LREC 2026

10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026

Workshop Proceedings

Editors
John P. McCrae, Katerina Gkirtzou, Fahad Khan, Patricia Martín Chozas, Sara Carvalho, and Erin Canning

12 May 2026

ISBN 978-2-493814-72-2

Preface

The Linked Data in Linguistics (LDL) workshop series has, over the years, established itself as an essential forum for the discussion of the latest trends in the application of Linked Data (LD) and Semantic Web technologies to linguistics and related fields such as digital lexicography, and the digital humanities. This year's edition of the workshop has a strong focus on the generation of LD datasets as well as the continued exploration of LD as a means of integrating heterogeneous data, both linguistic data as well as data deriving from related fields. In particular, LDL-2026 showcases a number of new projects and initiatives which aim to introduce new kinds of data and/or new kinds of resources into the linguistic linked data ecosystem, and which we therefore feel will be of great interest to the linguistic linked data community. For instance, Pellegrini et al's "Linked Open Data for West Nilotic Languages: The NILOMORPH project" presents a new project addressing the integration of complex, multilingual, and multimodal data on West Nilotic morphology via an LD approach. In "A Linguistic Ontology for Constructicography: the Research Constructicon and its Ontology Modules", instead, Winckel et al introduce the Research Constructicon (RCxn), a community-driven, ontology-based resource for representing constructions and their associated research processes. Meanwhile, McCrae, in "Open English NameNet: Extending English Wordnet with Names", presents a large-scale extension of WordNet with named entities from Wikidata, bridging lexical and encyclopedic knowledge. Billero's "Towards a Linguistic Linked Open Data Resource for Italian Cultural Heritage: The Lessico dei Beni Culturali Corpus" outlines efforts to represent the Lessico dei Beni Culturali as Linked Data, focusing on challenges in modeling historical and culturally-bound terminology. And in "Latin Represented Speech (LaReS): Linking LiLa and the DICES Database", Mambrini introduces a Linked Open Data (LOD) resource modeling represented speech in Latin literature through integration with the Linking Latin (LiLa) knowledge graph and established ontologies.

The workshop also showcases innovative uses of LD standards and technology as applied to linguistically relevant datasets. In "Towards the LinkEn Knowledge Base. A Neuro-Symbolic Approach to Build a Linked Data Hub for English Lemmas with Large Language Models", Augello and Passarotti introduce a lemma-centered knowledge base for English that combines OntoLex-Lemon modeling with a hybrid neuro-symbolic pipeline to support semi-automatic LD creation; Fiumanò et al's "Victim or Assailant? Exploring Narratives Through Knowledge Graph Queries" introduces DORIS, an ontology enabling cross-document framing analysis through knowledge graphs grounded in Frame Semantics; Chiarcos and Siewert's "Consolidating Syntactically Annotated Corpora with LLOD Technology. An Experiment in the Old Saxon Heliand" demonstrates how LLOD technologies can consolidate heterogeneous syntactic annotations of the Old Saxon Heliand using graph-based transformations; and Albertelli's "Modeling Topics as Linguistic Linked Open Data: a First Attempt Using BERTopic, Ontolex-Lemon and FrAC" explores the formalization of dynamically extracted topics from parliamentary discourse as Linked Data entities linked to linguistic and political metadata. Finally, McCrae et al's "Bridging the Gap Between Ontologies and Dictionaries: Requirements and Implementation of a New Core for OntoLex-Lemon" reports on a new core module for OntoLex-Lemon, addressing gaps in lexicographic modeling while ensuring compatibility with existing standards.

Taken together, the contributions to this year's workshop serve to highlight ongoing efforts to build interoperable language resources, to integrate heterogeneous linguistically relevant data, and to extend LD approaches to a greater variety of linguistic and cultural phenomena.

The Organizing Committee for LDL-2026

Organizing Committee

John P. McCrae, (University of Galway, Ireland)
Katerina Gkirtzou (Athena Research Center, Greece)
Fahad Khan (Consiglio Nazionale delle Ricerche and CLARIN-IT, Italy)
Patricia Martín Chozas (Universidad Politecnica de Madrid, Spain)
Sara Carvalho (University of Aveiro, Portugal)
Erin Canning (University of Oxford and Victoria and Albert Museum, United Kingdom)

Program Committee

Sina Ahmadi (University of Zurich, Austria)
Verginica Barbu Mititelu (Research Institute for Artificial Intelligence of the Romanian Academy, Romania)
Paul Buitelaar (Insight, Ireland)
Rute Costa (NOVA FCSH/NOVA CLUNL, Portugal)
Milan Dojchinovski (Czech Technical University, Czech Republic)
Francesca Frontini (CNR-ILC, Italy)
Frances Gillis Webber (University of Cape Town, South Africa)
Yoshihiko Hayashi (Waseda University, Japan)
Max Ionov (University of Zaragoza, Spain)
Alik Kirillovich (ex. Higher School of Economics, Russia)
Penny Labropoulou (Athena Research Center, Greece)
Chaya Liebeskind (Jerusalem College of Technology, Israel)
David Lindemann (University of the Basque Country, Spain)
Elena Montiel Ponsoda (Universidad Politécnica de Madrid, Spain)
Steven Moran (University of Neuchâtel, Switzerland)
Petya Osenova (IICT-BAS, Bulgaria)
Ana Ostroški Anić (Institute of Croatian Language and Linguistics, Croatia)
Antonio Pareja Lora (Universidad de Alcalá, Spain)
Giulia Pedonese (CNR-ILC, Italy)
Sigita Rackevičienė (Mykolas Romeris University, Lithuania)
Felix Sasaki (SAP, Germany)
Andrea Schalley (Karlstad University, Sweden)
Gilles Sérasset (University Grenoble Alpes, France)
Milena Slavcheva (IICT-BAS, Bulgaria)
Blerina Spahiu (Bicocca University, Italy)
Ranka Stanković (University of Belgrade, Serbia)
Armando Stellato (University of Rome, Italy)
Marieke van Erp (KNAW Humanities Cluster, The Netherlands)
Federica Vezzani (University of Padua, Italy)

Table of Contents

Workshop Program

09:00–09:15 ***Registration and Welcome***

09:15–10:00 *Invited Talk: "Making Knowledge Visible Again: Linguistic Linked Data in the Age of Large Language Models."*
Johanna Monti

10:00–10:30 ***Minute Madness and Poster Session***

10:30–11:00 ***Coffee Break and Poster Session***

11:00–11:20 *Towards the LinkEn Knowledge Base. A Neuro-Symbolic Approach to Build a Linked Data Hub for English Lemmas with Large Language Models*
Lorenzo Augello and Marco Passarotti

11:20–11:40 *Open English NameNet: Extending English Wordnet with Names*
John P. McCrae

11:40–12:00 *Victim or Assailant? Exploring Narratives through Knowledge Graph Queries*
Beatrice Fiumanò, Nicolas Lazzari, Simone Paolo Ponzetto and Valentina Presutti

12:00–12:20 *Modeling Topics as Linguistic Linked Open Data: A First Attempt Using BERTopic, Ontolex-Lemon and FrAC*
Lisa Sophie Albertelli

12:20–12:40 *Consolidating Syntactically Annotated Corpora with LLOD Technology. An Experiment in the Old Saxon Heliand*
Christian Chiarcos and Janine Siewert

12:40–13:00 *Bridging the Gap between Ontologies and Dictionaries: Requirements and Implementation of a New Core for OntoLex-Lemon*
John P. McCrae, Jorge Gracia, Fahad Khan and Philipp Cimiano

13:00–13:05 ***Closing remarks***

List of Posters

Linked Open Data for West Nilotic Languages: The NILOMORPH Project
Matteo Pellegrini, Matthew Baerman and Oliver Bond

A Linguistic Ontology for Constructicography: The Research Constructicon and its Ontology Modules
Elodie Winckel, Peter Uhrig and Stephanie Evert

Towards a Linguistic Linked Open Data Resource for Italian Cultural Heritage: The Lessico Dei Beni Culturali Corpus
Riccardo Billero

Latin Represented Speech (LaReS): Linking LiLa and the DICES Database
Francesco Mambrini

Modeling Topics As Linguistic Linked Open Data: a First Attempt Using BERTopic, OntoLex-Lemon and FrAC

Lisa Sophie Albertelli
Università Cattolica del Sacro Cuore
Largo Gemelli, 1, 20123 Milan, Italy
lisasophie.albertelli01@icatt.it

Abstract

Parliamentary discourse constitutes a key domain in which political actors publicly articulate policy positions and priorities through language. This study investigates debates from the Italian Chamber of Deputies (1948–2006) to identify and analyse latent semantic themes and their evolution using BERTopic-based dynamic topic modeling. The analysis relies on a subset of the ItaParlCorpus (Cova, 2025b), a large-scale, machine-readable corpus enriched with temporal, institutional, and political metadata. Beyond topic extraction, this work addresses a largely unexplored challenge: the formalization of topics derived from unsupervised, embedding-based topic modeling as Linked Data entities, adopting a linguistic perspective. Extracted topics are formalized as semantic entities reusing the OntoLex–Lemon model, its FrAC extension and declaring a dedicated ontology to link topics to speeches, speakers, political parties, and temporal information reusing standardized vocabularies and persistent URIs. This integration enables semantic querying through SPARQL, supporting analyses of topic distributions across political actors, parties and illustrating the analytical potential of the proposed approach. Moreover, the study highlights limitations in the formalization of topic modeling outputs, particularly regarding the representation of ambiguous word forms and their alignment with lexical concepts in OntoLex–Lemon.

Keywords: Linked Open Data, BERTopic, Dynamic Topic Modeling, Parliamentary Discourse, OntoLex-Lemon, FrAC, Semantic Web

1. Introduction

Parliamentary discourse represents a significant domain in which political actors publicly articulate policy positions, priorities, and social issues through language. This study presents a first attempt to formalize topics extracted from Italian parliamentary debates via unsupervised neural topic modeling approach, as interoperable semantic entities within a Linked Open Data (LOD) framework. By employing BERTopic (Grootendorst, 2022) to uncover latent thematic structures in Italian parliamentary debates, this study models the extracted topics as interoperable entities, adopting a linguistic perspective and leveraging established ontologies. In particular, the OntoLex–Lemon model (McCrae et al., 2017) and its FrAC extension (Chiarcos et al., 2022) are employed to formally represent topics internal lexical structure. The adopted ItaParlCorpus, a large-scale, machine-readable and richly annotated collection of speeches spanning 1948–2022, enables the integration of computationally derived topics with detailed socio-political metadata, including speakers, political parties and ideological families. Thus, topics move beyond being mere outputs of an unsupervised machine learning algorithm and become identifiable, interpretable, and reusable semantic entities. Through their formal representation, they are linked to external knowledge bases such as Wikidata (Vrandečić and Krötzsch, 2014), while their linguistic structure is explicitly modeled using LiITA (Litta et al., 2025) and BabelNet (Navigli and Ponzetto, 2010). SPARQL queries over the resulting Linked Data enable the exploration of both the linguistic structure of the modeled topics and their associated socio-political context. However, as only a limited subset of topics and documents has been formalized, the results of these queries cannot be considered generalizable and should instead be interpreted as exploratory findings. The current dataset is in fact intended to demonstrate the analytical potential of the framework rather than to support generalizable political conclusions. Beyond demonstrating the feasibility of integrating neural topic modeling outputs into the Semantic Web, this approach also highlights the methodological challenges involved in representing abstract, computationally derived topics within formal ontologies, particularly within the OntoLex–Lemon framework. Furthermore, it establishes a framework for future extensions, including multilingual, dynamic and cross-corpus applications. The paper is structured as follows: Section 2 reviews the state of the art; Section 3 presents the data used in the study; Section 4 describes the methodology adopted for extracting topics using BERTopic and for modeling them as Linked Open Data entities. Section 5 discusses the challenges that arose during topics formalization and outlines possible future work, while Section 6 presents the conclusions.

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 1–12
12 May 2026.

2. State of the Art

The computational analysis of parliamentary debates has increasingly relied on large-scale, machine-readable corpora enriched with structured metadata, supported by initiatives including ParlSpeech (Rauh et al., 2017), ParlSpeech V2 (Rauh, 2020), ParlaMint (Erjavec et al., 2023) and ParlaMint II (Erjavec et al., 2025), which provide structured access to parliamentary speeches enriched with metadata. In parallel, several projects have adopted LOD principles to represent parliamentary actors, speeches, and institutional information as interconnected knowledge graphs, including LinkedEP (van Aggelen et al., 2017), Linked-Saeima (Bojārs et al., 2019), PARLIAMENTSAMPO (Hyvönen et al., 2025), and national open data portals. These efforts demonstrate the value of modeling parliamentary data as interoperable semantic resources, enabling advanced querying and cross-dataset integration. At the same time, topic modeling techniques, ranging from probabilistic approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to more recent embedding-based methods such as BERTopic, have been widely applied to parliamentary debates to uncover latent themes. Within this line of research, topics are typically used as exploratory or analytical constructs to support further quantitative analyses of political discourse. However, their role has remained largely methodological and no attention has yet been given to their formalization as explicit semantic entities within a Linked Data framework. The present study explores this direction. Rather than focusing on the formalization of parliamentary data as Linked Open Data per se, it focuses on the semantic modeling of topics extracted through neural based topic modeling. Specifically, computationally derived topics are treated as Linguistic Linked Open Data entities, whose lexical composition, provenance and contextual associations are explicitly represented. This modeling strategy enables topics to function as structured and reusable semantic objects, creating a bridge between neural topic modeling and Semantic Web technologies.

3. Data

The employed dataset is the ItaParlCorpus (Cova, 2025b) , a large, machine-readable collection of speeches from the Italian Chamber of Deputies, obtained from the Harvard Dataverse repository [1]. For the period 1948–2006, three datasets were used: `camera_1948-1972.csv`, `camera_1972-1992.csv`, and `camera_1992-2006.csv`. Each row in these files corresponds to a single parliamentary intervention and includes rich metadata alongside the speech text. Metadata cover temporal information (`date`, `year`, `legislature`, `doc_id`, `row_id`), speaker details (name, unique identifier aligned with the Comparative Legislators Database), party affiliation (`party_name`, `party_family`, `party identifiers from ParlGov and ItaParlCorpus`), institutional roles (`chair`, `cabinet`), and the speech content itself (`text`).

4. Methodology

4.1. BERTopic Application

The first part of the analysis involved the application of BERTopic. A preprocessing step was done aggregating all interventions by the same speaker within the same parliamentary session into single documents to capture complete speech events. Moreover, procedural interventions, such as speeches by the chamber chair, were excluded because they lack substantive thematic content. Following BERTopic methodology, first document embeddings were generated using the multilingual SentenceTransformer model, `paraphrase-multilingual-MiniLM-L12-v2` [2], with longer texts split into overlapping segments. Then, dimensionality reduction and clustering were performed using UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017), with hyperparameters optimized via Optuna (Akiba et al., 2019) to maximize topic coherence and diversity. A custom CountVectorizer based on lemmatized texts restricted to nouns and adjectives, ensured transparent, human-interpretable c-TF-IDF topic representations, resulting in a total of 49 coherent topics. Table 1 shows some of one of the extracted topics together with their most representative words. Moreover, to investigate the temporal evolution of the extracted topics, the Dynamic Topic Modeling (DTM) functionality provided by BERTopic was exploited.

4.1.1. Quantitative Evaluation

Quantitative performance of the trained model was assessed using the OCTIS framework (Terragni et al., 2021) through Topic Coherence and Topic Diversity metrics. Topic Coherence, ranging from -1 to 1, evaluates the semantic consistency of each topic, while Topic Diversity, ranging from 0 to 1, measures the proportion of unique words across topics. The model achieved a Topic Coherence

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KUARWD

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Topic ID	Top 10 Words per Topic
5	terrorismo (terrorism), terrorista (terrorist), politico (political/politician), terroristico (terroristic), polizia (police), lotta (fight), internazionale (international), interno (internal), fermo (arrest), atto (act)
6	film (film), spettacolo (show), cinematografico (cinematographic), cinema (cinema), teatro (theatre), censura (censorship), turismo (tourism), musicale (musical), artistico (artistic), culturale (cultural)
14	lingua (language), linguistico (linguistic), minoranza (minority), bolzano (Bolzano), tedesco (German/german), statuto (statute), provincia (province), ladino (Ladin), tutela (protection), regione (region)
15	religione (religion), religioso (religious), cattolico (Catholic), insegnamento (teaching), scuola (school), chiesa (church), concordato (concordat), insegnante (teacher), intesa (agreement), confessione (denomination)
33	tossicodipendente (drug addict), tossicodipendenza (drug addiction), droga (drug), metadone (methadone), recupero (rehabilitation), terapeutico (therapeutic), sert (SERT), danno (harm), riduzione (reduction), sostanza (substance)

Table 1: shows a subset of extracted topics with their 10 most representative words

of 0.14 and a Topic Diversity of 0.73 with values that fall within the ranges observed for BERTopic across different dataset (Grootendorst, 2022).

4.1.2. Qualitative Evaluation

To move towards the formalization of topics as Linguistic Linked Open Data, a qualitative evaluation was conducted to assess whether the extracted topics were meaningfully interpretable. A Human–LLM annotation agreement approach was adopted, using GPT-5.2 as an auxiliary annotator. From the initial 49 topics, 21 were selected based on lexical coherence, internal consistency, and clear semantic focus, while excluding topics with highly mixed vocabularies or predominantly procedural language. A human annotator assigned descriptive labels, against which those generated by GPT-5.2 were compared. The model was prompted with the ten most relevant words per topic and five representative documents to generate one primary label, a short description, and three alternative labels [3]. Using the `all-MiniLM-L6-v2 model`[4], for each topic, cosine similarity was measured between the embedding of the human-assigned label and the embeddings of the LLM-generated labels, including both the primary label and the three alternatives. The maximum similarity value across these four comparisons was retained as the final similarity score for the topic. In this way, the agreement captured whether the model was able to produce at least one label semantically close to the human interpretation. A similarity threshold of 0.70 was used to determine semantic alignment. Topics with a maximum similarity equal to or above this value were therefore classified as semantically aligned. The resulting Human–LLM Semantic Agreement Rate indicated a substantial level of alignment between human judgments and model-generated labels, reaching a rate of 82.14%, with at least one of the labels suggested by the model sufficiently similar to the human-assigned label according to the threshold. This outcome suggested that the extracted topics were largely interpretable and could be meaningfully summarized using concise semantic labels. Importantly, this evaluation did not claim that the LLM reproduced human annotations perfectly. Rather, it demonstrated that the model was generally capable of proposing plausible and appropriate descriptions for the topics. The SKOS vocabulary was employed to model these topic labels. For each topic, a primary human-readable label was assigned through the `skos:prefLabel` property. When the cosine similarity between the human-assigned label and at least one LLM-generated label reached or exceeded the 0.70 threshold, the preferred label was selected as either the human label or the LLM-generated label with the highest similarity. If no LLM-generated label met the threshold, the human-assigned label was retained as the preferred label. All additional labels associated with the topic were represented using the `skos:altLabel` property, thereby preserving alternative valid formulations. In the present work, provenance information regarding the labels (e.g. whether they were produced by a human annotator or generated by an LLM), was not recorded. Explicitly tracking such information would enhance transparency and reproducibility, and thus represents a direction for future improvement.

[3] The exact prompt template used for topic labeling is provided in Appendix A.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

4.2. Modeling

The modeling of the extracted topics was done by declaring a dedicated TM (Topic Modeling) ontology. The ontology is formally declared as an OWL ontology under the identifier `:TMExtensionOntology`. It explicitly imports the OntoLex and FrAC vocabularies, indicating that the proposed model is designed to extend these frameworks. OntoLex provides the linguistic layer needed to represent lexical forms and their meanings, while FrAC supplies the observation-based structure required to model topic modeling results. In this study, OntoLex-Lemon is employed in its Core module (see Figure 1) to provide a precise linguistic interpretation of the lexical dimension of topics.

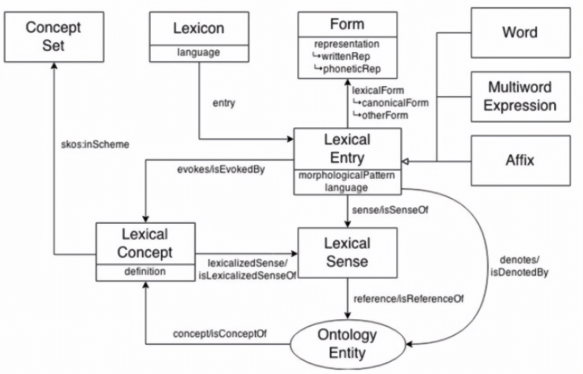

Figure 1: illustrates the core elements of the Ontolex-Lemon model (Cimiano et al., 2016).

Following the canonical structure of the Core module, the connection between word forms and their lexical entry is modeled using the property `ontolex:otherForm`, designed to link a lexical entry to a form, which is not a lemma and which realizes the given lexical entry. In particular, for each topic, the ten most highly associated words are modeled as OntoLex Forms. These forms correspond to the observed word forms in the corpus that are most strongly associated with a given topic according to the c-TF-IDF computed by BERTopic. Using the property ontolex:canonicalForm, each lexical entry is also linked to its canonical form identified using LiITA (Litta et al., 2025), a Linked Open Data knowledge base of interoperable linguistic resources for Italian. The linkage relies on LiITA's Lemma Bank, a collection of Italian lemmas modeled using the OntoLex-Lemon vocabulary and designed to interlink distributed lexical and textual resources. Moreover, the relationship between lexical entries and the lexical concepts they evoke is also modeled using the OntoLex-Lemon Core module, specifically through the `ontolex:evokes` property. This relation, in fact, is employed to align the lexical concepts with BabelNet synsets. BabelNet (Navigli and Ponzetto, 2010) is a large-scale multilingual encyclopedic dictionary and semantic network in which synset represents a distinct meaning and contains all synonyms expressing that meaning across multiple languages. By employing OntoLex-Lemon, topics are not merely sets of weighted word forms, but are explicitly linked to normalized lemmas and lexical concepts, enabling richer interpretation, supporting semantic interoperability with external resources, and allowing analysis at both the lexical and conceptual levels. Differently, the FrAC model (Chiarcos et al., 2022) extends OntoLex-Lemon by integrating corpus-based evidence and explicit links between lexical resources and corpora. FrAC is here employed specifically for the classes `frac:Observable` and `frac:Observation` which enable the explicit modeling of anything that can be observed, described, or quantified in relation to a lexical entity or concept (see Figure 2).

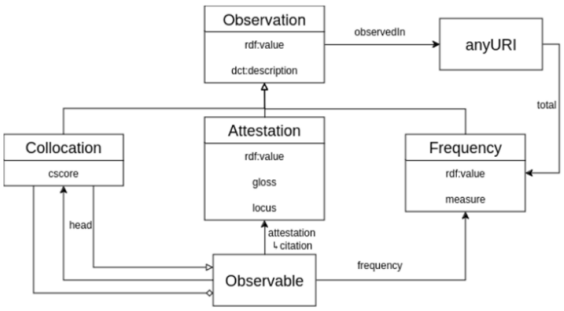

Figure 2: shows the OntoLex Module for Frequency, Attestation and Corpus Information (FrAC) (Chiarcos et al., 2025).

The class Observable represents any lexical entity that can be defined within the OntoLex vocabulary and potentially found in a corpus. This includes canonical and inflected forms, lexical entries, lexical senses and lexical concepts, all of which are treated as observables allowing the model to attach rich information to them. Complementing the observable, a `frac:Observation` represents any empirical measurement or annotation that is derived from or computed over a corpus and which concerns a certain observable. In the context of this study, FrAC is adopted to model topics as analytical results that groups word forms (`ontolex:Form`) and that are classified as subclasses of `frac:Observable`, `frac:Observation` and `rdfs:Containers`. Figure 3 provides an example of this modeling choice. Moreover, the FrAC treatment of collocations is particularly relevant for this study, as it directly informed the modeling choices adopted for topics in the proposed ontology. In FrAC, collocations are represented as `rdfs:Containers` of `frac:Observables`, capturing the aggregation observables based on their co-occurrence

within the same context window. This aggregation-based modeling strategy is reused for topics: similarly to collocations, topics are defined as `rdfs:Containers` of observables. In our ontology, these observables correspond specifically to `ontolex:Form`s, that is, the word forms extracted by the topic modeling process.

