 689–706. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Greta Hawes and Scott Smith. 2021. A dataset of mythical people with stable URIs. MANTO Blog.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*.

Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros Muñoz, and Ciprian-Octavian Truică. 2022. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web*, 13(6):987–1050.

Francesco Mambrini and Marco Carlo Passarotti. 2023. The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms. *Journal of Open Humanities Data*, 9(1).

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: development and applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 587–597, Brno. Lexical Computing.

Brian McHale. 2011. Speech Representation. In Peter Hühn, John Pier, Wolf Schmid, and Jörg Schöner, editors, *The Living Handbook of Narratology*. University of Hamburg.

Bice Mortara Garavelli. 1985. *La parola d'altri. Prospettive di analisi del discorso*. Sellerio, Palermo.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212.

Matteo Pellegrini, Valeria Irene Boano, Francesco Gardani, Francesco Mambrini, Giovanni Moretti, and Marco Carlo Passarotti. 2025. DynaMorph-Pro: A new diachronic and multilingual lexical resource in the LLOD ecosystem. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 208–220, Naples, Italy. Unior Press.

Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, and Martina Verdelli. 2022. Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension. *Prague Bulletin of Mathematical Linguistics*, 119(1):67–92.

Federico Pianzola, Luotong Cheng, Franziska Pannach, Xiaoyan Yang, and Luca Scotti. 2025. The GOLEM Ontology for Narrative and Fiction. *Humanities*, 14(10):193.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press, Austin. [Original edition: Leningrad. 1928].

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. Web Annotation Data Model. W3C Recommendation.

Neel Smith. 2009. Citation in Classical Studies. *Digital Humanities Quarterly*, 3(1).

Rachele Sprugnoli, Francesco Mambrini, Marco Carlo Passarotti, and Giovanni Moretti. 2023. The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace. *IJCOL - Italian Journal of Computational Linguistics*, 9(1):53–71.

Peter Stockwell and Michaela Mahlberg. 2015. Mind-modelling with corpus stylistics in David Copperfield. *Language and Literature*, 24(2):129–147.

Jochen Tiepmar and Gerhard Heyer. 2019. The Canonical Text Services in Classics and Beyond. In Monica Berti, editor, *Ancient Greek and Latin in the Digital Revolution*, pages 95–114. De Gruyter, Berlin, Boston.

7. Language Resource References

Austen, Jane, 1775-1817. 1988. *Pride and prejudice : (tagged version) / compiled by J.F. Burrows*. Oxford Text Archive. PID http://hdl.handle.net/20.500.14106/1229.

Fantoli, Margherita and Passarotti, Marco Carlo and Litta Modignani Picozzi, Eleonora and Ruffolo, Paolo and Moretti, Giovanni. 2023. *LASLA Corpus (RDF version)*. CIRCSE. Zenodo, v1.0.1. PID doi:10.5281/zenodo.8370759.

Forstall, Christopher and Verhelst, Berenice and Finkmann, Simone. 2025. *Raw Data for DICES Epic Speeches Database*. Mount Allison University Dataverse / Borealis. PID doi:10.5683/SP3/N8LS2Y.

Federica Iurescia and Giovanni Moretti and Marinella Testori and Marco Passarotti and Martina Verdelli and Flavio Massimiliano Cecchini. 2023. *Lucani Pharsalia*. CIRCSE. Zenodo, 1.0.0. PID doi:10.5281/zenodo.8027881.

Giovanni Moretti and Marco Passarotti and Rachele Sprugnoli and Paolo Ruffolo and Francesco Mambrini. 2023. *LiLa Lemma Bank*. CIRCSE. Zenodo, V1.2. PID doi:10.5281/zenodo.8300851.