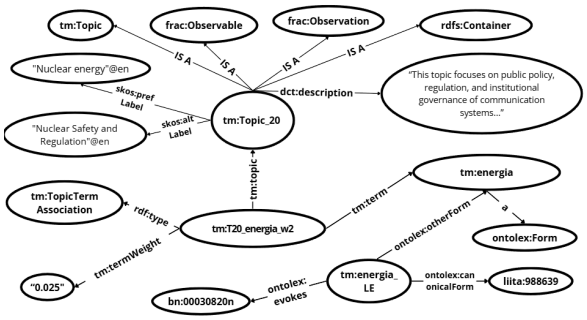

Figure 3: illustrates the modeling of Topic 20 as a `tm:Topic`, `frac:Observable`, `frac:Observation`, and `rdfs:Container`. The topic is associated with its preferred and alternative labels, as well as the LLM-generated description derived from the qualitative evaluation. Its lexical representation is formalized using specific relations from the TM ontology and standard OntoLex-Lemon properties.

4.2.1. Classes

Five classes are defined within the TMExtension ontology: `:DocumentTopicAssociation`, `:TopicTermAssociation`, `:Topic`, `:TopicModelRun` and `:Document`. The class `:DocumentTopicAssociation` represents an atomic observation linking a document to a topic. This observation is associated with a topic association weight that quantifies the strength of the relationship between the specific document and a certain topic. These scores capture the degree of closeness between a document and a topic as produced by the BERTopic. More specifically, these values reflect the distance between the embedding of the document and the centroid of the cluster that defines the corresponding topic. Therefore, these scores should be interpreted as soft-clustering confidence values rather than true probabilistic assignments. Complementarily, `:TopicTermAssociation` class models the internal structure of topics by formalizing the association between a topic and the lexical items that characterize it. This class represents atomic observations linking a topic to a lexical form, an OntoLex `ontolex:Form`, together with a weight computed by the topic modeling algorithm via the c-TF-IDF. As with document–topic associations, this modeling choice ensures that term relevance within topics is represented explicitly. `:Topic` is defined as a subclass of `rdfs:Container`, `frac:Observable`, and `frac:Observation`. As an `rdfs:Container`, a topic behaves as an aggregate entity collecting multiple members. Its members are explicitly constrained to be `OntoLex:Form` through an OWL restriction. The decision to model topic terms as `ontolex:Form` rather than `ontolex:LexicalEntry` is motivated by the need to avoid assigning ontological significance to what is merely an operational preprocessing choice. In particular, whether the corpus is lemmatized or not is a technical decision that should not determine the ontological status of the modeled entities. By adopting this approach, the representation remains ontologically neutral and directly reflects the behavior of topic modeling algorithms, which operate on the forms as they appear in the input corpus. Since a topic model simply processes the observed textual forms, if these forms are lemmas, this does not alter the nature of the computation. Representing topic terms as `ontolex:Form` therefore allows us to reflect the behavior of topic modeling algorithms, which operate over observed textual forms, without imposing additional lexical assumptions. Moreover, being defined also as subclass of `frac:Observable`, a topic is a phenomenon that can be observed in a corpus, allowing further observations to be made, including its frequency or distribution across documents. At the same time, as a subclass of `frac:Observation`, a topic is modeled as an analytical result produced by a computational process, capturing its provenance and reproducibility across multiple runs. In this way, the ontology formally links the output of topic modeling with structured and machine-readable lexical representations. Moreover, the computational provenance of topic modeling results is represented by the class `:TopicModelRun`, defined as a subclass of `prov:Activity`. Each instance corresponds to a concrete execution of a topic modeling algorithm with a specific configuration. This explicit representation supports transparency, reproducibility, and the systematic comparison of results across different modeling runs. Considering the class `:Document`, defined as a subclass of `rdfs:Resource`, it includes aggregated documents which are derived from the original documents collected in the ItaParl corpus and which are used as input for topic modeling. In fact, although the ItaParl Corpus is originally organized as individual speaking turns, with each row corresponding to a single intervention by a speaker in a parliamentary session, our analysis is conducted at a higher level. All interventions delivered by the same speaker during a single session are merged into

one document. Consequently, each :Document in the ontology represents the complete set of interventions made by a single speaker during one parliamentary session.

4.2.2. Properties

A set of properties is introduced to specify how entities are described, linked to one another, and quantitatively characterized. These properties are organized into annotation properties, object properties, and datatype properties, according to their modeling purpose. :termWeight is a datatype property that assigns a numerical value to a term within a topic. Its domain is :TopicTermAssociation, and its range is xsd:decimal. This value represents the relevance or contribution of a specific lexical form to a topic, typically computed using metrics such as c-TF-IDF, where higher values indicate stronger association between the term and the topic. :topicWeight is a datatype property that quantifies the strength of a topic within a document or text segment. Its domain is :DocumentTopicAssociation, and its range is xsd:decimal. Considering object properties, the relation :document links instances of :DocumentTopicAssociation, that is to say a soft-clustering weight associated to that document for that topic, to the document or text considered. Its range is rdfs:Resource, providing flexibility to represent documents at different levels of granularity, including full texts, paragraphs, or externally defined resources. The :term relation connects instances of :TopicTermAssociation to the lexical items that more prominently represent a topic. While no explicit range is imposed, it is modeled so that it typically points to OntoLex entities, such as ontolex:Form. The :topic property links an observation, either a :DocumentTopicAssociation or a :TopicTermAssociation, to the corresponding :Topic. By defining the domain as the union of these two association classes, the property reflects that both document–topic and topic–term relationships are modeled as FrAC-style observations. Moreover, speaker information is represented via the :speaker property. It associates a :Document with its author or speaker and its range is defined as owl:Thing, allowing alignment with external systems including FOAF or Wikidata. The :year relation is a datatype property associated with the class :Document and has a range of xsd:Year. It records the year corresponding to the document, allowing temporal metadata to be included in the ontology. Finally, :wasGeneratedByRun captures provenance information by linking any frac:Observation to the :TopicModelRun that produced it. Defined as a subproperty of prov:wasGeneratedBy, it preserves full compatibility with PROV-O while explicitly connecting observations to the computational run that generated them, supporting transparency and reproducibility. The overall structure of the ontology, including the relationships between topics, documents, lexical items, and computational runs, is summarized in Figure 4.

5. Discussion

5.1. Modeling Challenges

When modeling the linguistic-side of each topic in linked data, some critical aspects emerged. Topic modeling methods, including BERTopic, produce topics represented by decontextualized word forms ranked by representativeness, without preserving the syntactic or semantic context required to determine, for instance, a unique part-of-speech assignment. Consequently, when dealing with ambiguous words such as “*americano*”, multiple ontolex:LexicalEntry instances point to the same form via ontolex:otherForm, while ontolex:canonicalForm links each lexical entry to its own lemma. This modeling choice reflects the fact that topic modeling algorithms output isolated word forms; as a result, it is not possible to establish whether "*americano*", as it appears in a given topic, realizes a lexical entry whose canonical form corresponds to the lemma having part-of-speech ’noun’ or to the lemma having part-of-speech ’adjective’. Moreover, considering the ontolex:LexicalConcept instances evoked by each lexical entry, rather than linking all possible lexical concepts a lexical entry might evoke, which would effectively create a conventional dictionary, only those plausible lexical concepts that a given lexical entry may evoke within the specific context of its topic are recorded. This approach can be seen as a preliminary step toward potential future word sense disambiguation (WSD) experiments using topic modeling. At the same time, it highlights a limitation of the OntoLex–Lemon model, which does not allow the direct association of a word form, as collected within a specific topic, with a selected lexical concept, which would have been ideal for this use case. For example, it is not possible to directly link the word form “*onda*” (“*wave*”), appearing in a topic labelled as ‘*Electromagnetism and Health Risks*’, to the specific BabelNet concept corresponding to electromagnetic wave. Establishing such a connection would require defining a custom property whose semantic meaning would need to be explicitly specified. Importantly, the relation between a word and its meaning should not be interpreted as a direct link between an abstract form and a particular sense. Instead, the association should be understood as a relation between a lexical concept and the usage of that word across the documents

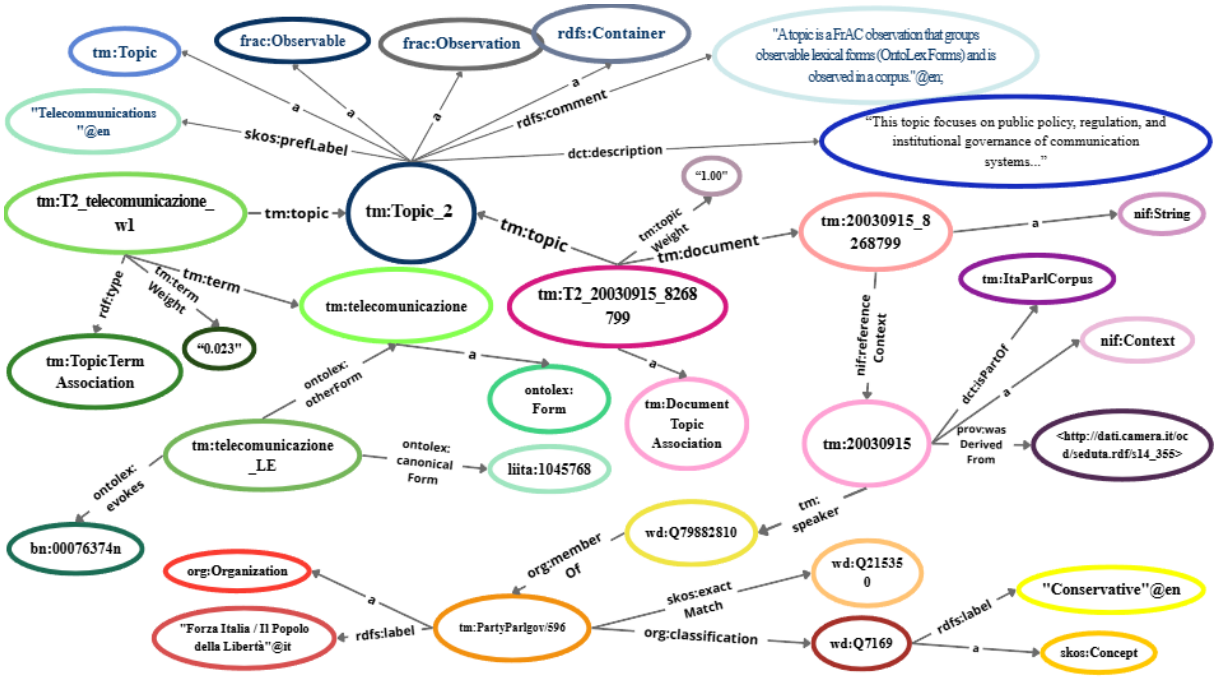

Figure 4: formalizes the representation of Topic 2, labeled as 'Telecommunications' and linked both to its lexical representation (here only the first most representative word is shown) and to one of the document that discuss this topic, together with information related to the speaker associated to the document, its political party and political ideology.

belonging to a given cluster. In other words, meaning selection emerges from the co-occurrence of words within the same document cluster, that is, within the same topic rather than from the word considered independently. This perspective also clarifies intuitions about meaning plausibility. In the case of "*onda*" within the Electromagnetic Health Risks topic, it is highly unlikely that the form evokes the concept of a sea wave. This judgment reflects the thematic coherence of the topic, which provides the contextual information necessary to resolve lexical ambiguity. These considerations highlight opportunities for further methodological refinement and motivate the discussion of possible extensions to the current modeling approach.

5.2. Future Work

Several directions for future research emerge from the present study. Topic modeling methods such as BERTopic generate topics as ranked lists of decontextualized word forms, which poses challenges when aligning them with lexical resources through OntoLex–Lemon. In particular, the model does not currently allow a direct association between a word form occurring within a specific topic and a selected lexical concept. Exploring the design and semantic definition of a dedicated property to represent context-dependent meaning selection therefore represents a promising direction for future work. Ways to represent the idea of '*plausibility*' of a certain lexical concept within the context of a certain topics could be defined more explicitly. One possibility is to model weighted or probabilistic associations between word forms and lexical concepts. Another possibility is to treat the selection of lexical concepts as an explicit interpretative decision, documented through additional metadata such as confidence scores. In this way, the inherently interpretative nature of lexical disambiguation would be made explicit and traceable, rather than remaining implicit in the data model. Furthermore, although the present analysis focuses on single-word units (nouns and adjectives), incorporating multi-word expressions (MWEs) could provide richer and more informative representations of complex topics. This approach could enhance topic interpretability, partially address issues arising from decontextualized words and ambiguous part-of-speech mappings, and remain fully compatible with OntoLex-Lemon modeling. Integrating MWEs therefore constitutes a valuable extension of the present work. Another promising direction for future work related to the representation of topics in Linguistic Linked Open Data could pose particular attention to the topics distributional semantics and temporal dynamics. This includes both the explicit modeling of documents embeddings derived from BERTopic, and the formalization of topics as dynamic and evolving entities. In fact, BERTopic identifies topics starting from clusters of contextual-

ized embeddings representing documents, on the assumption that documents discussing the same topic will be semantically similar and therefore positioned closer together in the vector space. The OntoLex-FrAC framework allows for explicitly representing numerical projections of linguistic data. In FrAC, the class `frac:Embedding` is defined as a representation of a given Observable in a numerical feature space, while the object property `frac:embedding` is used to link an attestation to its numerical representation. In this perspective, FrAC would provide a way of connecting textual evidence, numerical representations and higher-level semantic abstractions within Linguistic Linked Open Data. In this sense, FrAC could be used to link documents to their numerical representations, and to connect topics to clusters of documents embeddings. Each document would thus be associated with a corpus-based textual realization and a corresponding numerical embedding, while the topic would be defined by the set of its associated documents embeddings, reflecting the clustering mechanism underlying BERTopic. A more precise account of the relations and properties involved would require further refinement and should be addressed explicitly. Moving beyond static representations and considering topics as dynamic entities, another future work could address the temporal evolution of topics, in line with BERTopic's dynamic topic modeling approach. In particular, BERTopic distinguishes between global tuning and evolutionary tuning. While global tuning ensures that topics maintain a coherent identity across the entire corpus, evolutionary tuning refines topic lexical representations within individual time slices, allowing to capture shifts in topical vocabulary without changing document–topic assignments. The FrAC extension provides a suitable formal foundation for modeling this aspect through the class `frac:TimeSeries`. This class is defined as a subclass of `frac:Embedding` that represents an observable or its attestation as a sequence of fixed-size numerical representations recorded over time. This strategy can thus be used to model different temporally evolving observables which, in the proposed ontology, correspond to the extracted topics. Considering the prevalence of a topic over time, they could be captured as a sequence of numerical values, each indicating the proportion of documents assigned to that topic within a given time slice. In this representation, each value reflects how prominently the topic is represented in the corpus at a specific moment, for example by expressing the proportion of documents associated with the topic in a particular year. Moreover, the lexical representation of a topic as it changes over time could also be formalized. An idea could be the one of using a sequence of fixed-size vectors, where each vector corresponds to a time slice that encodes the relative importance of the topic's representative terms during that time period. For instance, a topic might be associated in one specific time step with a vector such as `[0.5, 0.3, 0.2]`, where each value encodes the weight of a particular term during that period. In this way, the representation makes it possible to track how the vocabulary characterizing a topic shifts across time, while maintaining a stable association between the topic and its document set. Modeling topics in this way would allow Linked Data representations to capture not only what a topic is about, but also how its prominence and lexical realization evolve over time. Further research could also enhance the analytical and comparative potential of the framework by expanding the number of documents formalized per topic, as the present study models only a limited subset. Extending the approach to multilingual parliamentary corpora would enable cross-country comparisons of political discourse, leveraging the OntoLex–Lemon model and shared conceptual resources such as BabelNet to align topic representations across languages. Finally, integrating structured representations of historical events constitutes another particularly interesting direction. Linking topic dynamics to external events (such as elections, governmental changes, or international crises) would strengthen the explanatory power of the analysis by situating parliamentary discourse within broader historical processes. Such integration would further reinforce the role of Linked Open Data as a bridge between computational linguistics analysis and historical knowledge, enabling complex queries that jointly consider topics, time, actors, and events.

6. Conclusions

This work examined parliamentary discourse as a domain in which political actors articulate policy positions and construct representations of social and political issues through language, focusing on debates in the Italian Chamber of Deputies between 1948 and 2006. Using a BERTopic-based dynamic topic modeling approach, latent semantic themes were identified and analysed, tracing their temporal evolution and lexical variation across decades of parliamentary activity. Quantitative and qualitative evaluations ensured the interpretability and reliability of the extracted topics, enabling their subsequent formalization within a semantic framework. Building on these results, the study proposed a systematic approach for representing computationally derived topics as Linked Open Data entities from a linguistic perspective, leveraging OntoLex-Lemon and its FrAC extension. Topics were modeled as structured semantic objects interconnected with

documents, speakers, political parties, and lexical elements, addressing the challenges of representing data-driven and abstract constructs within Semantic Web formalisms. The resulting knowledge graph, explored through SPARQL queries [5], demonstrates how the integration of dynamic topic modeling and Linked Data supports multi-level analyses of parliamentary discourse, linking thematic structures with temporal, institutional, and lexical dimensions. This contributes to linguistic research on parliamentary discourse by providing access to the lexical and semantic realizations of topics, allowing researchers to track term variation as well as identify emerging terminology and shared vocabularies. Overall, the integration of Natural Language Processing and Semantic Web technologies enhances the interpretability, interoperability, and reusability of topic modeling outputs, transforming them from isolated analytical results into semantically structured research data. The proposed framework is extensible to broader temporal coverage, additional parliamentary or heterogeneous corpora, and multilingual contexts, providing a foundation for future research at the intersection of computational text analysis and Linked Open Data.

7. Data and Code Availability

All resources produced in this study, including BERTopic implementation, the declared ontology, and the resulting knowledge graph, are publicly available in the project GitHub repository: https://github.com/Lisaalbertelli/tmextension-ontology

8. Ackowledgments

Grateful acknowledgment is made to Professors Federica Iurescia and Francesco Mambrini for their guidance.

9. References

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

U. Bojārs, R. Dargis, U. Lavrinovičs, and P. Paikens. 2019. Linkedsaeima: A linked open dataset of latvia's parliamentary debates. In *Semantic Systems: The Power of AI and Knowledge Graphs*, volume 11702 of *Lecture Notes in Computer Science*, pages 49–63, Cham. Springer.

C. Chiarcos, E.-S. Apostol, B. Kabashi, and C.-O. Truică. 2022. Modelling frequency, attestation, and corpus-based information with ontolex-frac. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

C. Chiarcos, J. P. McCrae, and M. Ionov. 2025. The ontolex module for frequency, attestation and corpus information. W3C Ontology-Lexica Community Group.

P. Cimiano, J. P. McCrae, and P. Buitelaar. 2016. Lexicon model for ontologies: Community report. W3C Community Group.

J. Cova. 2025a. ItaParlCorpus (Version 3.0) [Data set]. Harvard Dataverse.

J. Cova. 2025b. A new database for italian parliamentary speeches: introducing the itaparlcorpus dataset. *Italian Political Science Review/Rivista Italiana di Scienza Politica*, 55:1–10.

T. Erjavec, M. Kopp, N. Ljubešić, M. Ogrodniczuk, P. Pęzik, E. Sanders, and T. Wissik. 2025. Parlamint ii: Advancing comparable parliamentary corpora across europe. *Language Resources and Evaluation*, 59(2):2071–2102.

T. Erjavec, M. Kopp, M. Ogrodniczuk, P. Osenova, M. Agirrezabal, and D. Agnoloni, T. Fišer. 2023. Multilingual comparable corpora of parliamentary debates parlamint 4.0 (version 4.0) [data set]. CLARIN.SI.

M. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv.

E. Hyvönen, L. Sinikallio, P. Leskinen, J. Tuominen, H. Rantala, and M. Tamper. 2025. Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland. *Semantic Web – Interoperability, Usability, Applicability*, 16(1):1–25.

E. Litta, M. C. Passarotti, V. Basile, C. Bosco, A. Di Fabio, and P. Brasolin. 2025. Liita: a knowledge base of interoperable resources for italian. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 130–135. Unior Press.

J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, and P. Cimiano. 2017. The ontolex-lemon model: Development and applications. In *Proceedings*

[5] See Appendix B for a subset of representative SPARQL queries.

of the 5th Biennial Conference on Electronic Lexicography (eLex 2017), pages 587–597. Lexical Computing CZ s.r.o.

L. McInnes, J. Healy, and S. Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

L. McInnes, J. Healy, and J. Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv preprint.

R. Navigli and S. P. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

C. Rauh. 2020. The parlspeech v2 data set. Harvard Dataverse.

C. Rauh, P. De Wilde, and J. Schwalbach. 2017. The parlspeech data set. WZB Berlin Social Science Center.

S. Terragni, E. Fersini, et al. 2021. Octis: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

A. van Aggelen, L. Hollink, et al. 2017. The debates of the european parliament as linked open data. *Semantic Web*, 8(2):271–281.

D. Vrandečić and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

10. Appendix

10.1. Appendix A: Prompt Template

Table 2 reports the exact prompt template used in Human–LLM annotation agreement experiments (described in Section 4.1.2), where GPT-5.2 was employed. The prompt guided the model to generate one primary label, a short description, and three alternative labels for each topic.

10.2. Appendix B: SPARQL queries and results

This section presents a selection of SPARQL queries used to explore the knowledge graph. Once the data had been fully modeled, the resulting RDF triples were exported in Turtle format and queried using SPARQL. To this end, Apache Jena Fuseki[6], was employed as a SPARQL server, providing both a triple store and a web-based interface for query execution, running locally within a Java environment. The queries are designed to illustrate how different dimensions of the data can be accessed and combined. The complete set of queries and results is publicly available in the project's GitHub repository, highlighting the analytical potential of the proposed framework.

[6] https://jena.apache.org/

```
I have a topic from a topic model that I need to label.
Below are the most important words associated with this topic and the 5
most representative documents.
Topic Words (in order of importance):
{words_str}
Representative Documents:
{docs_str}
Based on this information, please provide:
1.  A concise, descriptive label for this topic (2-3 words)
2.  A brief explanation of what this topic represents
3.  Alternative label suggestions (2-3 options)
Please focus on capturing the main theme that connects both the words and
the document content.
Response format:
Primary Label:  [Your main label]
Explanation:  [Brief explanation]
Alternative Labels:  [Alternative 1, Alternative 2, Alternative 3]
```

Table 2: Prompt template used for topic labeling

```
PREFIX tm:  <http://example.org/tm/>
PREFIX org:  <http://www.w3.org/ns/org#>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?speaker ?partyLabel ?wikidataParty ?classificationLabel
?wikidataFamily

WHERE {
  ?docTopicAssoc tm:topic tm:topic_44 ;
                 tm:document ?document .
  ?document tm:speaker ?speaker .

  ?speaker org:memberOf ?party .
  ?party rdfs:label ?partyLabel ;
        skos:exactMatch ?wikidataParty .

  ?party org:classification ?wikidataFamily .
  ?wikidataFamily rdfs:label ?classificationLabel .
}

ORDER BY ?classificationLabel ?partyLabel ?speaker
```

Table 3: Example of a SPARQL query to identify the speakers who participated in parliamentary debates concerning the “Environmental and climate issues” topic, while also retrieving their political party affiliations and the corresponding political families.

```
PREFIX tm:  <http://example.org/tm/>
PREFIX org:  <http://www.w3.org/ns/org#>
PREFIX skos:  <http://www.w3.org/2004/02/skos/core#>
PREFIX wd:  <http://www.wikidata.org/entity/>
PREFIX wdt:  <http://www.wikidata.org/prop/direct/>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?topicLabel ?speaker ?genderLabel ?partyWD ?ideologyWD
?ideologyLabel

WHERE {
  ?dta a tm:DocumentTopicAssociation ;
      tm:topic ?topic ;
      tm:document ?document .
  ?topic skos:prefLabel ?topicLabel .
  ?document tm:speaker ?speaker .
  ?speaker org:memberOf ?party .
  ?party skos:exactMatch ?partyWD ;
        org:classification ?ideologyWD .
  FILTER (?ideologyWD IN (wd:Q7169, wd:Q76074))
  ?ideologyWD rdfs:label ?ideologyLabel .
  SERVICE <https://query.wikidata.org/sparql> {
    ?speaker wdt:P21 wd:Q6581097 .
    wd:Q6581097 rdfs:label ?genderLabel .
    FILTER (LANG(?genderLabel) = "en")
    }
}
ORDER BY ?topicLabel ?ideologyLabel ?speaker
```

Table 4: Example of federated SPARQL query combining speaker gender and political affiliation to retrieve topics discussed by male politicians belonging to conservative or right-wing parties.

Towards the LinkEn Knowledge Base. A Neuro-Symbolic approach to build a Linked Data hub for English lemmas with Large Language Models

Lorenzo Augello, Marco Passarotti
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123 Milan, Italy
lorenzo.augello01@icatt.it, marco.passarotti@unicatt.it

Abstract

This paper presents the first core component of LinkEn, a knowledge base of interoperable language resources for English adhering to Linked Open Data principles. With this initial step towards a broader infrastructure, we focus on the development of a lemma-centered hub designed to enable interoperability between distributed lexical resources, corpora, and linguistic annotations. The modeling is inspired by the LiLa Knowledge Base for Latin and the OntoLex-Lemon model, ensuring compatibility with existing lemma-centric knowledge graphs and enabling future cross-linguistic interoperability. Rather than relying solely on manual knowledge graph construction and significant human effort, the lemma bank has been developed through a hybrid neuro-symbolic pipeline that integrates large language models into the generation of RDF data under explicit ontological constraints. This approach combines automated generation with ontology-driven supervision and evaluation, enabling scalable yet controlled construction of structured lexical knowledge. By presenting the first steps towards the LinkEn Knowledge Base, this paper contributes both a new lemma bank for English and an experimental methodology for the semi-automatic creation of Linked Data based knowledge graphs.

Keywords: Linguistic Linked Open Data, English, Knowledge Base, Lemmas

1. Introduction

The increasing availability of digital language resources has highlighted the need for infrastructures that enable their integration, interoperability, and reuse. The Linguistic Linked Open Data[1] (LLOD) community addresses these challenges by promoting the publication of language resources according to the principles of the Linked Data paradigm.[2] In LLOD, the various components of a resource (e.g., lexical entries, sentences, words) are assigned URIs, enabling distributed resources to be linked and queried in an interoperable network.

Within this context, lemma banks have emerged as especially valuable linked collections of citation forms, distinguishing themselves from other lexical databases by serving as interlinking hubs between lexical entries and corpus annotations. A prominent example is the LiLa Knowledge Base for Latin (Passarotti et al., 2020), where each lemma is assigned a stable URI and used to link lexical resources and corpora within a unified Knowledge Base (KB).

Building on this paradigm, a few related lemma bank initiatives have recently appeared for other languages. For Italian, the LiITA KB (Litta et al., 2024) includes a lemma bank at its core, put together starting from selected lemma sets and designed to create interoperability between available resources for Italian. For Old Irish, the MOLOR project (Fransen et al., 2024) has similarly adopted the LiLa ontology and the OntoLex-Lemon framework (McCrae et al., 2017) to construct a lemma bank for a less-resourced language with low digital availability.

Despite these advances, English lacks a lemma bank published as linked data in a LiLa-style manner. Indeed, existing English lexical resources and networks such as WordNet (Fellbaum, 1998) are typically not shaped to function as hubs for lexical citations. The CLARIN Virtual Language Observatory[3] lists 26,790 resources dedicated to the English language, but they all encode information in different formats, and with various levels of granularity and annotation criteria; hence, the absence of a standardized lemma hub limits interoperability and cross-resource querying and linking.

Even though the LiLa project provides an inspiring example, the systematic construction and publication of lemma banks present significant challenges with a time-consuming procedure that requires a high degree of human effort and supervision. With this work, we propose a new methodology for the construction of a novel English lemma bank towards the creation of the LinkEn KB, including Large Language Models (LLMs) into a hybrid pipeline, contributing with a LLOD application to the broader field of Knowledge Graph (KG) construction assisted by LLMs (Zhu et al., 2024).

[1] https://linguistic-lod.org/

[2] https://www.w3.org/DesignIssues/LinkedData.html

[3] https://vlo.clarin.eu

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 13–21
12 May 2026.

We focus on generating a lemma bank for English, starting from selected lemma sets, modeled according to the LiLa ontological framework and the OntoLex-Lemon model. In doing so, we outline a specifically-tailored hybrid neuro-symbolic methodology towards the construction of LinkEn, an initial English KB aligned with the Linked Data principles. This is queryable and published together with a sparql endpoint[4] and a visualization tool.[5]

In Section 2 we provide a theoretical framework for lemma banks to which LinkEn is aligned, while in Section 3 we describe the LiLa Lemma Bank, and present recent studies on KG construction. Section 4 outlines our hybrid LLM-human methodology, and an evaluation of the final results is provided in Section 5. Details of the present lemma bank of the LinkEn KB are given in Section 6, while future efforts towards its expansion are proposed in Section 7.

2. Lemma Banks

Within the landscape of lexical resources, lemma banks occupy a distinctive role. A lemma is the canonical citation form of a word, i.e., the canonical form that is used (or may potentially be used) by language resources to lemmatize word forms or to index dictionary entries (Passarotti and Mambrini, 2022). Conventionally, lemmatization is defined in linguistics and lexicography as the task of reducing the multiple inflected forms of a word to a form unanimously recognized as canonical (i.e., the word form reported as an entry in dictionaries). Even though serving as a linked data hub is the purpose of a lemma bank, the existence of a lemma in a lemma bank does not depend on its linking to a lexical entry in other resources. Building on this notion, a lemma bank is a curated repository of lemmas that can be enriched with their associated grammatical, morphological, and lexical information. Lemma banks are essential for linguistic research and Natural Language Processing (NLP) downstream tasks, as they provide the anchor for linking inflected forms to their canonical representation. They also serve as entry points for lexical-semantic information, as lemmas are often the nodes that connect morphological data, syntactic dependencies, semantic relations and textual occurrences.

Lemma banks serve as a base and connection point that facilitates interoperability in the Linked Data ecosystem, as they offer stable identifiers for lexical items that can be linked across resources.

[4] https://linken-lod.eu/sparql
[5] https://lodview.it/

3. Related Work

3.1. LiLa Lemma Bank

The LiLa (Linking Latin) ERC-funded project (2018-2023) represents one of the most comprehensive implementations of the Linked Data paradigm for language resources. Its primary goal is to interconnect distributed lexical resources, annotated corpora, and NLP tools for Latin through a unified infrastructure based on shared Semantic Web standards (Berners-Lee et al., 2001). At the core of the LiLa KB lies a large lemma bank (Passarotti et al., 2020), designed as a central hub for interoperability across resources. By assigning stable identifiers to lemmas and linking lexical entries and corpus tokens to these identifiers, LiLa enables integrated querying and navigation across otherwise heterogeneous datasets. The current LiLa Lemma Bank comprises over 230,000 canonical forms for Latin,[6] demonstrating the scalability of lemma-centric modeling within the Linked Data framework.

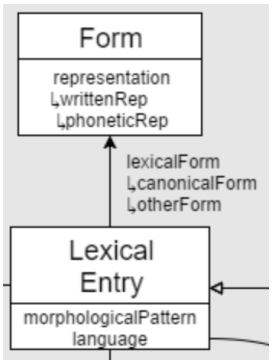

Figure 1: A section of the OntoLex-Lemon core model, specifically the relationship between the classes `ontolex:LexicalEntry` and `ontolex:Form`.

However, the LiLa Lemma Bank is not conceived as a lexical resource in the traditional sense. The modeling principles underlying LiLa are grounded in the OntoLex-Lemon model (McCrae et al., 2017), the de-facto standard for representing lexical information as Linked Data. Rather than instantiating lexical entries (`ontolex:LexicalEntry`), LiLa consists of entities representing canonical forms modeled as instances of the OntoLex class `ontolex:Form`. To support this approach, LiLa introduces the dedicated class `lila:Lemma`,[7] a subclass of `ontolex:Form`, representing canonical dictionary forms that serve as reference points for linking resources. Because each lemma is

[6] https://lila-erc.eu/lodview/data/id/lemma/LemmaBank
[7] https://lila-erc.eu/ontologies/lila/Lemma

modeled as a form, it can be connected to lexical entries in external resources through the property `ontolex:canonicalForm` (see Figure 1), thereby enabling interoperability across lexica and corpora. Each attestation of a word form in a textual resource (i.e., corpus token) can be linked to its respective lemma in LiLa through the property `lila:hasLemma`.[8]

Other than OntoLex-Lemon, the LiLa KB makes reference to classes and properties of already existing ontologies to model relevant information, such as POWLA for corpus data (Chiarcos, 2012) and OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015).

3.2. Large Language Models for Knowledge Graph Construction

Recent research has increasingly explored the use of LLMs for KG construction, motivated by their ability to encode large amounts of relational knowledge implicitly within their parameters (Pan et al., 2023).

From a methodological perspective, KG construction can be defined as a mapping from data sources and background knowledge to a structured graph representation (Zhong et al., 2023). Approaches to KG construction are commonly classified as supervised, semi-supervised, or unsupervised. Supervised and semi-supervised systems, such as Knowledge Vault (Dong et al., 2014) and Stanford OpenIE (Angeli et al., 2015), rely on predefined schemas, extraction patterns based on linguistic features, and varying degrees of human intervention. In contrast, fully unsupervised approaches such as MAMA (Wang et al., 2020) aim to recover factual knowledge directly from pretrained language models without explicit human supervision, offering insights into the knowledge encoded by neural models.

Different approaches to KG construction are characterized by the level of information provided to LLMs. In zero-shot methods (Carta et al., 2023), the LLM is prompted to identify relationships between data and define the schema for representing triples in the KG. Despite eliminating the need for a predefined representation schema, the drawback is that the output schema may not align with the desired level of granularity for information representation. More informed methods, based on zero/one/few-shot prompts and/or RAG techniques (Yang et al., 2025), provide the LLM with detailed instructions that enforce adherence to a predefined structure in order to construct the KG.

The LAMA benchmark (Petroni et al., 2019) was among the first systematic efforts to evaluate factual knowledge retrieval in pretrained language models. Subsequent studies, such as Zhong et al. (2021) and the KAMEL experiments (Kalo and Fichtel, 2022), have shown that while LLMs can recall many factual triples, their behavior often reflects memorization rather than reasoning, and their knowledge access remains limited compared to symbolic KBs.

Despite their potential, LLM-based approaches to KG construction exhibit notable limitations. Frameworks such as AutoKG (Zhu et al., 2024), which propose autonomous KG construction and reasoning via multi-agent LLM architectures, demonstrate promising results but also confirm that reliable KG construction still requires careful instruction design, high-quality input data, and robust evaluation methodologies. These limitations are particularly critical in the context of Linked Open Data, where formal correctness, ontological alignment, and interoperability are essential.

4. Methodology

Constructing a lemma bank for English using a LLM is a task that contained various challenges, from the data collection phase, to the modeling choices, the prompt design and the evaluation of the results.

Data collection and cleaning was performed from a more general "lemma" perspective and with more tailored devices for hypolemmas (see Section 4.2 for the distinction between lemmas and hypolemmas), which were generated separately starting from words with specific parts of speech (PoS). LLM prompting was also divided into two separate stages, one for hypolemma generation and one for RDF triples generation and PoS tagging for the rest of the input words. The outputs were evaluated inspecting both their formal validity and syntactic compliance to Turtle syntax (RDF validation, parsing and ontology alignment), and their content and encoded information, testing the LLM linguistic competence. For incorrect outputs, a looped process of reprompting and correction with manual supervision was conducted, before accepting and storing everything in the final lemma bank. The whole pipeline is reported in Figure 2.

Completely relying on a LLM was at the foundation of this work, as we wanted to assess its capability in (i) generating meta-linguistic knowledge starting from raw lemma sets, and (ii) structuring that information in a syntactically valid way.

For this task, we use Gemini 2.5 Flash[9] of GoogleAI, which provides a favourable combination of performance and cost.

[8] https://lila-erc.eu/ontologies/lila/hasLemma

[9] https://deepmind.google/models/gemini/flash/

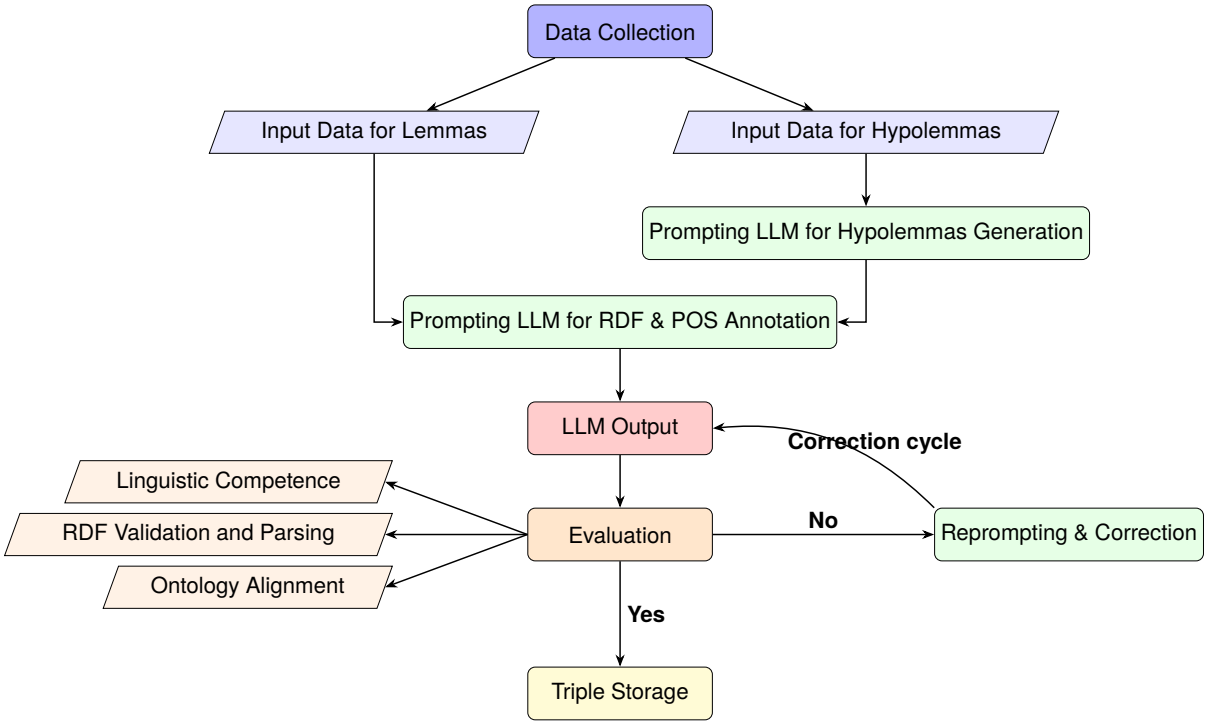

Figure 2: Workflow diagram of the proposed neuro-symbolic pipeline towards the construction of LinkEn.

4.1. Lexical Data

We started by gathering lexical data, and initial experiments were performed on a first set of lemmas that was obtained from the Kilgarriff list (Kilgarriff, 1997), a lemmatized frequency list for the 6,318 words with more than 800 occurrences in the whole 100M-word British National Corpus (BNC) (Consortium, 2007).

After this, a larger set of lemmas was chosen in order to expand the coverage of the lemma bank to more lexical items of the English lexicon, moving from words with more than 800 occurrences to words which occur at least 5 times in the BNC, for a total of 20,437 head words.[10] We gathered those from an English lemmas database[11] compiled by referencing the BNC, NodeBox Linguistics[12] (a Python library to do linguistic analysis) and other lemma lists[13] where word tokens are combined into lemma groups. Entries are listed and structured following the below format:

book -> booked,booking,books

write -> writes,writest,writing,written,wrote

After taking out all the overlaps between the Kilgarriff list and the second larger one (2,154), we also removed all the words that were not included in Open English WordNet (OEWN) (McCrae et al., 2019) (3,410). The alignment with OEWN was used as a criterion for reducing the size of the data to be handled in this initial stage of lemma bank creation, but its future expansion will not be limited to this and aims to cover as much as the English lexicon as possible. The exclusion process included highly specialized and technical words (e.g., *agribusinessman*, *baculovirus*), abbreviations (e.g., *smth*), acronyms (e.g., *pca*) and borrowings from other languages (e.g., *cruzeiro*), getting to a final list of 14,873 head words to give in input to the model.

4.2. Ontological and modeling choices

We started from the LiLa ontology,[14] but we deemed unnecessary to include it all, as that was specifically tailored for the modeling of the Latin language. The LiLa ontology includes classes, individuals and properties as detailed as a morphologically rich language such as Latin requires. Indeed, much of the meta-linguistic information that is necessary to describe a Latin lemma is rather superfluous for English, including inflectional classes and gender. Only the classes and properties related to the lemma, hypolemma, PoS and written representations have been picked and considered as essen-

[10] https://lexically.net/wordsmith/support/lemma_lists.html

[11] https://github.com/skywind3000/lemma.en/

[12] https://www.nodebox.net/code/index.php/Linguistics

[13] https://lexically.net/downloads/BNCwordlists/elemma.txt

[14] https://github.com/CIRCSE/LiLaOntologies

tial to create the core lemma bank for the LinkEn KB. In fact, unlike Latin, English nouns and adjectives do not have declensions, inflectional classes or gender, and verbs are not classified into distinct conjugations. The reduced custom ontology was provided to the model directly within the prompt, without pointing to any external file.[15]

Apart from the `lila:Lemma` class, an important subset of lemmas is categorized under the `lila:Hypolemma` class.[16] As well as a lemma, a hypolemma is still a form that is used (or may be used) by a lemmatizer to lemmatize a token. However, this form can also be analyzed as an inflected (or otherwise derived) form of another lemma. Prototypical examples of this are deadjectival adverbs and past and present participles used adjectivally. In fact, deadjectival adverbs are derived from adjectives (e.g., English adverb *slowly* derived from the adjective *slow*, see Figure 3) and assigned PoS ADV, while participles are inflected forms of their root verbs (e.g., English past and present participles *broken* and *breaking* are part of the inflectional paradigm of the verb *to break*) and assigned PoS ADJ in the lemma bank. Hence, the `lila:Hypolemma` class is a sub-class of `lila:Lemma`, and individual hypolemmas are linked to their related lemmas through the property `lila:isHypolemma`[17] (opposite to `lila:hasHypolemma`).

lila:POS Class	UPOS tag
lila:Adjective	ADJ
lila:Adposition	ADP
lila:Adverb	ADV
lila:CoordinatingConjunction	CCONJ
lila:SubordinatingConjunction	SCONJ
lila:Determiner	DET
lila:Interjection	INTJ
lila:Noun	NOUN
lila:Pronoun	PRON
lila:Verb	VERB

Table 1: Alignment of parts of speech to the UPOS tagset.

Since PoS make for the most important and only meta-linguistic information encoded for each canonical form in LinkEn, we aligned them to the well-defined standard represented by the UPOS tagset of Universal Dependencies[20] (see Table 1), and we mapped PoS for English aligning them with the

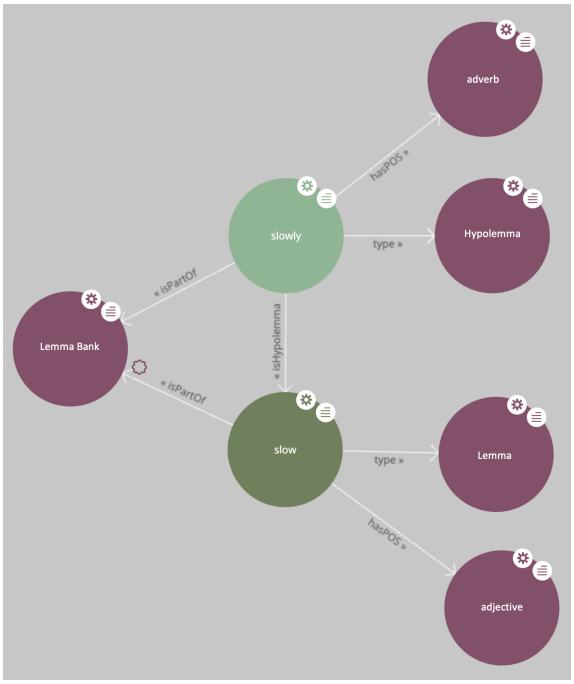

Figure 3: LodLive visualization of the deadjectival adverb *slowly*,[19] hypolemma of the base adjective *slow*.

vocabulary used in the LiLa ontology (Table 2 reports PoS frequencies in LinkEn). Every lemma and hypolemma are linked to their PoS through the `lila:hasPOS` property.[21]

Following an OntoLex-Lemon constraint on forms, a one-to-one correspondence must hold between forms and PoS, i.e., each lemma and hypolemma must have one and only one PoS. If a word could be tagged with more than one PoS, as many lemmas as the number of possible PoS must be created. For instance, the English word *book* could be both a noun and a verb, so instead of having only one lemma pointing to two different PoS, there must be two distinct lemmas with the same label but different IDs, one with PoS `lila:noun`[22] and one with PoS `lila:verb`.[23]

As well as LiLa, we use the OntoLex-Lemon framework to link each canonical form to its written representation, recorded as a literal, through the `ontolex:writtenRep` property.[24] Written representations are ortographical variants, which should be represented as different representations of the same form. For example, for the word *cen-*

[15] The ontology is available within the LinkEn repository at: https://github.com/lorenzoaugello/LinkEn

[16] https://lila-erc.eu/ontologies/lila/Hypolemma

[17] https://lila-erc.eu/lodview/ontologies/lila/isHypolemma

[20] https://universaldependencies.org/u/pos/

[21] https://lila-erc.eu/ontologies/lila/hasPOS

[22] https://linken-lod.eu/data/id/lemma/3060

[23] https://linken-lod.eu/data/id/lemma/971

[24] http://www.w3.org/ns/lemon/ontolex#writtenRep

PoS	Nodes	Frequency
ADJ	4,682	18.9%
ADP	63	0.25%
ADV	1,014	4.1%
CCONJ	7	0.03%
DET	43	0.2%
INTJ	26	0.11%
NOUN	13,022	52.6%
PRON	48	0.19%
SCONJ	25	0.11%
VERB	5,808	23.5%

Table 2: Total count and distribution of all parts of speech encoded in LinkEn, including both lemmas and hypolemmas.

tre, we would have two different representations of the same form, one for the British English written representation *centre* and one for the American English written representation *center*. If a word could have more than one written representation, only one lemma is created with multiple representations, without duplicating it (see Figure 4).

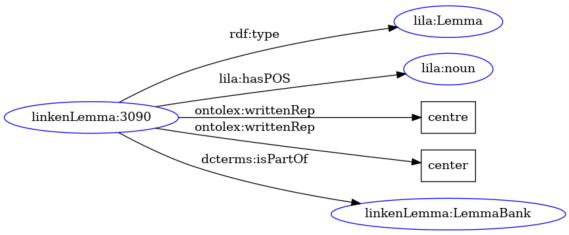

Figure 4: Visualization of the lemma *centre* with the `ontolex:writtenRep` data property pointing to two different strings.

4.3. Prompting

Having all the necessary starting lexical data as lemma lists, we asked the model to generate their PoS. For words that allowed for more than one PoS, a different lemma was to be created for each possible PoS. Beside PoS, the model was also prompted to generate multiple graphical variants, modeled through the `ontolex:writtenRep` property, for lemmas that allowed for more than one. Once all this linguistic and meta-linguistic information was produced, the LLM was also tasked with its structuring through well-defined web standards, organizing it all in RDF triples conforming to Turtle[25] syntax.

The following is a prompt template for a single lemma entry:

> Given the lemma "{lemma}", structure the following information in RDF format according to the example and the rules contained in the reported ontology. For words that can have more than one part of speech, generate as many lemmas as the possible parts of speech. All the possible parts of speech are included in the ontology. The lila:hasPOS Property does not have the lila:POS Class as range, but rather an individual of that class, as listed here: lila:noun, lila:adjective, lila:determiner, lila:adverb, lila:verb, lila:coordinating_conjunction, lila:subordinating_conjunction, lila:adposition, lila:pronoun, lila:interjection. If the lemma "{lemma}" could have more than one part of speech, generate two distinct lemmas with the corresponding parts of speech and with different unique identifiers. If the lemma "{lemma}" could have more than one written representation, generate them with the ontolex:writtenRep Property, as in the example of "analyse" (ontolex:writtenRep "analyse" , "analyze"). The output must be only the Turtle code and nothing else. Only the information about the lemma must be generated once and not repeated, and nothing else that is contained in the ontology. The information about the pre- fixes must not be generated in output. The following string "linkenLemma:LemmaBank a lila:LemmaBank ." must not be generated. Include everything between the string "'turtle at the beginning and "' at the end, without stopping and interrupting the triples, as in the following example:
>
> "'turtle
> linkenLemma:123 a lila:Lemma ;
> rdfs:label "absolute" ;
> lila:hasPOS lila:adjective ;
> ontolex:writtenRep "absolute" ;
> dcterms:isPartOf linkenLemma:LemmaBank .
> "'
>
> The following is the ontology to be compliant with, which must not be generated in output: [...]

The generation of hypolemmas was conducted in a separate setting. Two separate sets of base adjective (1,124) and verb (1,281) head words were derived from the Kilgarriff list in order to prompt the model to respectively produce their related deadjectival adverbs and past and present participles, where possible. This generated 936 deadjectival adverbs and 2,527 participles to be modeled as hypolemmas. For each of them, the LLM was asked to generate RDF triples, following a similar prompt to the one for simple lemmas. The following is an output example:

```
linkenIpoLemma:829 a lila:Hypolemma ;
rdfs:label           "slowly" ;
lila:hasPOS          lila:adverb ;
lila:isHypolemma     linkenLemma:677 ;
dcterms:isPartOf linkenLemma:LemmaBank;
ontolex:writtenRep   "slowly" .

linkenLemma:677    a lila:Lemma ;
rdfs:label           "slow" ;
lila:hasPOS          lila:adjective ;
dcterms:isPartOf linkenLemma:LemmaBank;
ontolex:writtenRep   "slow" .
```

[25] https://www.w3.org/TR/turtle/

5. Evaluation and Results

The validation of the results for the input 14,873 head words and the evaluation of the model's performance were carried out from two points of view. A syntactic evaluation was performed to check that all outputs were structured correctly, according to RDF syntax and the ontology that was provided in input. RDF structure and each ontological rule were always respected, including the appropriate usage of the necessary classes and properties, as well as the naming and indexing of individuals, structured correctly according to Turtle syntax. Hence, concerning the capability of the model to be compliant with an ontological schema for structuring multiple lexical data, the results were 100% well-formed.

A second semantic evaluation regarded the content of the outputs, which specifically concerned the accuracy of PoS assignment to each lemma. This was the most nuanced and variable task that was asked to the model, leading to potential variability and interpretation. In order to perform an accuracy evaluation of the PoS assigned by the model, we divided the outputs into three categories, and performance metrics for each of them are reported in Table 3:

1. Input words corresponding to one lemma and one PoS only in output (8,269).
2. Input words for which two lemmas were created in output, with one different PoS each (5,482).
3. Input words for which three or more lemmas were created in output, with one different PoS each (138).

The model was tasked to produce PoS only for adjectives, nouns and verbs, while we relied on the Kilgarriff list as a gold standard for function words, without asking them to the model.

We chose OEWN as a gold standard for evaluation, as it organizes lemmas into synsets and assigns multiple PoS to each lemma. We chose this method, instead of using NLP toolkits such as Stanza from CoreNLP[26], NLTK[27] or SpaCy,[28] because this is a type-based task, rather than token-based, and we needed an out-of-context approach, where any input word must be associated to any possible PoS it could have in discourse.

The choice of using OEWN for the identification of errors was useful as it provided a specific benchmark to compare the LLM outputs, but at the same time it led to some limitations. For instance, in the case of *importune*, the model gave two PoS in output (ADJ and VERB), while according to OEWN it should have only been VERB.[29] But looking at another lexical resource, namely the Oxford English Dictionary,[30] *importune* can be also an adjective. Here, we rely on OEWN, but this was done for practical and evaluation reasons only, and what we call "errors" by following OEWN may not be such according to other resources. We report this as a limitation of this work, which can be kept as such at this stage where only a limited portion of the English lexicon is included in the lemma bank, but will need to be addressed when expanding it to a larger coverage.

n of PoS	PoS	P	R	A
1	ADJ	92.9	77.3	99
	NOUN	98.7	98.2	99
	VERB	97.9	96.2	99
2	ADJ	56.8	93.5	85.9
	NOUN	92.2	100	92.2
	VERB	72.2	97.5	78.8
3	ADJ	69.6	100	100
	NOUN	92.0	100	100
	VERB	75.4	100	100

Table 3: Scores reported in percentages of correct PoS generation for adjectives, nouns and verbs in the three categories: when one lemma was generated in output with only one PoS, when two lemmas were generated in output with two different PoS, when three lemmas were generated in output with three different PoS. In the third category, recall and accuracy are equal to 1.0 as there are no cases where a PoS is not generated by the model.

As confirmed by the numbers in Table 3, words that can have only one PoS are not ambiguous and the tagging is quite straightforward, so the model shows high performance scores, even if out of context. In the second and third categories, the lower scores in precision are influenced by the fact that the model was overproductive, so it tended to assign more PoS even when only one was required. This can be linguistically motivated by the conversive derivational processes of verbalization (e.g., NOUN *a table*/VERB *to table something*), nominalization (e.g., VERB *to change*/NOUN *make a change*) and adjectivization (e.g., NOUN *an antioxidant*/ADJ *antioxidant properties*), which are very frequent in English and can influence the model towards over-generation when not necessary. This affects precision and recall differently: producing more than necessary makes precision lower while increasing recall, as there will be more misclassified entities, but less missed ones.

Apart from the one related to PoS, another er-

[26] https://stanfordnlp.github.io/stanza/
[27] https://www.nltk.org/
[28] https://spacy.io/usage/linguistic-features
[29] https://en-word.net/lemma/importune
[30] https://www.oed.com/

ror type that needed manual intervention concerns incompleteness. This does not involve incorrect syntax or vocabulary, but incomplete answers: for a given input word, the model started generating the triples for the related lemma or hypolemma, but interrupted the generation at some point. The stage at which the generation was interrupted was not constant across all cases, and it was more frequent for hypolemmas (17% of input words) than for lemmas (2.3%). Most of them (76%) were related to words that should have been assigned more than one PoS, and hence more than one lemma was to be generated for the given word, introducing more variability and lexical information to process for the model.

6. The LinkEn Lemma Bank

The Lemma Bank[31] of the LinkEn KB represents the final output of the neuro-symbolic workflow developed and followed in this research. Its design aims to integrate data-driven lexical generation from LLMs with symbolic knowledge representation grounded in the OntoLex-Lemon model and the LiLa ontology. Each word is uniquely identified and represented as a `lila:Lemma` (or `Hypolemma`) instance as the entry point of an RDF labeled graph consisting of a central lexical node and all its metalinguistic information encoded through well-defined properties.

At the time of completion, the lemma bank is composed as shown in Table 4.

Category	Count
Total RDF triples	127,833
Distinct lemmas	21,275
Distinct hypolemmas	3,463
Total number of nodes	24,765
Distinct written variations	25,368
Distinct rdfs classes	2
Distinct properties	5
Average triples per lemma/hypolemma node	5.17

Table 4: Statistics and numbers of the lemma bank of the LinkEn KB for English.

7. Future Work

As the Lemma Bank represents only the initial phase of a much broader effort that inherently needs to be carried on continuously towards the enrichment of the LinkEn KB, we propose the following directions for future development. The long-term objective is to extend the coverage to the entire English lexicon. This is planned to be achieved in a double fashion:

- Refining and reapplying the proposed methodology to larger and more diverse lemma sets, leveraging our hybrid LLM-KG pipeline in the task of KB expansion and enrichment.
- Following a resource-driven approach, where each newly identified lemma from external lexical or textual sources is continuously integrated into the existing lemma bank. Such an incremental and iterative updating process will ensure the dynamic growth and long-term sustainability of the resource, aligning it with the LLOD principles.

The LinkEn interoperable design, combined with Linked Data best practices, allows to create richer KGs going beyond isolated datasets, and enabling advanced queries and data mining operations at scale. In order to enhance LinkEn's coverage and robustness, our goal is to link an expanded set of lexical and textual resources for English, harmonizing diverse data sources and stimulating collaborative research.

[31] https://github.com/lorenzoaugello/LinkEn

8. Bibliographical References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *ScientificAmerican.com*.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction.

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *The Semantic Web: Research and Applications*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Chiarcos and Maria Sukhareva. 2015. Olia – ontologies of linguistic annotation. *Semantic Web*, 6:379–386.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 601–610, New York, NY, USA. Association for Computing Machinery.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR lemma bank: a new LLOD resource for Old Irish. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 37–43, Torino, Italia. ELRA and ICCL.

Jan-Christoph Kalo and Leandra Fichtel. 2022. Kamel: Knowledge analysis with multitoken entities in language models. In *4th Conference on Automated Knowledge Base Construction*.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio, and Cristina Bosco. 2024. The lemma bank of the LiITA knowledge base of interoperable resources for Italian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

John P. McCrae, Julia Bosque Gil, Jordi Gràcia, Paul Bitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: Development and applications.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges.

Marco Carlo Passarotti and Francesco Mambrini. 2022. Linking latin: Interoperable lexical resources in the lila project. In *Building new resources for historical linguistics*, pages 103–124. avia University Press.

Marco Carlo Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Maria Gabriella Litta Modignani Picozzi, 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(1):177–212.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs.

Rui Yang, Boming Yang, Aosong Feng, Sixun Ouyang, Moritz Blum, Tianwei She, Yuang Jiang, Freddy Lecue, Jinghui Lu, and Irene Li. 2025. Graphusion: A rag framework for knowledge graph construction with a global perspective.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4).

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

9. Language Resource References

BNC Consortium. 2007. *British National Corpus 1994*. Literary and Linguistic Data Service.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *Int. J. Lexicogr*, 10.

McCrae, John P. and Rademaker, Alexandre and Bond, Francis and Rudnicka, Ewa and Fellbaum, Christiane. 2019. *English WordNet 2019 – An Open-Source WordNet for English*. Global Wordnet Association.

Towards a Linguistic Linked Open Data Resource for Italian Cultural Heritage: The *Lessico dei Beni Culturali* Corpus

Riccardo Billero
Free University of Bozen
riccardo.billero@unibz.it

Abstract

We present an ongoing effort to bridge the *Lessico dei Beni Culturali* (LBC), a multilingual lexicographic project covering Italian cultural heritage terminology, with the Linguistic Linked Open Data (LLOD) ecosystem. The LBC corpus spans five centuries of art-historical writing, from fifteenth- and sixteenth-century treatises by Alberti, Leonardo, and Vasari to nineteenth-century works by Stendhal and Burckhardt and contemporary tourist guides to Florence, with source texts in several European languages alongside their translations. The resource has already undergone automatic linguistic annotation and term extraction, but lacks structured lexical representation in any standard LLOD formalism. We describe the current state of the resource, identify the main challenges for its publication as Linked Data — including the modelling of culturally-bound terms (*realia*), historical proper nouns, and multilingual source texts of different registers — and outline a roadmap towards its representation in OntoLex-Lemon (McCrae et al., 2017) and its alignment with existing LLOD resources such as the Getty Vocabularies (Getty Research Institute, 2024a) and Wikidata (Vrandecic and Krötzsch, 2014). By sharing this work with the LLOD community, we expect input on best practices for historical-artistic and cultural heritage lexicons that will raise interoperability between resources from different sources, generating new information and increasing the value of existing data.

Keywords: Linguistic Linked Open Data, cultural heritage, lexicography, multilingual corpus, digital humanities, corpus annotation

1. Introduction

The Linguistic Linked Open Data (LLOD) cloud has grown substantially over the past decade, yet certain specialised domains remain conspicuously underrepresented. Among these is the language of cultural heritage: the rich, historically layered vocabulary used to describe artworks, monuments, artists, and artistic practices. This gap is especially striking in the case of Italian, whose cultural heritage terminology forms the backbone of art history discourse worldwide yet lacks a structured, interoperable lexical representation.

The *Lessico dei Beni Culturali* (LBC; "Lexicon of Cultural Heritage") project directly addresses this gap from the perspective of multilingual lexicography and specialised translation (Garzaniti and Farina, 2013; Billero and Nicolás Martínez, 2017; Billero, 2020a; Billero et al., 2020; Ballestracci, 2023; Flinz et al., 2024). The project has assembled a corpus of parallel and comparable texts spanning five centuries of art-historical writing, in Italian and seven other languages (Chinese, English, French, German, Portuguese, Russian, Spanish).

What makes the LBC corpus particularly distinctive is its multilingual nature and the broad diachronic range and variety of genres of its source texts: it ranges from foundational Italian treatises — Alberti's *I dieci libri dell'architettura* (Alberti, 1550) and *Della pittura* (Alberti, 1436), Leonardo da Vinci's *Trattato della pittura* (da Vinci, 1651), and Vasari's *Le Vite de' più eccellenti pittori, scultori e architettori* (Vasari, 1568) — to works originally composed in other European languages, such as Stendhal's *Rome, Naples et Florence* (Stendhal, 1817) and Jacob Burckhardt's *Geschichte der Renaissance in Italien* (Burckhardt, 1867), through to contemporary tourist guides to Florence (Ciccarelli Roming et al., 2016).

2. Related Work

Within the LLOD ecosystem, significant efforts in the cultural heritage domain have focused on museum collections and archival metadata rather than on the language used to describe them. The Europeana Data Model provides a framework for aggregating cultural heritage objects across European institutions (Europeana Foundation, 2017), while CIDOC-CRM (Doerr, 2003) offers a rich ontology for cultural heritage events and objects; neither, however, is designed to represent the multilingual lexical variation that characterises art-historical discourse across centuries.

At the lexicographic level, OntoLex-Lemon (McCrae et al., 2017) has become the de facto standard for representing lexical resources as Linked Data, with applications ranging from general-purpose resources such as English WordNet (McCrae et al., 2014, 2020) to domain-specific terminological databases. Its application to art-historical and cultural heritage lexica specifically, however, remains relatively underexplored.

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 22–28
12 May 2026.

The LBC project addresses precisely this gap: its diachronic, multilingual corpus and existing lexicographic infrastructure make it a strong candidate for OntoLex-Lemon encoding, though substantial work remains before a full LLOD publication can be achieved.

3. The LBC Resource

3.1. Corpus composition

The LBC corpus spans roughly five centuries of art-historical writing, forming a continuum from Renaissance treatises to contemporary tourist publications, with source texts in different European languages.

Renaissance and early modern treatises. The earliest layer of the corpus consists of Italian treatises that established the vocabulary of Western art history: Leon Battista Alberti's *I dieci libri dell'architettura* (Alberti, 1550) and *Della pittura* (Alberti, 1436), Leonardo da Vinci's *Trattato della pittura* (da Vinci, 1651) (compiled posthumously), and Benvenuto Cellini's *Vita* (Cellini, 1728) (written c. 1558–1563). These texts are available in Italian with translations in the project's target languages.

Vasari's *Vite*. Vasari's *Le Vite de' più eccellenti pittori, scultori e architettori* (Vasari, 1568) is the foundational text of Western art historiography. Written in sixteenth-century Tuscan Italian, it provides technical descriptions and related terminology of paintings, sculptures, and architectural works alongside biographical accounts of Renaissance artists. The LBC project has assembled the Italian original together with translations in other project languages, forming a set of *parallel corpora* of exceptional depth.

Nineteenth-century critical writing. Texts such as Stendhal's *Rome, Naples et Florence* (Stendhal, 1817) and Burckhardt's *Geschichte der Renaissance in Italien* (Burckhardt, 1867) introduce a non-Italian scholarly perspective on Florentine cultural heritage and expand the source language range of the corpus beyond Italian.

Contemporary tourist guides. Tourist guides to Florence (Ciccarelli Roming et al., 2016) provide an accessible, descriptive register sharing the same terminological core as the historical sources while operating under very different communicative constraints. Together with the earlier texts, they form a *comparable corpus* that enables diachronic and register-based analysis.

3.2. Corpus access and current processing

Corpus query interface. Of the full LBC corpus, six sub-corpora have so far been published as individual open-access volumes by Firenze University Press (Lanini, 2024; Carpi and Pano Alamán, 2024; Ballestracci et al., 2024; Natali, 2024; Farina, 2024; Rossi and Zhukova, 2024), all searchable via a local installation of NoSketchEngine (Kilgarriff et al., 2014) hosted at `http://corpora.lessicobeniculturali.net/` (Billero, 2020b).

Users can submit simple keyword queries or formulate complex searches using the Corpus Query Language (CQL), which allows fine-grained retrieval based on word forms, lemmas, and part-of-speech tags. CQL queries can be combined with metadata filters to restrict searches to specific source texts, time periods, or language pairs, making it possible to trace the diachronic evolution of terms across the corpus.

Expert concordance collection. A curated collection of KWIC (Key Word in Context) concordances, selected and annotated by domain experts, is accessible via the lexicon interface at `http://lexicon.lessicobeniculturali.net/`. These concordances are linked to specific lemmas and are currently available for four of the seven target languages: French, German, Russian, and Spanish, providing high-quality usage examples of domain-specific terms across languages and registers.

Linguistic annotation and term extraction. The corpus texts have already undergone automatic linguistic annotation, including tokenisation, part-of-speech tagging, and lemmatisation, carried out with TreeTagger (Schmid, 1994, 1995) and related tools for the languages supported. Domain-specific terms have subsequently been extracted and are linked to the lemma entries in the lexicon interface. These preprocessing steps represent a significant investment and provide a solid foundation for subsequent LLOD modelling, though the accuracy of automatic annotation for historical and specialised texts poses known challenges that will require targeted post-correction before LLOD encoding.

Prototype dictionary. A prototype dictionary, organised as bilingual entries from Italian to each target language, is under development and is intended for open-access publication by Firenze University Press (Farina et al., 2024).

3.3. Lexical content and scope

The LBC targets three main categories of lexical items, each posing distinct challenges for LLOD representation:

- **Common nouns of art**: technical terms describing techniques, materials, styles, and genres (*affresco*, *chiaroscuro*, *pietra serena*, *predella*). Many of these terms originate in Italian and have been borrowed or calqued into other European languages, making Italian the primary source for terminological standardisation in art history.
- **Proper nouns specific to Florence and its history**: artists, patrons, political figures, place names, street names, monuments, and titles of artworks.
- **Florentine *realia***: culturally bound items including foods, festivals, and practices that are specific to the Florentine cultural context and resist direct translation.

This tripartite structure, together with the diachronic variation of the language across five centuries of texts, maps naturally onto the distinctions drawn in LLOD modelling between lexical entries, named entities, and culture-specific concepts. Beyond these lexical categories, two further challenges arise from the corpus itself: the multilingual origin of its source texts (translation directionality) and its five-century diachronic span (register variation), both discussed in Section 4.

4. Challenges for LLOD Modelling

4.1. Historical and multilingual proper nouns

The LBC corpus contains a large number of historically attested proper nouns recorded under different orthographic forms and in different scripts across the seven languages. *Michelangelo Buonarroti* appears as *Michel-Ange* in French, *Miguel Ángel* in Spanish, and under Cyrillic transliterations in Russian.

Wikidata (Vrandecic and Krötzsch, 2014) already provides multilingual labels for most major artists and monuments and is available as LLOD. Linking LBC entries to Wikidata Q-items via `owl:sameAs` or `skos:exactMatch` would provide a ready-made multilingual backbone. However, less prominent figures — minor painters, local patrons, Florentine guild officials — may be absent from Wikidata or poorly described, requiring new entity creation or the use of more specialised resources such as the Getty Union List of Artist Names (ULAN) (Getty Research Institute, 2024c). A similar challenge arises for historical place names: Florentine toponyms appear under different forms across languages and centuries — from *Firenze* to *Florence*, *Florenz*, and *Fiorenza* — and their alignment to the Getty Thesaurus of Geographic Names (TGN) (Getty Research Institute, 2024b) will require careful historical disambiguation.

4.2. Modelling *realia*

Realia — culture-specific items that lack direct equivalents in other languages — are one of the central concerns of the LBC project (Farina, 2014) and one of the most theoretically interesting challenges for LLOD modelling. A term such as *calcio storico* (a historical Florentine ball game) or *canto* (a civic district of Florence) does not simply translate; it *migrates* into target languages with varying degrees of adaptation, borrowing, paraphrase, or loss.

Standard LLOD formalisms such as OntoLex-Lemon (McCrae et al., 2017) are well suited to representing translation equivalents when equivalence holds, but are less expressive for capturing *degrees of cultural approximation*. One possible approach is to model *realia* as `ontolex:LexicalConcept` instances with asymmetric `skos:closeMatch` or `skos:relatedMatch` relations to their approximate translations, supplementing these with `skos:note` annotations that capture the nature of the mismatch.

The same framework handles specialised art vocabulary more straightforwardly: *affresco* would be encoded as an `ontolex:LexicalEntry` linked to a `ontolex:LexicalConcept`, with translation equivalents (*fresco* in Spanish, *fresque* fresque in French, *Fresko*) in German) in their respective lexicons all pointing to the same concept node, itself aligned via `skos:exactMatch` to Getty AAT 300177433 (*frescoes*).

The following sketches illustrate the proposed encoding for *affresco*:

```
:affresco_it a ontolex:LexicalEntry ;
  ontolex:language "it" ;
  ontolex:sense :affresco_sense .
:affresco_sense
  ontolex:reference :fresco_concept .
:fresco_concept a ontolex:LexicalConcept ;
  skos:exactMatch aat:300177433 .
:fresco_en a ontolex:LexicalEntry ;
  ontolex:language "en" ;
  ontolex:sense :fresco_en_sense .
:fresco_en_sense
  ontolex:reference :fresco_concept .
```

4.3. Multilingual source texts and translation directionality

Unlike most parallel corpora used in LLOD projects, the LBC corpus includes texts originally composed in several languages, not only Italian. This complicates standard assumptions about translation directionality: rather than a single source language with multiple target translations, the corpus presents a web of source-target relationships that varies by text. For instance, Burckhardt's German text on the Italian Renaissance has been translated into Italian, making Italian both a source and a target language depending on the text pair considered. The LIME metadata module (Fiorelli et al., 2015) for OntoLex-Lemon (McCrae et al., 2017) would need to be applied carefully to represent these relationships without forcing Italian into the role of sole pivot language. Concretely, each source text would be associated with a distinct `lime:Lexicon` instance carrying its own `lime:language` and `lime:linguisticCatalog` properties. Translation relations between lexicons would be expressed at the `ontolex:LexicalSense` level using `vartrans:translatableAs` from the OntoLex Vartrans module, allowing Burckhardt's German lexicon to relate directly to its Italian and other translations without Italian acting as intermediary.

4.4. Diachronic register variation

The juxtaposition of fifteenth-century treatises with twenty-first-century tourist guides raises a fundamental issue for lexical representation: the same term may carry different denotations, connotations, or pragmatic values across registers and centuries. *Maniera*, for instance, is a key Vasarian aesthetic concept but functions as a generic stylistic descriptor in contemporary tourist texts.

OntoLex-Lemon's `ontolex:usage` and `lexinfo:register` (Cimiano et al., 2011) properties can capture register distinctions, while separate `lime:Lexicon` instances could encode the different sub-corpora as distinct lexical perspectives on the same domain ontology.

4.5. Alignment with existing LLOD resources

The cultural heritage domain is one of the few areas where substantial, professionally curated LLOD-compatible resources already exist:

- **Getty Art & Architecture Thesaurus (AAT)** (Getty Research Institute, 2024a): hierarchical vocabulary of artistic concepts and techniques, available as Linked Data at `https://vocab.getty.edu/aat/`.
- **Getty Union List of Artist Names (ULAN)** (Getty Research Institute, 2024c): authority file for artists and architects with multilingual labels, at `https://vocab.getty.edu/ulan/`.
- **Getty Thesaurus of Geographic Names (TGN)** (Getty Research Institute, 2024b): geographic thesaurus covering historical and contemporary place names, at `https://vocab.getty.edu/tgn/`.
- **Wikidata** (Vrandecic and Krötzsch, 2014): encyclopaedic knowledge graph with strong multilingual coverage of artists, artworks, and monuments, queryable via SPARQL at `https://query.wikidata.org/`.

Aligning LBC entries with these resources would immediately embed the lexicon into a rich semantic context. For example, the LBC entry for *pietra serena* — the grey sandstone characteristic of Florentine Renaissance architecture — could be linked to the corresponding Getty AAT concept, while *Brunelleschi* could be linked via `skos:exactMatch` to Getty ULAN 500018169 and Wikidata Q174330, simultaneously enriching all three knowledge graphs with the Italian-language perspective of the LBC. A key challenge is that the Getty vocabularies are English-centric and do not always provide the nuanced Italian-language perspective and multilingual dimension that the LBC is designed to offer, and many minor historical figures in the corpus are absent from all existing authority files.

5. Towards an LLOD Roadmap for LBC

We propose the following incremental steps towards the publication of the LBC as an LLOD resource, starting from the processing already completed.

Step 1 — Quality assessment and post-correction of existing annotation. The corpus has already been tokenised, POS-tagged, and lemmatised using TreeTagger (Schmid, 1994, 1995) and related tools, and domain terms have been extracted and linked to lemma entries in the lexicon interface. TreeTagger offers robust models for several of the languages covered by the LBC, but coverage is uneven across all languages of the project, and for some target languages dedicated tools are required. A further source of degradation is the diachronic range of the corpus: TreeTagger models are trained on contemporary corpora, and their performance degrades on historical texts where orthographic and morphological variation is

substantial, as well as on domain-specific nominal compounds. Quality assessment will therefore require targeted manual post-correction by specialists with both technical and linguistic competence in the individual languages. Error typologies identified during this phase will inform post-correction heuristics applicable to similar multilingual historical corpora.

For the historical Italian sub-corpus in particular, the main difficulties stem from pre-standardisation graphemic and orthographic variability; archaic or obsolete morphological forms not always recognised by models trained on contemporary Italian; specialist humanistic vocabulary not found in standard dictionaries; and interference arising from modern editorial conventions. The most frequent errors relate to POS disambiguation in the presence of archaic spellings, incorrect lemmatisation of historical verb forms, and incorrect recognition of proper names and Latinisms. Phase 1 therefore does not involve fully automated annotation, but rather a hybrid workflow: post-correction will be supported by specific guidelines for historical Italian, developed in collaboration with Italian linguists, with the aim of achieving a level of reliability consistent with the objectives of semantic querying and LLOD modelling.

Step 2 — OntoLex-Lemon encoding of extracted terms. Encode the existing extracted terms in OntoLex-Lemon (McCrae et al., 2017), with separate `lime:Lexicon` instances for each language and for each diachronic sub-corpus. The expert-curated concordances already available in the lexicon interface for French, German, Russian, and Spanish can serve as high-quality usage examples within the `ontolex:LexicalSense` entries.

Step 3 — Entity linking and LLOD alignment. Link named entities (artists, places, artworks) to Getty ULAN (Getty Research Institute, 2024c), Getty TGN (Getty Research Institute, 2024b), and Wikidata (Vrandecic and Krötzsch, 2014) via `owl:sameAs` and `skos:exactMatch`. Link conceptual terms (art techniques, styles, genres) to the Getty AAT (Getty Research Institute, 2024a).

Step 4 — Modelling of the specialised artistic lexicon, *realia*, and translation asymmetries. Develop a lightweight extension to the OntoLex-Lemon model to represent degrees of translational equivalence for both specialised artistic vocabulary and culture-specific *realia*, building on the SKOS mapping vocabulary (Miles and Bechhofer, 2009) and drawing on literature on the lexicographic treatment of culture-specific items (Aixelà, 1996).

Step 5 — Publication and interlinking. Publish the resulting lexicon as RDF under an open licence, with a SPARQL endpoint, and submit it for inclusion in the LLOD cloud diagram (Chiarcos et al., 2012).

Steps 1–2 are the immediate next objectives; Steps 3–5 represent medium-term goals for which community input is actively sought.

6. Conclusion

We have presented the LBC project as a candidate for LLOD publication, described its corpus and current state of processing, and outlined the principal challenges its conversion to Linked Data poses. The cultural heritage domain is underrepresented in the LLOD cloud despite the existence of high-quality complementary resources such as the Getty Vocabularies; the LBC has the potential to fill a significant gap, particularly with respect to Italian-centric multilingual coverage and the lexicography of historical texts.

We therefore invite members of the LDL community to share experience with OntoLex-Lemon modelling for specialised or historical lexica, strategies for representing culture-specific concepts and translational asymmetries in RDF, and alignment methodologies for parallel corpora and knowledge graphs.

We hope this paper will serve as a starting point for a productive discussion at LDL-2026, contribute to the growing intersection of digital humanities and the Semantic Web, and offer a replicable model for other historical multilingual cultural heritage lexica currently outside the Linked Data cloud.

7. Limitations

The work presented here is at an early stage. No LLOD implementation has yet been produced, and the modelling proposals in Section 5 remain programmatic. The accuracy of TreeTagger on historical Italian is known to be lower than on contemporary language: lemmatisation errors are particularly frequent for sixteenth-century verb forms and domain-specific nominal compounds, and will require targeted manual post-correction before high-quality LLOD encoding can be achieved. The proposed treatment of *realia* via SKOS mapping properties is theoretically motivated but has not yet been validated against the full range of cases in the corpus. Finally, the alignment with Getty and Wikidata has been illustrated with selected examples only; systematic alignment will require semi-automatic methods and expert validation.

8. Acknowledgements

The *Lessico dei Beni Culturali* project is hosted at the Department of Education, Languages, Intercultures, Literatures and Psychology (FORLILPSI) of the University of Florence and is published by Firenze University Press.

9. Bibliographical References

Javier Franco Aixelà. 1996. Culture-specific items in translation. In Román Álvarez and M. Carmen África Vidal, editors, *Translation, Power, Subversion*, pages 52–78. Multilingual Matters, Clevedon.

Leon Battista Alberti. 1436. *Della pittura*. Three books. Italian version of *De pictura*.

Leon Battista Alberti. 1550. *I dieci libri dell'architettura*. Lorenzo Torrentino, Florence. Italian translation of *De re aedificatoria* (1452).

Sabrina Ballestracci. 2023. Da un segno a tanti segni. L'emergere della polisemia del termine "disegno" nelle traduzioni tedesche delle "Vite" di Vasari. In Valeria Zotti and Monica Turci, editors, *Nuove strategie per la traduzione del lessico artistico: da Giorgio Vasari a un corpus plurilingue dei beni culturali*, pages 21–38. Firenze University Press, Firenze.

Riccardo Billero. 2020a. Cultural heritage lexicon: A case study. In Ana Pano Alamán and Valeria Zotti, editors, *The Language of Art and Cultural Heritage: A Plurilingual and Digital Perspective*, pages 86–103. Cambridge Scholars Publishing.

Riccardo Billero. 2020b. Implementazione di software per la gestione dei corpora LBC. In Riccardo Billero, Annick Farina, and María Carlota Nicolás Martínez, editors, *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*, pages 19–32. Firenze University Press, Firenze.

Riccardo Billero, Annick Farina, and María Carlota Nicolás Martínez. 2020. *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Firenze University Press, Firenze.

Riccardo Billero and María Carlota Nicolás Martínez. 2017. Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. *CHIMERA: Romance Corpora and Linguistic Studies*, 4(2).

Jacob Burckhardt. 1867. *Geschichte der Renaissance in Italien*. Ebner & Seubert, Stuttgart.

Benvenuto Cellini. 1728. *Vita*. Naples. Written c. 1558–1563; first printed edition.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. 2012. Linked data in linguistics: Representing and connecting language data and language metadata. In *Proceedings of the 1st Workshop on Linked Data in Linguistics (LDL-2012)*, Frankfurt am Main, Germany. CEUR Workshop Proceedings.

C. Ciccarelli Roming, T. Jepson, and T. Fisher. 2016. *Florenz. Perfekte Tage in der Toskana – Metropole*. Taschen, München.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Leonardo da Vinci. 1651. *Trattato della pittura*. Compiled posthumously; first printed edition, Paris: Giacomo Langlois.

Martin Doerr. 2003. The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92.

Europeana Foundation. 2017. Europeana data model documentation. https://pro.europeana.eu/page/edm-documentation. Accessed: February 2026.

Annick Farina. 2014. Descrivere e tradurre il patrimonio gastronomico italiano: le proposte del lessico plurilingue dei Beni Culturali. In Francesca Chessa, Cosimo De Giovanni, and Maria Teresa Zanola, editors, *La terminologia dell'agroalimentare*, pages 55–66. Franco Angeli, Milano.

Annick Farina, Riccardo Billero, and María Carlota Nicolás Martínez. 2024. Presentation of the LBC database. In I. Natali, editor, *The LBC English Corpus*. Firenze University Press, Firenze. E-ISBN: 979-12-215-0307-4. https://corpora.lessicobeniculturali.net/en/.

Manuel Fiorelli, Armando Stellato, John P. McCrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. LIME: The metadata module for OntoLex-Lemon. In *Proceedings of the 12th European Semantic Web Conference (ESWC 2015)*, pages 321–336, Portorož, Slovenia. Springer.

Carolina Flinz, Valeria Zotti, D. Henkel, and Sabrina Ballestracci. 2024. A multilingual parallel corpus for the lexical information system LBC:

Recent progress and future perspectives. In *Lexicography and Semantics*. Cambridge Scholars Publishing.

Marcello Garzaniti and Annick Farina. 2013. Un portale per la comunicazione e la divulgazione del patrimonio culturale: progettare un lessico multilingue dei beni culturali on-line. In A. Filipovic and W. Troiano, editors, *Strategie e Programmazione della Conservazione e Trasmissibilità del Patrimonio Culturale*, pages 500–509. Edizioni scientifiche Fidei Signa, Roma.

Getty Research Institute. 2024a. Art & architecture thesaurus online. https://vocab.getty.edu/aat/. Accessed: February 2026.

Getty Research Institute. 2024b. Thesaurus of geographic names online. https://vocab.getty.edu/tgn/. Accessed: February 2026.

Getty Research Institute. 2024c. Union list of artist names online. https://vocab.getty.edu/ulan/. Accessed: February 2026.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. In *Lexicography*, volume 1, pages 7–36.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon model: Development and applications. In *Proceedings of eLex 2017*, pages 587–597, Leiden, The Netherlands.

John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking WordNet using Lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

John P. McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. European Language Resources Association (ELRA).

Alistair Miles and Sean Bechhofer. 2009. SKOS simple knowledge organization system reference. W3C Recommendation, https://www.w3.org/TR/skos-reference/.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Stendhal. 1817. *Rome, Naples et Florence*. Delaunay, Paris.

Giorgio Vasari. 1568. *Le Vite de' più eccellenti pittori, scultori e architettori*. Giunti, Florence. Second enlarged edition; first edition 1550.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57(10):78–85.

10. Language Resource References

Sabrina Ballestracci, Claudia Buffagni, and Carolina Flinz. 2024. Das deutsche LBC-korpus. E-ISBN: 979-12-215-0311-1. https://corpora.lessicobeniculturali.net/de/.

Elena Carpi and Ana Pano Alamán. 2024. Corpus LBC español. E-ISBN: 978-88-5518-035-1. https://corpora.lessicobeniculturali.net/es/.

Annick Farina. 2024. Corpus LBC français. E-ISBN: 979-12-215-0309-8. https://corpora.lessicobeniculturali.net/fr/.

Ludovica Lanini. 2024. Corpus LBC italiano. E-ISBN: 979-12-215-0305-0. https://corpora.lessicobeniculturali.net/it/.

Ilaria Natali. 2024. The LBC English corpus. E-ISBN: 979-12-215-0307-4. https://corpora.lessicobeniculturali.net/en/.

Valentina Rossi and Natalia Zhukova. 2024. Russkij korpus «Leksika kul'turnogo nasledià». E-ISBN: 979-12-215-0313-5. https://corpora.lessicobeniculturali.net/ru/.

Consolidating Syntactically Annotated Corpora with LLOD Technology. An Experiment in the Old Saxon Heliand

Christian Chiarcos[a], Janine Siewert[a,b]
[a]Applied Computational Linguistics (ACoLi), University of Augsburg, Germany
[b]Department of Digital Humanities, University of Helsinki, Finland
christian.chiarcos@uni-a.de, janine.siewert@helsinki.fi

Abstract

The humanities are methodologically and technologically diverse, and independent projects often produce complementary but technically incompatible digital editions from the same sources. We show how Linguistic Linked Open Data (LLOD) technology, and in particular, Fintan and CoNLL-RDF can support the post-hoc consolidation of such resources by using SPARQL updates for selection, aggregation and consolidation operations over heterogeneous annotations. This is illustrated for the Old Saxon (Old Low German) *Heliand*, a 9th-century gospel harmony annotated for different syntactic aspects in three independent projects and across different editions and manuscripts. We describe the derivation of a UD-compliant corpus through transformation into corpus-specific CoNLL formats, cross-version alignment, and annotation integration. A central challenge is the consolidation of incomplete and heterogeneous annotations.

Keywords: annotation consolidation, syntax, Fintan, RDF/SPARQL, Old Saxon (Old Low German)

1. Background and Motivation

The increasing availability of digitally annotated linguistic resources has profoundly reshaped research in historical linguistics and philology. At the same time, the humanities remain methodologically and technologically heterogeneous: different theoretical assumptions, annotation traditions and research questions regularly lead to parallel, independent annotation efforts over (often different editions or witnesses of) the same primary texts. While such diversity is scientifically productive, it results in resources that are complementary in scope but technically incompatible, leaving their combined analytical potential largely untapped.

This situation is particularly evident in historical syntax, where constituency-based treebanks inspired by generative grammar coexist with multi-tier annotations in the tradition of interlinear glossed text (Bow et al., 2003, IGT). For the Old Saxon *Heliand*, these approaches differ substantially in design and coverage. Treebanks model hierarchical structure exhaustively, whereas tier-based annotations allow flexible layering, for instance for the study of information-structural phenomena at the syntax–pragmatics interface. Re-annotation is therefore common practice – not due to deficiencies in existing resources, but because reuse across frameworks is technically difficult. As a consequence, high-quality annotations remain siloed.

The *Heliand*, a 9th-century Gospel harmony and the most extensive Old Saxon text, occupies a central position in early West Germanic philology. Its linguistic profile is crucial for reconstructing early syntactic change (Petrova and Solf, 2009; Lühr, 2025). Reflecting its importance, it has been annotated repeatedly in independent projects: the HeliPaD treebank (Walkden, 2016), following Penn-style constituency annotation; the Heliand DDD corpus (Referenzkorpus "Altdeutsch") with tier-based morphosyntactic and clausal annotation; and the Heliand B4 corpus (Linde, 2009), developed for diachronic information-structural research. These corpora differ in granularity. HeliPaD provides full phrase-structure parses; DDD and B4 offer partial, non-recursive annotations (POS, clause boundaries, and in B4 nominal and prepositional phrases). Consolidation therefore requires (i) conversion to CoNLL(-U), (ii) enrichment of partial annotations, (iii) alignment across divergent textual bases, and (iv) merging of alternative syntactic analyses.

We focus on conversion and merging, implemented as graph rewriting operations using SPARQL over on-the-fly transformations between CoNLL sentence blocks and RDF graphs. While earlier work proposed RDF for storage and querying of multi-layer corpora (Burchardt et al., 2008; Mazziotta, 2010), later research demonstrated its suitability for transformation tasks (Fäth et al., 2020; Chiarcos et al., 2021), including integration of heterogeneous syntactic annotations (Chiarcos et al., 2022). In this paper, we extend the application of RDF and Linked Open Data (LLOD) technologies to the combination of conflicting syntax annotations into a coherent CoNLL-U representation. Although RDF and LLOD have been applied previously to harmonizing annotations, the post-hoc consolidation of conflicting legacy annotations for pre-modern corpora remains underexplored and is addressed in this paper for the specific case of the Heliand.

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 29–39
12 May 2026.

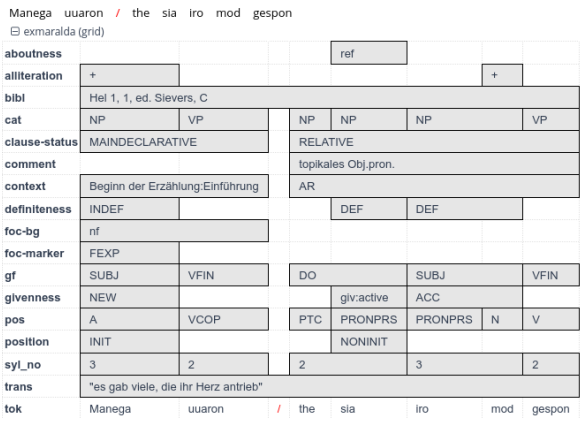

Figure 1: Heliand B4 annotation, first verse, visualized with ANNIS3

2. From Sources to Syntax Annotation. The Story so far

Old Saxon (OS) is attested primarily in two major 9th-century alliterative poems, the *Heliand* and the *Genesis*. Apart from a small number of liturgical texts and tax lists, these constitute the core of the Old Saxon textual record. With roughly 6,000 long verses, the *Heliand* is by far the more extensive text and has therefore become the principal basis for corpus-based research. It is preserved in two major manuscripts, C (London, British Library, Cotton Caligula A. VII) and M (Munich, Bayerische Staatsbibliothek, Cgm 25), as well as in four fragments.

The first syntactically annotated corpus of the *Heliand* was the *Heliand B4 corpus* (Petrova, 2006; Linde, 2009), developed within the Collaborative Research Center 632 “Information Structure”. Its annotation methodology, described in detail by Petrova et al. (2009), relies on the tier-based tool EXMARaLDA (Schmidt and Wörner, 2014). Tokens form the primary timeline; additional layers provide morphosyntactic categories, syntactic functions, information-structural features, and metadata. These layers are attached to spans of tokens in a stand-off architecture. This design allows overlapping and independent hierarchies without enforcing a single dominant structure, and thus accommodates conflicting annotations and discourse-level relations. The data are converted to the PAULA interchange format and integrated into ANNIS for querying and visualization (Petrova et al., 2009). The same multi-tier strategy has been applied to Old Saxon, resulting in a richly annotated but comparatively small corpus: Heliand B4 (Fig. 1) comprises approximately 3,500 tokens, covering less than 10% of the text. Despite its limited size, it remains a valuable complement of the later DDD corpus in providing rich manual annotation, in particular for nominal and prepositional phrases and their grammatical functions.

```
( (IP-MAT
    (CODE <P_7>)
    ...
    (NP-SBJ *exp*)
    (NP-PRD (Q^N^PL Manega-manag)
            (CP-REL *ICH*-1))
    (BEDI^3^PL uuaron-wesan)
    (CODE <C>)
    (CP-REL-1
      (WNP-SBJ-2 0)
      (C the-the)
      (IP-SUB
        (NP-SBJ-RSP-2
          (PRO^A^3^PL sia-he))
        (NP-OB1 (PRO$^N^3^SG iro-his)
                (N^N^SG mod-mod))
        (GE+VBDI^3^SG gespon-spanan)
        (, ,-,)
        ... ))))
```

Figure 2: HeliPaD annotation, first verse, PTB bracketing format

The *Old German Reference Corpus (Referenzkorpus Altdeutsch, DDD)* (Linde and Mittmann, 2013) project aimed to digitize and annotate all extant Old High German and Old Low German (Old Saxon) texts, using a tier-based approach similar to the Heliand B4 corpus, but different tools. For the *Heliand*, DDD is based on the synoptic 55,080-token edition of Behaghel and Taeger (1984), digitized via the TITUS project (Gippert, 2011), enriched with morphological information and lemmatization from Sehrt (1925), and manually revised. Annotation was carried out in ELAN (Brugman and Russel, 2004), another tier-based tool. DDD provides detailed morphosyntactic annotation and shallow syntactic structure in the form of clause chunks and linking relations, but it does not encode full phrase structure trees. Syntactic annotation remains non-recursive and limited to higher-level segmentation. Compared to Heliand B4, DDD covers the complete text but at a shallower syntactic level.

In contrast to both tier-based corpora, the *HeliPaD* (Heliand Parsed Database) (Walkden, 2016) introduces full syntactic annotation for Old Saxon. It follows the conventions of the Penn Historical Corpora (Pintzuk and Taylor, 1997; Taylor, 2003; Booth et al., 2020; Kulick et al., 2022) and encodes explicit constituency structures and grammatical functions (Fig. 2). HeliPaD is based on Sievers (1878), and unlike the synoptic edition underlying DDD, this represents a single manuscript (Ms. C) and thus preserves manuscript-specific orthography and variation. The corpus comprises approximately 46,000 tokens and is smaller than DDD but syntactically more explicit.

Together, these three corpora present a heterogeneous landscape. Each contributes unique in-

formation: HeliPaD offers full constituency structure; DDD provides comprehensive coverage with consistent morphosyntax and clause segmentation; B4 supplies fine-grained manual phrase-level and information-structural annotation. At the same time, they differ in textual basis (single manuscript vs. synoptic edition), tokenization and orthography, annotation depth, tools (EXMARaLDA, ELAN, Penn bracketing), and formats. While conversion of Penn-style bracketing to dependency formats has been addressed in previous work (Johansson and Nugues, 2007; Arnardóttir et al., 2020), tier-based corpora pose a different problem. Their shallow, non-recursive structure cannot simply be converted; it must be enriched with additional syntactic information. In our approach, HeliPaD provides this enrichment both through its direct conversion to UD and through a parser trained on its UD representation and applied to DDD and B4 data.

Two challenges arise: (1) alignment and projection across divergent textual versions of the same work, and (2) consolidation of multiple and incomplete layers of syntax annotation into a consistent representation. The following sections describe how these can be addressed in a unified workflow.

3. Technological foundations

For consolidating the diverse Heliand annotations, we use four primary technical core components: (1) the Flexible Integrated Annotation eNgineering platform (Fäth et al., 2020, Fintan) allows to use SPARQL for transformation tasks, (2) the CoNLL-RDF customization of Fintan (Chiarcos and Fäth, 2017; Chiarcos et al., 2021) for reading and writing CoNLL data, (3) the UDPipe parser (Straka and Straková, 2017) for training CoNLL-U parsers and analysing unseen data, and (4) CoNLL-Merge (Chiarcos and Schenk, 2018) for merging CoNLL files with multiple layers of annotation from different sources. In addition, custom converters transform native corpus formats (PTB bracketing, ELAN, EXMARaLDA) into CoNLL-style TSV representations.

Fintan is a flexible framework for converting, enriching and transforming heterogeneous linguistic annotations by mapping them to RDF graphs and applying SPARQL queries and updates. It is modular and separates loading, transformation and serialization: Loaders convert native formats into RDF graphs that mirror their original structure; transformation modules use SPARQL to perform graph rewriting operations; writers serialize results back into conventional formats, including CoNLL. Because Fintan operates on RDF graphs that represent single sentences, not the complete data set, it can parallelize SPARQL updates, enabling scalable graph rewriting. In doing that, Fintan is agnostic to linguistic data models and vocabularies. This flexibility allows us to harmonize structurally divergent resources within a single graph-based workflow.

For corpus processing, we operate with the CoNLL-RDF vocabulary (Chiarcos et al., 2021), summarized in Fig. 3. CoNLL(-TSV) formats represent sentences as blocks of tab-separated rows, one token per line, optionally preceded by comment metadata. Columns encode annotations such as form, lemma, POS, morphological features, dependency relations or task-specific labels. CoNLL-RDF maps each token to a `nif:Word` node, linked via `nif:nextWord`, with column values represented as literal properties in the `conll:` namespace. Some columns receive dedicated treatment: `HEAD`, for example, is resolved into explicit `conll:HEAD` relations that link tokens (`nif:Word`s) with each other (for syntactic dependencies) or the sentence (for roots). The CoNLL-RDF tree extensions (Chiarcos and Glaser, 2020) extend the basic CoNLL-RDF model with the POWLA vocabulary to represent hierarchical structures such as phrase-structure trees. This extension allows Penn Treebank-style parses to be encoded as additional nodes linked via `powla:hasParent` and `powla:next`. Figure 4 shows a fragment with a token connected to a phrase node, which participates in a larger tree. Any kind of TSV-compatible annotation can be represented in a unified RDF graph, which can be easily traversed and manipulated with SPARQL, even if the data includes multiple layers of styles of syntax annotation, be it phrase structure trees or dependency syntax.

CoNLL-Merge complements this graph-based transformation layer by aligning different witnesses or editions encoded in CoNLL-style formats. It supports token-based alignment, merging and splitting operations, and concatenation of aligned annotations into additional columns. This is crucial for integrating independent annotations of the same text that differ in tokenization or textual basis. Once merged into a common CoNLL representation, annotations can be consolidated through SPARQL-based graph rewriting in CoNLL-RDF.

Overall, the combination of format conversion, alignment (CoNLL-Merge), graph-based normalization and rewriting (Fintan/CoNLL-RDF), and parser-based enrichment (UDPipe) provides a modular and extensible infrastructure for consolidating heterogeneous annotations into a unified CoNLL-U representation. We illustrate this for the Heliand corpora where annotation depth varies considerably: HeliPaD provides full constituency parses; DDD offers clause-level segmentation and morphosyntax; B4 contains non-recursive nominal and prepositional chunks. To compensate for missing syntactic structure, we train UDPipe (Straka and Straková, 2017) on a CoNLL-U conversion of HeliPaD and apply the resulting parser to DDD and

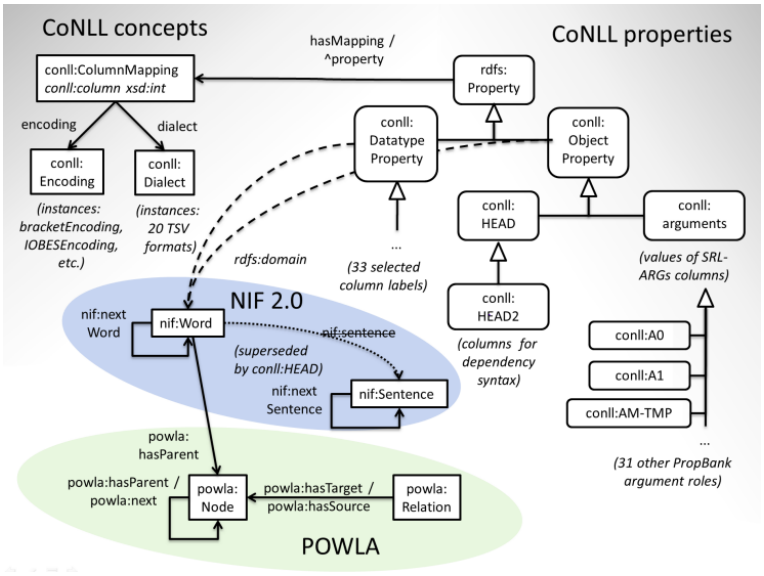

Figure 3: CoNLL-RDF Data Model (Chiarcos et al., 2021)

```
:s1_0    a                nif:Sentence .
:s1_7    a                nif:Word ;
         conll:HEAD       :s1_0 ;
         conll:WORD       "Manega" .
         conll:POS        "Q^N^PL" ;
         conll:LEMMA      "manag" ;
         nif:nextWord     :s1_8 ;
         powla:hasParent  :bPARSE_4 .
:bPARSE_4  a              conll:PARSE ;
         conll:CAT        "NP" ;
         conll:ROLE       "PRD" ;
         powla:hasParent  :bPARSE_2 ;
         powla:next       :bPARSE_5 .
```

Figure 4: CoNLL-RDF fragment for the first word of HeliPaD, with tree extensions

B4 data. The automatically generated dependency structures serve as a baseline that complements projected annotations and fills structural gaps.

4. From heterogeneous sources to a CoNLL-U

Figure 5 illustrates the different conversion, training and merging steps: We convert HeliPaD into CoNLL-U and train UDpipe over it. We convert DDD and B4 to CoNLL representations and align these with both HeliPaD-UD and automatically parsed dependencies to derive an initial CoNLL-U representation. In subsequent processing (Sect. 5), the CoNLL-U editions of all three corpora are merged into full annotations for two major versions of the Heliand, Ms. C (text basis from HeliPaD) and the synoptic BT text (text basis from DDD).

In Fig. 5, corpora without boxes designate datasets in their original format, corpora in boxes indicate CoNLL-U data, corpora in dashed boxes indicate corpus-specific CoNLL (one-word-per-line TSV) formats. Arrows indicate conversion (e.g., from HeliPaD to HeliPaD UD), training (e.g., from HeliPaD UD to the parser), the application of a tool to data (only for the parser) or merging (e.g., from Heliand B4 CoNLL (with converted manual annotations) and Heliand B4 Parsed (with automated annotations) to Heliand B4 UD. For merging, the thickness of an arrow indicates the priority, e.g., manual annotations from the Heliand B4 CoNLL corpus take priority over automated annotations from Heliand B4 Parsed – but only when the former are actually available.

4.1. HeliPaD conversion

For converting HeliPaD from the Penn bracketing format, we apply a straight-forward replication of rule-based approaches such as Arnardóttir et al. (2020) in SPARQL. As for this transformation, we would argue that Fintan and SPARQL allow to easily replicate existing rules, but that it does so in a declarative, less idiosyncratic and more portable manner, because it is the first tool to perform that task that allows to decouple transformation logic (in SPARQL) from the actual format conversion (by Fintan loader and writer modules), whereas earlier systems are characterized by merging conversion and transformation logic and thus, difficult to re-use and adapt. In fact, Arnardóttir et al. (2020) emphasize that their approach is specific to the IcePaHC corpus, and also, that they felt the need to develop it from scratch because existing software for converting the Penn bracketing format (such as the one by Johansson and Nugues (2007) and the Univer-

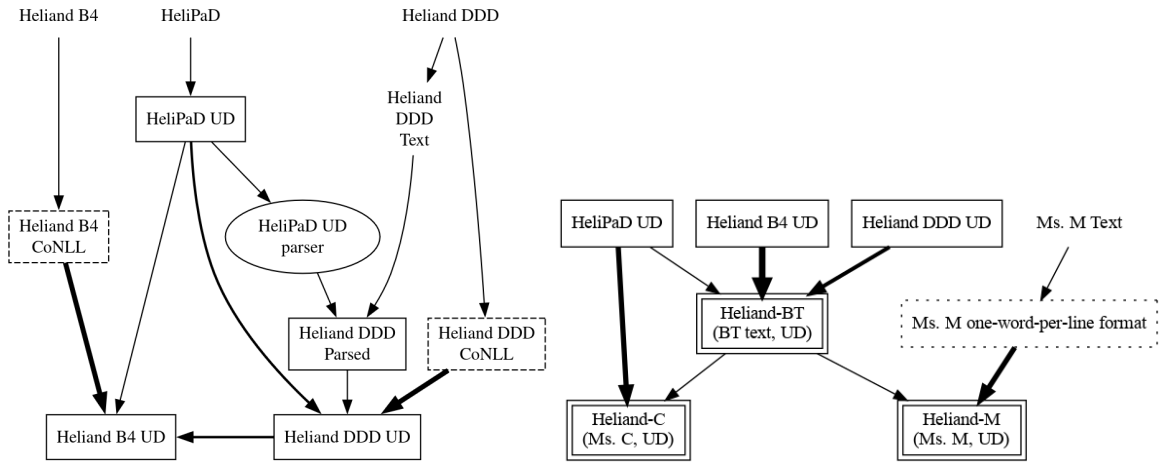

Figure 5: Conversion and consolidation workflows. **left:** conversion of source corpora (Heliand B4, HeliPaD, Heliand DDD) to CoNLL-U format (HeliPaD UD, Heliand DDD UD, Heliand B4 UD). **right:** consolidation of CoNLL-U conversions of source corpora into CoNLL-U editions of different text versions (BT UD: text according to the critical edition by Behaghel-Taeger, with text based on DDD; Ms. C UD: Manuscript C, with text based on HeliPaD; Ms. M UD: Manuscript M, with text based on Schmeller, 1830)

salDependenciesConverter of StanfordNLP[1]) could not be adapted. Beyond that, however, the transformation is otherwise equivalent. We describe details of the SPARQL-based conversion in Chiarcos and Siewert (2026). The focus of this paper is more on innovative aspects of SPARQL, in particular to formulate rules that can operate over *multiple* levels of conflicting syntax annotation.

With HeliPaD converted to CoNLL-U, it is now possible to train the dependency parser of your choice over the Heliand. We conducted our experiments with UDpipe (Straka and Straková, 2017). The trained parser was then used to 'fill in the gaps' of the manual DDD and B4 annotations.

4.2. DDD and B4 conversion

Although they were created with different tools and are distributed in different formats, both the DDD corpus and the Heliand-B4 corpus use a tier-based annotations, where an abstract timeline is annotated with spans, where a span can either represent a token or a sequence of tokens that is then assigned up to one label per layer of annotation (tier). Both DDD and Heliand-B4 provide morphosyntactic annotations, lemmas, clause segments and labels for clause linking, but only B4 also provides a tier for phrases (for NPs, VPs or PPs) and grammatical functions (within the clause). It is to be noted, however, that this is not a full-fledged parse tree as tier-based annotation does not support recursive structures. For the description of the conversion process, we focus on B4, because it provides richer annotations. The general principles and technologies we used are, however, the same for DDD.

We first transform the (DDD and B4) source data into a CoNLL representation. Then, we use CoNLL-Merge to force-align this source data with other ('target') CoNLL annotations which – for tokens that can be aligned – are appended to the original CoNLL columns. For tokens that cannot be aligned, the corresponding number of target columns is filled up with empty annotations. For target tokens that cannot be aligned with source tokens, new rows are added and the columns representing the source data and its annotations are filled up with placeholder symbols (here `*`) by CoNLL-Merge. In our implementation, these are just filtered out. The process can be iterated to align more than two types of CoNLL source source annotations.

For the case of DDD, we merge (a CoNLL representation of) the original DDD annotations with an automated CoNLL-U parse of this data (using a UDpipe parser trained over HeliPaD UD) and with the CoNLL-U conversion of the HeliPaD. Technically, all HeliPaD (ms. C) is included in the (text edition underlying the) DDD corpus. The resulting CoNLL data features a total of 35 columns, and using user-defined labels for all of them, these are mapped to properties in the `conll:` namespace by CoNLL-RDF. Using SPARQL rules as described below, these are then merged by projecting HeliPaD (and, where not available, automatically parsed) dependencies onto DDD data, and using DDD clause labels for inferring dependency labels. The result is then serialized as CoNLL-U.

The B4 corpus was created in EXMARaLDA but has been published in the ANNIS format, so that it can be readily converted to RDF via the SALT

[1] https://nlp.stanford.edu/software/stanford-dependencies.shtml, accessed: 2026-02-15.

component of Fintan (Fäth and Chiarcos, 2022) and serialized into a custom CoNLL format where each tier corresponds to exactly one column. In subsequent processing, this is merged with the resulting CoNLL-U representation of DDD as well as with HeliPaD UD, and a selection (`cut`) of the resulting 58 columns (38 original tiers and 20 CoNLL-U columns) is then further processed by CoNLL-RDF. Although the original corpus does not provide phrasal structures, we could derive a phrase-structural representation from the extent of spans on the respective tiers: We create `powla:Node` entities for every annotation of the respective tier, and `powla:hasParent` relations between overlapping segments according to the following hierarchical organization of tiers: `tok` → `pos` → `cat` (phrase types) → `gf` (grammatical function) → `clause`. The resulting CoNLL-RDF graph is illustrated in Fig. 6. In addition to syntax and glossing, B4 provides annotations for rhyme, bibliographical and context information, as well as annotation for information structure, e.g., aboutness, definiteness, focus-background organization and givenness, excluded from the figure to facilitate readability.

Heliand B4 has been aligned with HeliPaD UD as well as with the CoNLL-U version of Heliand DDD. As its text basis is (except for variance in transliteration) identical to that of DDD, it is not directly parsed automatically, but incorporates major aspects of the automated parses via the DDD alignment.

For inferring consolidated dependencies from the Heliand B4 tree of `powla:Node`s and the DDD- and HeliPaD UD dependencies, three primary transformation steps need to be performed: (1) for every node in the B4 tree, detect the child element that contains the UD head and mark it as `temp:HEAD`, (2-3) for every word that is not a `temp:HEAD`, copy the dependencies from DDD (`conll:DHEAD`), resp. (where not available) HeliPaD-UD (`conll:HHEAD`), and (4-5) collapse the phrase structure to the words that serve as (transitive) `temp:HEAD`s of the respective phrases:

(1) Within every `powla:Node` and every `nif:Sentence`, we identify the syntactic head as the `nif:Word` that has the largest number of dependents in DDD or HeliPaD annotation (as a property path: `(^conll:HHEAD|^conll:DHEAD)* []`). If there are multiple candidate heads, we use the first. In Fig. 6, the `temp:HEAD` edges pointing from `powla:Node`s to the respective head of each phrase are marked in bold.

(2) For siblings in n-ary phrases whose heads (`temp:HEAD*`) are linked in aligned DDD annotation, we copy the DDD dependency and remove the sibling that contains the dependent of the relation.

(3) For siblings in remaining n-ary phrases whose heads are linked in aligned HeliPaD annotation, we copy the HeliPaD dependency and remove the sibling that contains the dependent.

(4) For every binary phrase, we establish a link between the (head of the) non-head child and the (head of the) `temp:HEAD` child and remove the sibling that contains the dependent.

(5) For every remaining n-ary phrase, we link the (heads of) non-head children with the (head of the) `temp:HEAD`.

Note that all these rules exploit the special capabilities that SPARQL property paths provide to navigate within RDF graphs: We use the *transitive closure*, e.g., for the repeated lookup of `temp:HEAD` in (2), marked by `*`, we use the *directional inversion* of `conll:HHEAD` to aggregate over all dependents of a given word in (1), marked by `^`, we use the *disjunction* between different RDF properties in (1), marked by `|`, etc. In code, each of these steps is represented by (at least) one SPARQL Update, appropriately formatted and commented, separated by empty lines and `;` and aggregated into a single `*.sparql` file. Fintan then applies one or more `*.sparql` files in their sequential order to every RDF graph, and, optionally, also to perform loops.

Dependency labels (`conll:EDGE`) are also induced in this process. In (2-3), these are adopted from DDD, resp. HeliPaD along with the transfer of dependencies. For unlabelled dependencies, steps (4-5) provide a mapping from the annotations of the `powla:Node`s that the respective `nif:Word` is `temp:HEAD` of (`(^temp:HEAD)*`). Primarily, this exploits the grammatical function (`gf`) annotation. For words without `gf` annotation, we resort to DDD labels. `UPOS` is adopted from DDD. Figure 7 shows the consolidated result graph.

5. Consolidating Annotations

As the corpora cover different parts from different manuscripts, these need to be regrouped so that we provide individual annotations for the different textual witnesses. Fig. 5 summarizes the annotation consolidation workflow, with source corpora in CoNLL-U marked in boxes, corpora in other CoNLL (one-word-per-line TSV) formats in dotted boxes, and resulting corpora in double boxes. The arrows indicate transformation and merging processes.

Overall, we focus on both major versions of the text for which we possess manual annotations: Heliand-C (Ms. C) primarily based on HeliPaD and Heliand-BT (synoptic text) primarily based on DDD (with additions from B4 and HeliPaD).

Heliand-BT provides the text of the critical edition by Behaghel and Taeger (1984), the most compre-

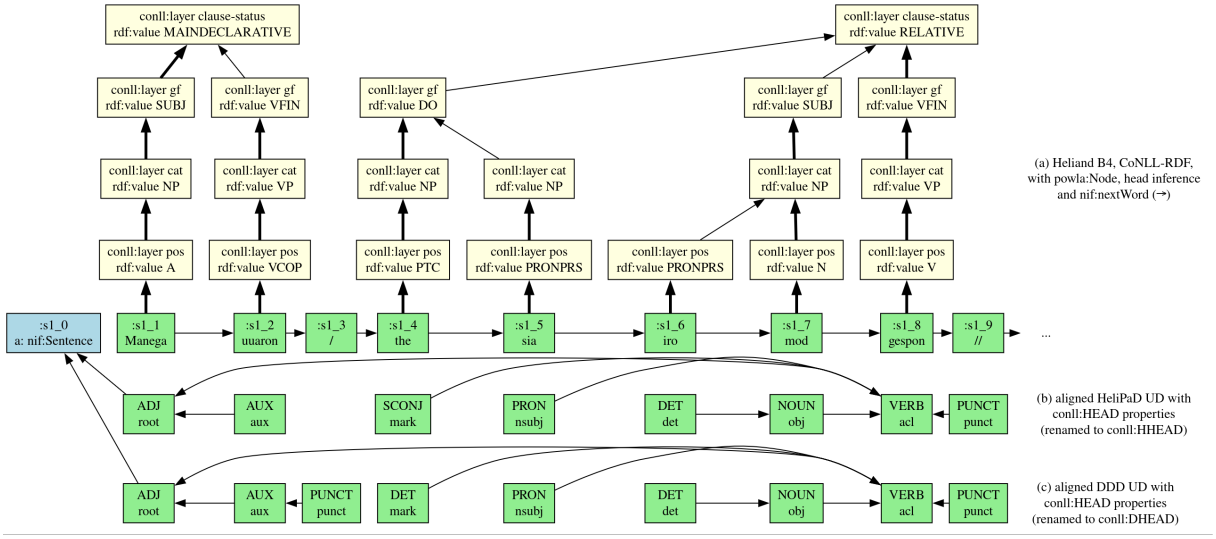

Figure 6: First verse of Heliand B4, compact visualization of a single CoNLL-RDF graph consisting of (a) B4 annotations with inferred `powla:hasParent` (↑, **↑**) and `temp:HEAD` properties (inverse of bold **↑**), (b) merged HeliPaD UD annotations over the same `nif:Word`s, and (c) merged DDD UD annotations.

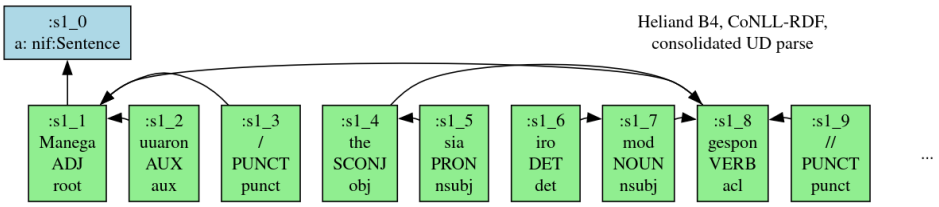

Figure 7: Heliand B4, consolidated CoNLL-RDF graph as constructed from Fig. 6 with SPARQL Updates, with `conll:HEAD` properties (←, →, ↑); `nif:nextWord` not shown.

hensive text version of the Heliand, but, in comparison to the source manuscripts, amended at several occasions. It forms the basis of both Heliand DDD and Heliand B4, but it should also contain all textual material covered by Ms. C (i.e., HeliPaD), as well as Ms. M (not annotated, so far), the most extensive witnesses. Building a consolidated and internally consistent Heliand UD corpus over all three corpora thus required to first align all subcorpora with the BT text (as provided by the DDD corpus) and to aggregate their information, and then to project (parts of) their information into the annotated corpora, resp., Ms. M.

As our CoNLL-U edition of the DDD corpus already builds on an alignment with HeliPaD, the DDD CoNLL-U corpus was taken as a basis and aligned with the B4 CoNLL-U corpus. The input to CoNLL-RDF is thus a 20-column format that provides DDD CoNLL-U, followed by B4 annotations (where available) for all DDD words, and using DDD sentence splits.

With SPARQL Update, the merged dependency annotations are copied from B4 (because this provides richer manual annotations), or from DDD if no B4 annotations are available. In case of cycles, we remove all B4 annotations for the current sentence and resort to DDD. The values for `UPOS`, `XPOS` (i.e., original `POS` of HeliPaD and `pos` and `morph` annotations of DDD concatenated) and `LEMMA` are taken from DDD. For debugging purposes, the `MISC` column is used for tracking where which piece of annotation came from. In release data, this information is subsequently removed. The result is serialized in CoNLL-U. Figure 8 illustrates an excerpt of the merged corpus.

Effectively, the resulting parse is 99% (54735/55080 tokens) identical with DDD (because of the limited size of B4 Heliand, because of using DDD as fallback where alignment fails, and because DDD parses are mostly confirmed by B4). This does not mean that they are error-free, because the B4 UD annotation is partially derived from DDD, and because both source corpora provide *partial* manual annotations for syntax only, and new errors may have been introduced both by the alignment with HeliPaD and the automated parsing (unless superseded by the partial annotations provided by DDD or B4), whereas morphosyntax is (mostly) based on original DDD annotations.

Heliand-C is primarily based on the previously con-

```
1 *      ,      PUNCT CODE|_                              _ 2 punct _ DDD
2 Manega manag  ADJ   Q^N^PL|DIS.MASC_PL_NOM_ST           _ 0 root  _ B4=DDD
3 uuâron wesan  AUX   BEDI^3^PL|VVFIN.IND_PAST_PL_3       _ 2 aux   _ B4=DDD
4 ,      ,      PUNCT ,|$,                                _ 2 punct _ DDD
5 the    de     DET   C|DDSREL.MASC_PL_ACC                _ 9 mark  _ DDD
6 sia    he     PRON  PRO^PL^A^3|PPER.MASC_PL_ACC_3       _ 5 nsubj _ B4
7 iro    he     DET   PRO$^SG^3^N|DPOS.MASC_PL_GEN_3      _ 8 det   _ DDD
8 môd    mod    NOUN  N^N^SG|NA.SG_NOM                    _ 9 nsubj _ B4
9 gespôn spanan VERB  GE+VBDI^SG^3|VVFIN.IND_PAST_SG_3   _ 2 acl   _ B4
```

Figure 8: Merged Heliand-BT annotations derived from DDD and B4 (and, indirectly, HeliPaD), with provenance information in the `MISC` column

verted HeliPaD corpus, aligned with CoNLL-Merge with the Heliand-BT corpus and enriched in `XPOS` and for dependency labels for clausal juncture: The HeliPaD dependencies were left untouched, but for cases in which the same dependency relation received a different label in Heliand-BT, we resort to the Heliand-BT label. If Heliand-BT `XPOS` annotations were compatible (i.e., starting with) HeliPaD `XPOS` annotations, the Heliand-BT `XPOS` was used.

For **Heliand-M**, a plain text edition of Ms. M (Schmeller, 1830) was annotated experimentally both by the HeliPaD parser and by alignment with Heliand-BT. However, we observed a large number of alignment errors: At many occasions, the scribe decided to concatenate morphological words that Ms. C and BT would consider to be independent words, at others, a word is split. So, we read *inatorht lico* in Ms. M for *ina torhtlîco* '(that reminded) him (of) shining (times)' in BT (Ms. C: *ina torohtlico*). As a result, the alignment identifies *inatorht* as pronoun and direct object (correct for *ina*, only), and *lico* as adverb. A more correct analysis that stays true to the text and that is consistent with UD should introduce a multi-word annotation for *inatorht*, and create a `flat` link between the sub-token *torht* and *lico*. However, these kind of corrections would require manual oversight, and unless this can be provided, the Heliand-M build scripts and its textual source file are provided, but the data will not be released.

6. Results and Perspectives

In this paper, we described the end-to-end consolidation of heterogeneous syntactic annotations for the Old Saxon *Heliand* across multiple corpora, annotation styles, and textual witnesses. The workflow comprises (i) the conversion of several corpora with different syntactic coverage and formalism into CoNLL-style formats and, for the fully parsed HeliPaD corpus, into CoNLL-U, (ii) the training of a UD parser on the resulting HeliPaD-UD data, (iii) the enrichment of partially annotated corpora (Heliand B4 and Heliand DDD) by complementing manual annotations with automatically generated dependency parses, (iv) the merging of all annotations into a master representation based on the most comprehensive consolidated text (Behaghel/Taeger, BT), and (v) the alignment and projection of annotations from this master edition onto different textual witnesses (notably manuscripts C and M).

These steps yield three main contributions. First, we provide the first UD-compliant dataset for Old Saxon. Second, we demonstrate the application of CoNLL-Merge to the alignment of divergent witnesses and editions of the same source text while preserving their respective annotations, even where these are incomplete or structurally incompatible. Third, we show how CoNLL-RDF, Fintan and SPARQL can be used for rule-based merging and consolidation of multiple layers of syntactic annotation, including conflicting and partial analyses, within a unified graph-based framework.

In the broader landscape of historical Germanic corpora, several syntactically annotated resources have been created in recent years. Within the Universal Dependencies ecosystem, however, only Old Icelandic, Gothic, Old English and Middle High German are currently covered. Old English UD treebanks are largely based on conversions of Penn Treebank-style corpora. From a purely technical perspective, converting another Penn-style treebank such as HeliPaD to UD is therefore not fundamentally novel, and could be achieved with dedicated conversion scripts. Likewise, training a UD parser on the converted corpus follows established methodology. The more innovative aspect lies in the consolidation of multiple overlapping annotation layers from independent projects, combining HeliPaD, DDD, B4 and parser-based annotations. Other than graph technologies as used here, we are not aware of an existing solution that could have been used for this purpose, because existing tools typically specialize either in dependency or in constituency representations and offer limited support for handling both simultaneously. Moreover, independently created historical corpora usually differ in transliteration, tokenization and editorial principles, which makes their integration difficult and often requires substantial manual effort.

Our workflow addresses these obstacles by combining token-level alignment with graph-based representation and rewriting. CoNLL-Merge supports automated alignment across divergent tokenizations and editions and produces joint CoNLL representations with parallel annotation columns. CoNLL-RDF and Fintan convert these tabular structures into RDF graphs that can host multiple syntactic layers in parallel. Distinct annotation layers are represented through user-defined column labels and properties, enabling their separation and joint processing. SPARQL `SELECT` queries support cross-layer search and aggregation, while SPARQL Update rules enable rule-based consolidation, repair and restructuring of annotations, including the correction of alignment errors and the controlled prioritization of manual over automatic analyses. Because Fintan processes sentences as independent RDF graphs, this can be executed efficiently and in parallel with minimal memory requirements.

Related work has shown that CoNLL-RDF and Fintan can support complex annotation engineering tasks, including the merging of complementary syntactic and semantic layers into theory-specific representations, rule-based enrichment and rewriting, and joint querying across heterogeneous annotation styles. The present study extends this line of work to the consolidation of annotations that differ not only in scheme and depth but also in their underlying primary data. Our resources cover manuscripts and editions with partial textual overlap, normalized versus manuscript-faithful spelling, and divergent principles of transliteration, punctuation and word segmentation. We show that even under these conditions, large-scale consolidation is feasible using RDF-based representations and SPARQL-driven transformation.

So far, evaluation focused on technological feasibility rather than linguistic quality, as the contribution of this paper is primarily methodological and infrastructural: we establish a reproducible, extensible workflow for integrating heterogeneous legacy annotations into a coherent UD-style corpus. We adhere to manual annotations wherever available, but the exact rules employed require additional evaluation, or, adjustment on the basis of manually annotated dependency data. For Old Saxon, such data is currently prepared by the authors.

Three core technologies are central to this workflow: CoNLL-Merge for cross-version alignment and column-wise integration of annotations, CoNLL-RDF/Fintan for conversion between CoNLL(-U) and RDF and for multi-layer graph representation, and UDPipe for training a baseline parser that fills structural gaps in partially annotated corpora. Together, these components form a modular toolkit for post-hoc annotation consolidation.

From a linguistic perspective, the resulting resources open perspectives beyond the *Heliand*. For the closely related Old High German and Old Low Franconian (Old Dutch), no dependency corpora are known to exist, and only partial syntax annotations are available from DDD. As far as historical German is concerned, existing corpus initiatives and annotation experiments primarily address younger stages of the language (Sapp et al., 2024; Dipper et al., 2024; Haiber, 2024). Directly applying a the HeliPaD parser (or a novel Old Saxon parser) to these is difficult due to orthographic and dialectal variation, but where consistent lemmatization, part-of-speech tags and agreement features are available (as in DDD), it is straightforward to construct a representation that abstracts away from surface spelling. Training parsers over Old Saxon text normalized accordingly might offer a direction toward broader parsed resources. Even if automatically generated, such corpora would substantially improve empirical coverage for the study of historical West Germanic syntax.

We publish code and data described here as the Consolidated Old Saxon (ConOS) corpus `https://github.com/nds-spraakverarbeiden/ConOS` under different open source licenses. The prospective goal is to provide a UD-compliant edition of the Old Saxon data of the DDD and HeliPaD corpora, which includes the Heliand, the Old Saxon Genesis and a number of smaller fragments. At the moment, we plan to provide automatically processed data only, as created with the workflows described above. As for licensing data, we are bound to the licenses of the respective source corpora, for aggregated data, this means to follow the most restrictive source license. Fortunately, DDD, B4 and HeliPaD licenses are compatible with each other, so that this is legally possible.

In order to evaluate not only the viability of the conversion, but also the linguistic quality of the converted and consolidated annotations, we created a small training set of about 2,000 tokens based on DDD text and morphosyntax, and released it as part of the Universal Dependencies under `https://github.com/UniversalDependencies/UD_Old_Saxon-ConOS`. The UD ConOS corpus currently consists of the first fit (chapter) of Heliand (806 tokens, based on B4 text and DDD morphosyntax) and the second fragment of the Old Saxon genesis (1,145 tokens, based on the DDD corpus). Initial results on the evaluation of the consolidated annotations against the Heliand part and on linguistic aspects of the UD annotation of Old Saxon have been reported in Chiarcos and Siewert (2026). The Genesis part has been prepared for a future evaluation of automated parsing and its consolidation with DDD morphosyntax on unseen text.

7. Bibliographical References

Þórunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigurðsson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinþór Steingrímsson. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW-2020)*, pages 16–25, Barcelona, Spain (online).

Otto Behaghel and Burkhard Taeger. 1984. *Heliand und Genesis*, 9 edition. Max Niemeyer Verlag, Tübingen.

Hannah Booth, Anne Breitbarth, Aaron Ecay, and Melissa Farasyn. 2020. A Penn-style Treebank of Middle Low German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC-2020)*, pages 766–775, Marseille, France (online).

Cathy Bow, Baden Hughes, and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings of E-MELD Workshop 2003: Digitizing and annotating texts and field recordings*, pages 11–13, East Lansing, Michigan.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media / multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*, pages 2065–2068, Lisbon, Portugal.

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, Anette Frank, and Ulrich Heid. 2008. Formalising multi-layer corpora in owl dl-lexicon modelling, querying and consistency control. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge (LDK-2017)*, pages 74–88, Galway, Ireland. Springer.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022. Querying a dozen corpora and a thousand years with Fintan. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC-2022)*, pages 4011–4021, Marseille, France.

Christian Chiarcos and Luis Glaser. 2020. A Tree Extension for CoNLL-RDF. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, pages 7161–7169, Marseille, France (online).

Christian Chiarcos, Maxim Ionov, Luis Glaser, and Christian Fäth. 2021. An Ontology for CoNLL-RDF: Formal Data Structures for TSV Formats in Language Technology. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Christian Chiarcos and Niko Schenk. 2018. The ACoLi CoNLL libraries: Beyond tab-separated values. In *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Christian Chiarcos and Janine Siewert. 2026. Towards a Universal Dependency corpus for Old Saxon (Old Low German). In *Ninth Workshop on Universal Dependencies (UDW-2026)*, Palma de Mallorca, Spain. Co-located with LREC 2026.

Stefanie Dipper, Cora Haiber, Anna Maria Schröter, Alexandra Wiemann, and Maike Brinkschulte. 2024. Universal Dependencies: Extensions for modern and historical German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17101–17111.

Christian Fäth and Christian Chiarcos. 2022. Spicy salmon: Converting between 50+ annotation formats with Fintan, Pepper, Salt and POWLA. In *Proceedings of the 8th Workshop on Linked Data in Linguistics (LDL-2022), held in conjunction with LREC-2022*, pages 61–68, Marseille, France.

Christian Fäth, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 2020. Fintan - Flexible, integrated transformation and annotation engineering. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, pages 7212–7221, Marseille, France (online).

Jost Gippert. 2011. The TITUS project. 25 years of corpus building in ancient languages. In *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. Internationale Tagung des Akademienvorhabens Altägyptisches Wörterbuch.

Cora Haiber. 2024. A Crosslingual Approach to Dependency Parsing for Middle High German. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 23–31, Vienna, Austria.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA*

2007), pages 105–112, Tartu, Estonia. University of Tartu, Estonia.

Seth Kulick, Neville Ryant, and Beatrice Santorini. 2022. Penn-Helsinki parsed corpus of early Modern English: First parsing results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 578–593, Seattle, United States. Association for Computational Linguistics.

Sonja Linde. 2009. Aspects of word order and information structure in Old Saxon. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change: New approaches to word order variation in Germanic*, pages 367–389. Walter de Gruyter.

Sonja Linde and Roland Mittmann. 2013. Old German Reference Corpus: Digitizing the knowledge of the 19th century. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J Whitt, editors, *New Methods in Historical Corpora*, pages 235–246. Narr Francke Attempto Verlag, Tübingen.

Rosemarie Lühr. 2025. Reflexivität im Altsächsischen. In Norbert Kössinger, editor, *Altsächsisch: Beiträge zur altniederdeutschen Sprache, Literatur und Kultur*. Walter de Gruyter.

Nicolas Mazziotta. 2010. Building the syntactic reference corpus of Medieval French using NotaBene RDF annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pages 142–146, Uppsala, Sweden.

Svetlana Petrova. 2006. A discourse-based approach to verb placement in Early West-Germanic. *ISIS| Working Papers of the SFB 632| 5 (2006)*, page 153.

Svetlana Petrova and Michael Solf. 2009. On the methods of information-structural analysis in historical texts: A case study on Old High German. In Roland Hinterhölzl and Svetlana Petrova, editors, *Information structure and language change: New approaches to word order variation in Germanic*, pages 121–203. Walter de Gruyter.

Svetlana Petrova, Michael Solf, Julia Ritz, Christian Chiarcos, and Amir Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. In *Traitement Automatique des Langues, Volume 50, Numéro 2: Langues anciennes [Ancient Languages]*, pages 47–71.

Susan Pintzuk and Ann Taylor. 1997. Annotating the Helsinki Corpus: The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English and the Penn-Helsinki Parsed Corpus of Middle English. In *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop, New College, University of Toronto, Toronto, May 1995*, pages 91 – 104. Brill, Leiden, Niederlande.

Christopher D. Sapp, Elliott Evans, Rex Sprouse, and Daniel Dakota. 2024. Introducing a Parsed Corpus of Historical High German. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9224–9233, Torino, Italia. ELRA and ICCL.

Johann Andreas Schmeller. 1830. *Heliand: oder die altsächsische Evangelienharmonie*. Cotta.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In *The Oxford handbook of corpus phonology*.

Edward Henry Sehrt. 1925. *Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis*. Vandenhoeck & Ruprecht.

Eduard Sievers. 1878. *Heliand*. Buchhandlung des Waisenhauses, Halle.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Ann Taylor. 2003. The York—Toronto—Helsinki parsed corpus of Old English Prose. In *Creating and digitizing language corpora: Volume 2: Diachronic Databases*, pages 196–227. Springer.

George Walkden. 2016. The HeliPaD: a parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21(4):559–571.

Victim or Assailant? Exploring Narratives Through Knowledge Graph Queries

Beatrice Fiumanò[*1], **Nicolas Lazzari**[*], **Simone P. Ponzetto**[†], **Valentina Presutti**[*]
University of Bologna, Italy[*], University of Mannheim, Germany[†]
{beatrice.fiumano,nicolas.lazzari3,valentina.presutti}@unibo.it, ponzetto@uni-mannheim.de
Corresponding author[1]

Abstract

Our understanding of social reality is shaped by the specific ways in which that reality is framed by different sources. Analyzing framing means examining how these sources are able to convey particular worldviews by foregrounding or downplaying certain aspects of experience. Current computational approaches address this task by automatically identifying communicative patterns (e.g., topic selection or rhetorical strategies) that characterize individual artifacts. However, they often remain document-bound, overlooking the comparative dimension that enables the uncovering of convergent or conflicting narratives about the same actor, event, or issue. In this paper, we propose DORIS, an ontology that supports both document-level and cross-document framing analysis using SPARQL queries on automatically constructed Knowledge Graphs. We validate the proposed approach through a case study of historical news articles, exploring multiple framings of a real-world event using Fillmore's Frame Semantics and the FrameNet resource. Code and data are available on GitHub at https://github.com/beatrice-f/DORIS/.

Keywords: Knowledge Graph, Ontology, Framing Analysis, Frame Semantics

1. Introduction

Decades of research across multiple disciplines, from linguistics to the social sciences, have highlighted that communication is never neutral, neither in the production nor in the interpretation of meaning (Hall, 1973; Foucault, 2013).

Any communicative artifact promotes a specific worldview that embeds social and political structures, personal and collective biases, ideology, and stance (Fairclough, 1995). This constitutes an act of *framing*, understood as the ad-hoc construction of meaning by selectively foregrounding or silencing certain aspects of experience (Entman, 1993). As a result, framing shapes how reality is perceived.

Consider, for instance, these two excerpts from different news sources covering the anti-Nazi protest that took place in New York on 20 February 1939[1], extracted from the NewsWire (NW) dataset (Silcock et al., 2024):

> "*Outside Madison Square Garden policemen had a six-hour struggle with throngs of anti-Nazis who repeatedly charged their lines trying to fight their way inside.*"
> *NW5635*

> "*..., a moving throng of anti-Nazis, theatergoers and the merely curious milled about in the streets. About 1,500 police reserves stood guard over the area, while violence spurted up inside the Garden and out.*"
> *NW523*

Excerpt from *NW5635* emphasizes the physical confrontation between police officers and anti-Nazi protesters, who are presented as a disruptive force. In *NW523*, the emergence of violence does not have a clear agent: Protesters are portrayed as part of a bigger, heterogeneous crowd, whereas police officers are represented as a surveilling presence. We observe here that communication is, in itself, an act of perspectivization or interpretation of reality.

To analyze framing dynamics, scholars have adopted a variety of methodological approaches. Traditionally, research in discourse and narrative analysis has relied on qualitative, human-driven investigations (Entman, 1993; Fairclough, 1995; Wodak, 2001).

These approaches guarantee high-quality, nuanced insights, but they are both time- and effort-intensive, and susceptible to researcher bias (Parks and Peters, 2023). By contrast, computational approaches ensure reproducibility and scalability, enabling large-scale analyses that would otherwise be impractical (Hamborg, 2023; Parks and Peters, 2023).

A number of approaches have been proposed in the NLP field, allowing scholars to systematically analyze framing (Otmakhova et al., 2024). Nonetheless, they are generally bound to a single document, hampering a broader understanding of framing across multiple documents (Otmakhova et al., 2024; Ali and Hassan, 2022).

In this work, we propose overcoming this limitation by representing artifacts, their metadata, and automatically derived observations in a structured way using Knowledge Graphs (KGs) (Hogan et al., 2021). Our approach is based on the *encoding-*

[1] https://en.wikipedia.org/wiki/1939_Nazi_rally_at_Madison_Square_Garden

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 40–49
12 May 2026.

decoding model (Hall, 1973), where a real-world situation (e.g., a protest) undergoes an *encoding* process that produces a semiotic artifact (e.g., a news article) that is finally *decoded* by a cognizer (e.g., a reader of the news article).

We propose organizing knowledge in the KG using a novel ontology, DORIS, based on the established Description and Situation Ontology Design Pattern (Gangemi and Mika, 2003; Gangemi and Presutti, 2009). DORIS allows dynamics of framing to be modeled within a broader relational system that supports cross-artifact and cross-language investigations and higher-level pattern detection. By leveraging KGs, semiotic items are linked through disambiguated actors, events, and topics, enabling systematic comparison across sources, languages, and contexts, and the exploration of how events, actors, and issues are constructed across artifacts.

We demonstrate this approach by leveraging Fillmore's Frame Semantics (FS) (Fillmore, 1976), operationalized in FrameNet (Baker et al., 1998), as the guiding framework for analyzing discourse and semiotic representations, since it enables exploration of the discursive conceptualization of events and participants (Ziem, 2014). For example, FS shows that excerpt *NW5635* interprets the circumstance as a HOSTILE_ENCOUNTER between a `Side_1` (police officers) and a `Side_2` (anti-Nazi protesters), while text *NW523* employs a metaphor that casts violence as a `Fluid` within the FLUIDIC_MOTION frame, hence providing scholars with insights that "go beyond just words" (Ali and Hassan, 2022).

Nonetheless, our approach is not bound to a specific theory and can be adapted to different operational frameworks, such as the Moral Foundation Theory (Graham et al., 2013) or the Narrative Policy Framework (Shanahan et al., 2018), as well as a combination of them.

In summary, our contribution is threefold:

C1) we present DORIS (**D**omain-specific **O**ntology for **R**epresenting and **I**nterpreting **S**emiotic artifacts), a novel ontology modeling the relationship between a semiotic artifact, the real-world situation it represents, and the interpretation of that situation conveyed in the artifact;

C2) we show that DORIS supports the exploration of semiotic representations and framing strategies across different artifacts through the use of SPARQL queries, lowering the barrier to the use of complex computational methods;

C3) we present a use case that illustrates ontology-grounded KG construction and exploration, leveraging FS at the interpretative layer, using historical news articles.

The rest of the paper is organized as follows: in Section 2, we provide an overview of the main approaches to framing analysis in NLP. In section 3, we present DORIS and describe its main modeling choices, which directly influence the construction of the KG and how it is queried. We demonstrate the feasibility of our approach through a case study based on historical documents, presented in Section 4. Finally, we summarize our contributions and highlight limitations and future extensions in Section 5.

2. Related Work

As introduced in Section 1, computer-driven techniques for framing analysis have gained traction in the past few decades, particularly within the NLP and computational social science fields. In this section, we provide an overview of the main computational approaches to framing analysis, specifically focusing on those that, like our proposal, rely on graph-based methods and FS.

Computational Approaches As noted by Otmakhova et al. (2024), computational approaches to the study of framing dynamics are heterogeneous, varying across disciplines and the focus of analysis. In the NLP community, most research efforts focus on emphasis framing, i.e., issue dimensions explored through topic modeling or predefined codebooks (Ali and Hassan, 2022). Although popular, this line of research tends to simplify frames as topics. Another line of research adopts a more fine-grained approach, focusing on lexical (Wicke and Bolognesi, 2025), syntactic, and discursive choices (Reinig et al., 2024). Among these, several studies have relied on (frame-)semantic parsing to explore semantic relations in text. Adopting an approach similar to the one proposed in this paper, both Postma et al. (2020) and Minnema et al. (2022) use FrameNet frames to analyze variations in the framing of events and social issues.

Furthermore, Postma et al. (2020) underscore the importance of anchoring narratives to real-world entities and events, and of resolving them across documents. In line with this intuition, some studies have shown that a more effective approach for inter-document exploration of narratives is to structure them as graphs.

Graph-based Approaches Following this argument, Baden (2018) proposes to model news discourse as a network of entities and applicability relations among them, enabling the exploration of entity-centric framing patterns across news items. Similarly, Pournaki and Willaert (2025) rely on semantic dependencies to extract narrative signals from a corpus of political speeches and represent

them as graphs. As in our approach, framing can be analyzed through graph queries. Focusing on event-centric representations, Rovera et al. (2021) employ a graph-based model that encodes both surface-level information (e.g., actors, locations, and explicit relations) and the deeper semantic argument structure of events. Motta et al. (2025) propose a formal, ontology-based typology for news classification that captures the topic or issues addressed in an article and the issue-specific claims and viewpoints it conveys, supporting fine-grained content analyses.

Our approach follows this stream of research with a specific focus on adopting LOD, as we will show in Section 4.4.

3. The DORIS Ontology

The main objective of this work is to demonstrate how KGs can support query-driven narrative exploration and extract meaningful insights into the framing strategies encoded within a given artifact. To achieve this, knowledge must be organized in the KG to support answering relevant research questions, such as "How is the police framed in the document *NW5635*?", and "How do role attribution and agency of anti-Nazi protesters vary across documents?"

In this section, we describe DORIS, a novel ontology designed to support answering similar research questions. As introduced in Section 1, DORIS draws on Hall's *encoding/ decoding* theory (Hall, 1973). This theory conceptualizes communication as a process in which a state of affairs (e.g., a protest) is encoded into a semiotic artifact (e.g., a news article) by an agent operating within a specific socio-cultural and ideological context. This encoding process involves selecting a preferred reading or interpretation of events, which results in framing by means of specific lexico-syntactic and narrative choices (Entman, 1993). The encoded artifact is subsequently decoded by some agent, whose interpretation is also mediated by their own beliefs, values, and background.

DORIS, shown in Figure 1, models the encoding-decoding process by reusing the Description And Situation (Gangemi and Mika, 2003, DnS) Ontology Design Pattern (Gangemi and Presutti, 2009, ODP) and the DOLCE ontology (Borgo et al., 2022).

Conceptually, the ontology is organized into four layers modeling different steps of the process: the *grounding layer*, the *encoding layer*, the *decoding layer*, and the *theory layer*.

The *grounding layer* models the fact that a DUL:ENTITY (e.g., a protest) is encoded into multiple artifacts, collected in an :ENCODINGCOLLECTION (e.g., a set of news articles), all realized through the same :MODE (e.g., natural language). Note

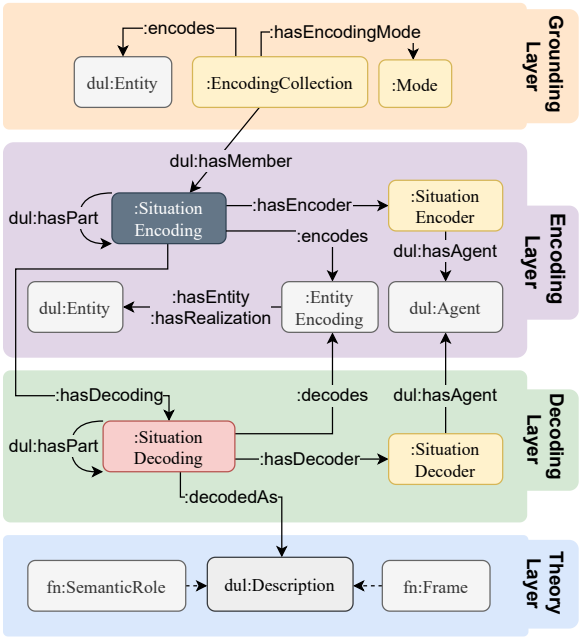

Figure 1: The DORIS ontology represented using Graffoo. Grey boxes refer to reused ontology classes.

that due to the generality of DUL:ENTITY instances, DORIS allows modeling the framing of concrete entities (e.g., a politician) as well as abstract concepts (e.g., gender equality) or social entities (e.g., the government).

Each member of an :ENCODINGCOLLECTION is a semiotic artifact (e.g., a news item or a painting), formalized by the :SITUATIONENCODING class in the *encoding layer*. A :SITUATIONENCODING instance is a DnS Situation. Informally, a DnS Situation represents a relation over multiple entities, where each entity takes a specific role.

In a :SITUATIONENCODING, these entities are: (i) the agent (or possibly a group of agents) that produced the artifact, represented by the :SITUATIONENCODER class, (ii) the :ENTITYENCODING, which is itself a DnS Situation involving the entity encoded and how it has been realized in the artifact (e.g., its textual mentions) and (iii) the decoding(s) of an artifact, represented by the :SITUATIONDECODING class. Finally, a :SITUATIONENCODING might be decomposed into smaller parts. For instance, in the case of a news article, each of its paragraphs is again :SITUATIONENCODING instances.

The decoding output is represented by the :SITUATIONDECODING class, which is also a DnS Situation. Similar to a :SITUATIONENCODING, a :SITUATIONDECODING instance involves the agent performing the decoding and the entity being decoded. Unlike a :SITUATIONENCODING, a :SITUATIONDECODING instance involves a DnS Description that is used to decode it. A DnS Description instance represents a class of situations that share some aspect between

them (e.g., the set of all individuals decoded as victims).

The DnS Descriptions and the relations holding between them are formalized in the *theory layer*. In Figure 1, we show a possible instantiation of a theory layer using FS as formalized in Framester (Gangemi et al., 2016) (as further illustrated in Section 4). More generally, this layer is theory-agnostic, and any suitable theoretical framework can be adopted, provided that it is formalized as an ontology that re-uses the DnS ODP. In Section 4, we demonstrate that this approach enables the integration of multiple theories, yielding deeper insights into the framing of an entity. Finally, a :SituationDecoding can also be decomposed into smaller parts, similarly to a :SituationEncoding instance.

Figure 2 shows an example of how DORIS is used to represent the encoding and decoding of the social entity `police` in excerpts *NW5635* and *NW523* from Section 1. Although the two documents use different surface forms ("policemen" vs "police"), they both refer to the same conceptual entity (e.g., a general police entity represented by node Q35535 on Wikidata). DORIS allows comparing how the same entity is decoded across documents by explicitly decoupling a superficial mention from the actual entity it describes, overcoming the limitation of analyses bound to a single article. Figure 2 also shows how the choice of a theory layer does not influence the analysis of an artifact.

4. Case Study: Analyzing Media Framing in Historical Newspapers

For our use case, we focus on entity framing (Mahmoud et al., 2025), and in particular on how media sources represent an event and its participant entities. As previously illustrated, highlighting certain aspects of an event, such as confrontations between police officers and protesters, while downplaying or omitting others, such as the actors responsible for violence escalation, promotes different worldviews. Consequently, entities may acquire different connotations depending on the actions they are described as performing or undergoing. Within this use case, our objective is to examine how the constituent happenings of an event are portrayed across news articles and, based on these portrayals, how participant entities are assigned different semantic roles.

In this section, we outline a methodology for constructing a KG, structured according to DORIS, from a news article. Nonetheless, it is possible to apply a similar pipeline to other modalities beyond text, as also discussed in Section 5.

4.1. KG Construction

Constructing a KG from text has been widely explored as part of the Knowledge Extraction subfield of NLP, where one is interested in identifying a set of entities and the relations that hold between them starting from a natural language sentence (Zhong et al., 2024).

Given a set of documents describing the same event (e.g., excerpts *NW5635* and *NW523*), we rely on the set of named entities annotated in the documents and perform document-level coreference resolution, by relying on ChatGPT 5.1[2] and manually refining predictions. This step normalizes all textual mentions to refer to the same entity, ensuring consistency in the KG. Despite manual effort, we highlight that this approach can be automated using recent advances in NLP, e.g., Martinelli et al. (2025).

We then identify semantic relations between entities in the documents using the LOME frame semantic parser (Xia et al., 2021), which has been trained on FrameNet. Finally, we convert LOME's predictions to a KG in RDF representation compliant with DORIS. To perform this step, we use SPARQL Anything (Asprino et al., 2023), which enables KG construction from heterogeneous document types. The result of this step produces a KG that follows the structure of the ones shown in Figure 2.

4.2. Dataset

In order to demonstrate how one can query the KG to analyze the framing in an article, we base our case study on a selection of five news pieces extracted from the NewsWire dataset[3] (Silcock et al., 2024), which contains 2.7 million historical news wire articles in English. Since we are interested in identifying texts covering the same historical event, we filter the dataset by year and topic, selecting articles published in 1939 and classified under the topic Protest. Among this subset, we select five articles reporting on the Nazi rally held in New York on February 20, all issued by U.S. news wire services. Excerpts from *NW5635* and *NW523* reported in Section 1 have been extrapolated from these articles.

4.3. Qualitative Analysis

Before delving into the query-based framing analysis, we report a qualitative analysis conducted by one of the authors with a background in linguistics. This analysis serves a dual purpose. On one hand, it provides a reference for evaluating whether query

[2] https://openai.com/gpt-5/

[3] https://huggingface.co/datasets/dell-research-harvard/newswire

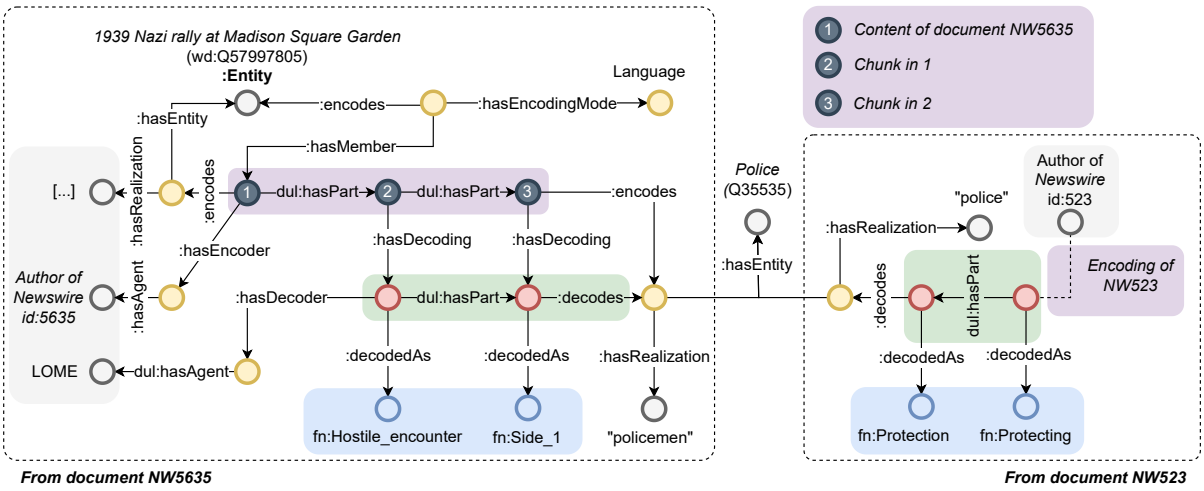

Figure 2: Example usage of DORIS to represent the decoding of the entity *police* in the excerpts *NW5635* and *NW523*. The block on the left shows how the metadata of the article is encoded in a KG (grey box) as well as how the event is encoded (purple box) in a news article (node 1), which contains a sentence from it (node 2) mentioning the entity police as "policemen" (node 3). Using FS, the sentence evokes the frame HOSTILE_ENCOUNTER with police taking the role of SIDE_1 (i.e., one of the forces of the hostile encounter). Similarly, the block on the right shows that a sentence mentioning "police" evokes the frame PROTECTING, with police taking the role of PROTECTION (i.e., the entity that prevents harm to another entity). For the sake of compactness, we omit metadata and encoding/decoding instances related to this node.

results yield meaningful insights for framing analysis. On the other hand, it allows to assess whether the knowledge derived from the KG is consistent with human observations[4].

In the following paragraphs, we examine the semiotic portrayal of the main actors involved in the news coverage of the rally, including the German American Bund (GAB), its paramilitary forces, and its leader, Fritz Kuhn; U.S. police; and U.S. citizen Isadore Greenbaum. Specifically, each article reports on two key episodes of the rally: the confrontations between police officers and counter-protesters, discussed above, and the episode involving Kuhn and Greenbaum, who leapt on the stage while the former was delivering an antisemitic speech. For each article, we analyze the stance adopted toward the actors and the discursive construction of both episodes.

NW523 This article attributes different evaluative weights to the key social actors. The Bund is construed as an ideologically marked and conflict-generating presence. Ideological non-alignment is encoded, for instance, in the structure *Bund* → *denounces* → *American alliance with European democracies*, where "European democracies" functions as a positively loaded term whose alliance is, however, positioned as the target of denunciation. Storm troopers are framed as a source of violence in the Kuhn-Greenbaum episode, in which they are described as "setting upon" and 'knocking down" Greenbaum. Conversely, the police are portrayed as a rescuing force that intervenes to remove Greenbaum from the troopers' attack. Both the anti-Nazi protesters and the crowd are framed through limited agency, as violence is described as "spurting up" and the outside crowd is grouped into a heterogeneous group of people who "mill about".

NW3763 A similar framing of storm troopers and police forces emerges in *NW3763*. Here, Greenbaum is explicitly construed as a victim of the trooper's violence, saved by police intervention. While his actions are presented with little evaluative loading (he "advanced toward" Kuhn), Kuhn is framed as an antagonist through the depiction of his speech as a bitter attack on the Jews. Although no anti-Nazi group is explicitly mentioned, the outside crowd is here portrayed as a source of public disorder, characterized as quarrelsome and as disrupting the order maintained by the police.

NW5635 In this item, greater focus is given to the order-maintaining role of the police, as illustrated in Section 1, and on the disruptive conduct of counter-protesters. The GAB is again ideologically marked through a detailed account of the rally cere-

[4]We note that, given the limited scope of the study and the potential influence of annotator bias, this analysis is intended solely as a proof of concept. Further research is required for a systematic evaluation, as discussed in Section 5.

mony, which included “denunciations of Jews” and “salutes to swastikas”. The Greenbaum episode is given limited space and is explicitly framed as an “attempt to attack” the Bund leader, with omission of the ongoing antisemitic speech. A narrative reorientation occurs only at the last sentence of the article, where Bund members are construed as patients of an `attack` event with unspecified agency.

NW6072 As in the previous item, the core of this article lies in the positive construal of the police, here depicted as “a wall of invulnerable blue”. With limited attention devoted to the rally itself, the police role is emphasized by narrative minimization of conflict (“general rioting had not developed”, “inside the Garden itself only one fight occurred”). In contrast, the crowd is positioned as an antagonistic force, being politically identified as anti-Fascist and depicted as actively attempting to breach a “forbidden area”.

NW19221 In this article, police forces are again foregrounded as an order-containing force (“a solid ring of shoulder-to-shoulder foot and mounted police”) against antagonist anti-Nazi protesters. The GAB is strongly ideologically marked through explicit labeling of the rally as “anti-Jewish” and “anti-communist”. The Greenbaum episode is narrated in detail and framed as the central moment of violence. While his advance toward Kuhn is described with limited evaluative loading, storm troopers are assigned explicit and graphic violent agency (“felled”, “seized”, “hurled”), positioning Greenbaum as the patient of violence, who is later rescued by police. Greenbaum is subsequently also framed through institutional culpability, as he is “arraigned” and charged with disorderly conduct. At the end, his request for a doctor after the arraignment reintroduces vulnerability.

4.4. Framing Analysis by Querying a Knowledge Graph

Although we rely on a limited number of articles, the analysis in the previous section highlights several differences in how the same event is framed by different authors (in our setting, different newswires).

In this section, we demonstrate how SPARQL queries over a DORIS-based KG can support the exploration of relevant research questions in discourse and framing analysis. For each query, we report a representative subset of results, selected based on their relevance to the qualitative analysis conducted in Section 4.3[5]. However, we note that, due to the number of less relevant or incorrect frames detected by the frame semantic parser, a considerable amount of noise emerges from each query. We further discuss these limitations in Section 5.

In this section, we focus on three main operations that support framing analysis: document-level analysis, cross-document analysis, and Linked Open Data (LOD)-enriched browsing. For each scenario, we present one or more queries related to our use case, and discuss the results in relation to the human-driven analysis outlined in Section 4.3.

Document Level: How is entity *X* framed in a given document? We examine here the framing of the social actor Isadore Greenbaum within the news item *NW3763* using the following Query 1:

```
SELECT ?role WHERE {
 ?articleSituation a :SituationEncoding;
    :encodes [ :hasRealization [
          provo:wasGeneratedBy <NW3763>
    ] ];
    :hasDecoding ?articleDecoding .
 ?articleDecoding dul:hasPart [
      :decodes [
        :hasEntity <Isadore Greenbaum>
      ];
      :decodedAs ?role] .
}
```

Query 1: What are the roles assigned to Isadore Greenbaum in document NW3763?

Semantic Frame	Role
Experience_bodily_harm	Experiencer
Cause_impact	Impactee
Cause_harm	Victim
Attack	Assailant

Table 1: Roles associated with Isadore Greenbaum in document NW3763 in the context of a semantic frame. Wrong parser predictions are highlighted in gray.

A relevant subset of the results, consistent with human observations, is reported in Table 1. The VICTIM role in the CAUSE_HARM frame, the EXPERIENCER role in the EXPERIENCE_BODILY_HARM frame, and the IMPACTEE role in the CAUSE_IMPACT frame align with the fact that the article largely construes Greenbaum as a victim. Nonetheless, the ASSAILANT role is incorrectly assigned to Greenbaum instead of Kuhn in the sentence “Kuhn had been bitterly attacking the Jews”, due to an erroneous prediction of LOME.

Cross-Document Level: What is the framing of Entity X across different documents? Query 2

[5]We report the complete set of query outputs in the GitHub repository, including less relevant and noisy results

investigates the construal of a given entity across all available documents. In this paragraph, we retrieve information relevant to analyzing the framing of the police and stormtrooper entities.

```
SELECT ?source ?description WHERE {
 ?articleSituation a :SituationEncoding;
    :encodes [ :hasRealization [
        provo:wasGeneratedBy ?source
    ]];
    :hasDecoding ?articleDecoding .
  ?articleDecoding dul:hasPart [
    :decodes [
      :hasEntity <police/storm troops>
    ];
    :decodedAs ?description ] .
}
```

Query 2: What are the roles assigned to police across documents?

Article	Semantic Frame	Role
NW5635	Hostile_Encounter	Side_1
NW5635	Hostile_Encounter	Depictive
NW5635	Hostile_Encounter	Side_2
NW19221	Lvl_of_Force_Res.	Resisting_Entity
NW523	Protecting	Protection

Table 2: Roles associated with the police across documents. Wrong parser predictions are highlighted in gray.

Table 2 shows the results of Query 2 for the entity "police". The framing emerging from the query results corroborates human observations. Across documents, the police are construed as active agents in physical confrontations, who oppose an antagonistic crowd for the maintenance of social order. This pattern is reflected in the police's realization of the SIDE_1 role within the HOSTILE_ENCOUNTER frame and the RESISTING_ENTITY role, defined in FrameNet as an entity that "is capable of resisting or does resist a physical OPPOSING_FORCE". It is further instantiated through the PROTECTOR role in the PROTECTING frame, where "Protection prevents a Danger from harming an Asset".

Similarly, the query results for the storm troopers (Table 3) align with their qualitative portrayal as a violent actor. Across nearly all documents, they are instantiated in the AGENT role within the CAUSE_HARM and CAUSE_IMPACT frames, as well as in the ASSAILANT role in the ATTACK frame, consistently encoding them as a source of physical aggression.

Refinement and Expansions using External Knowledge Bases In this paragraph, we highlight some advantages of leveraging KGs to represent discourse and discourse dynamics, namely

Article	Semantic Frame	Role
NW3763	Cause_Harm	Agent
NW3763	Cause_Impact	Agent
NW5635	Cause_Harm	Agent
NW523	Cause_Impact	Agent
NW523	Attack	Assailant
NW523	Arrest	Charges
NW19221	Cause_Harm	Agent

Table 3: Roles associated with the storm troopers across documents. Wrong parser predictions are highlighted in gray.

the use of external Linked Open Data (LOD). The integration of LOD enables access to external knowledge bases (KB), which can be exploited to refine or expand an initial query with additional semantic information. In Query 3, we show how the Framester KB (Gangemi et al., 2016) can be used to retrieve only those instances in which Isadore Greenbaum instantiates a Frame Element of the ATTACK frame.

```
SELECT ?source ?role ?domain WHERE {
 ?articleSituation a :SituationEncoding;
    :encodes [ :hasRealization [
      provo:wasGeneratedBy ?source
    ]];
    :hasDecoding ?articleDecoding .
 ?articleDecoding dul:hasPart [
  :decodes [
    :hasEntity <Isadore Greenbaum>
  ];
  :decodedAs ?role ] .
 SERVICE <Framester SPARQL endpoint> {
   ?role rdfs:domain fsframe:Attack .
  }
}
```

Query 3: In which documents does Isadore Greenbaum instantiate a role belonging to the ATTACK frame?

Article	Semantic Frame	Role
NW3763	Attack	Assailant
NW5635	Attack	Assailant
NW523	Attack	Victim

Table 4: Roles associated to Isadore Greenbaum and belonging to the ATTACK frame. Wrong parser predictions are highlighted in gray.

By filtering the query, we observe in Table 4 how Greenbaum instantiates both the ASSAILANT and the VICTIM roles, depending on the article. This variation reflects a difference in construal, whereby item *NW5635* frames Greenbaum's leap onto the stage as an "attempt to attack" Fritz Kuhn, while item *NW523* foregrounds his patient role in the storm trooper's assault.

Finally, in Query 4, we exploit supplementary knowledge available in Framester where FrameNet frames are aligned to MFT Graham et al. (2013), as formalized in the ValueNet ontology (Giorgis et al., 2022).

```
SELECT ?source ?value (COUNT(*) AS
    ?count) WHERE {
 ?articleSituation a :SituationEncoding;
   :encodes [ :hasRealization
     [ provo:wasGeneratedBy ?source
   ]];
   :hasDecoding ?articleDecoding .
 ?articleDecoding :decodedAs
    ?description .
  SERVICE <Framester SPARQL endpoint> {
    ?description vcvf:triggers ?value.
  }
  FILTER CONTAINS(STR(?value),
    "HaidtValues")
}
GROUP BY ?source ?value
```

Query 4: What is the number of occurrences of Moral Foundations in each document?

Article	Care	Oppression	Harm	Loyalty
NW523	3	2	1	1
NW5635		1	7	1
NW3763			6	
NW19221		1	2	1
NW6072				1

Table 5: Occurrences of Moral Foundations per article.

Overall, the distribution of MF (Table 5) endorses our qualitative analysis.

As many articles foreground physical confrontation, the *Harm* moral foundation predominates, typically activated by frames such as ATTACK. References to Greenbaum's arrest also activate *Oppression*. By contrast, the order-maintaining role of the police evokes *Care*, which in article *NW523* is instantiated through the PROTECTING frame. The *Loyalty* value is recurrent across documents, mostly triggered by the MEMBERSHIP frame associated with Bund members and the COME_TOGETHER frame describing the rally crowd.

5. Limitations and Future Work

In this paper, we presented DORIS, an ontology based on Hall's encoding-decoding theory that supports cross-document framing analysis by representing an artifact and its theory-grounded interpretation in a KG that can be easily queried via SPARQL. Using a case study on historical news articles, we have shown that the retrieved results align with human-derived findings. Section 4.4 shows that, although we experimented with Fillmore's FS as our theoretical lens, it is possible to expand the analysis to other relevant theories (e.g., the Moral Foundation Theory). This approach paves the way to the development of a human-in-the-loop paradigm that leverages different NLP tools to support framing analysis.

Nonetheless, our results also highlight a key limitation of the proposed approach, namely the frame semantic parser's incorrect predictions. Further research is required to investigate the impact of such errors on the overall quality of the analysis, and to identify strategies for minimizing expert effort, for instance by automatically refining LOME's outputs or by developing more accurate frame semantic parsers, especially for multilingual settings. Along similar lines, we plan to perform a more systematic evaluation of the devised method using established framing and discourse analysis benchmarks. Additional future work includes enhancing the KG construction pipeline described in Section 4.1 by integrating Entity Linking systems, which would greatly enhance the potential of this approach (akin to state-of-the-art machine readers such as Text2AMR2FRED (Gangemi et al., 2026)), as well as using LLMs to elicit implicit knowledge, as illustrated in De Giorgis et al. (2025).

Finally, we highlight that although our case study focuses on the linguistic domain, DORIS has been designed to be medium- and domain-agnostic. Following a similar approach to the one in Section 4, it is possible to formalize the representation and interpretation process across different domains. For instance, Ciroku et al. (2024) relies on FS to analyze the implicit meaning of images by leveraging the relations among the entities they depict. By relying on DORIS, it is possible to harmonize such analyses across multiple images, allowing scholars to explore image collections through theoretically grounded interpretations.

Moreover, DORIS bridges the gap between large-scale *distant reading* techniques and more qualitative *close reading* approaches by preserving an explicit link between semiotic or discursive features and their analytical interpretations. Unlike many automated methods, the use of KGs allows scholars to derive more general, cross-artifact observations while retaining the possibility of progressively refining the analysis down to the surface (e.g., textual) level to examine the specific instantiations of a given phenomenon (Rovera et al., 2021).

6. Bibliographical References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Con-*

ference on Empirical Methods in Natural Language Processing, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Luigi Asprino, Enrico Daga, Aldo Gangemi, and Paul Mulholland. 2023. Knowledge graph construction with a façade: A unified method to access heterogeneous data sources on the web. *ACM Transactions on Internet Technology*, 23(1):1–31.

Christian Baden. 2018. Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In *Doing news framing analysis ii*, pages 3–26. Routledge.

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. *Appl. Ontology*, 17(1):45–69.

Fiorela Ciroku, Stefano De Giorgis, Aldo Gangemi, Delfina Sol Martinez Pandiani, and Valentina Presutti. 2024. Automated multimodal sensemaking: Ontology-based integration of linguistic frames and visual data. *Comput. Hum. Behav.*, 150:107997.

Stefano De Giorgis, Aldo Gangemi, and Alessandro Russo. 2025. Neurosymbolic graph enrichment for grounded world models. *Information Processing & Management*, 62(4):104127.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Norman Fairclough. 1995. *Critical Discourse Analysis: The Critical Study of Language*. Longman, London.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Michel Foucault. 2013. *Archaeology of knowledge*. Routledge.

Aldo Gangemi, Arianna Graciotti, Antonello Meloni, Andrea Giovanni Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, and Alessandro Russo. 2026. Text2amr2fred, converting text into rdf/owl knowledge graphs via abstract meaning representation. *Knowledge and Information Systems*, 68(1):47.

Aldo Gangemi and Peter Mika. 2003. Understanding the semantic web through descriptions and situations. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE - OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003*, volume 2888 of *Lecture Notes in Computer Science*, pages 689–706. Springer.

Aldo Gangemi and Valentina Presutti. 2009. Ontology design patterns. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 221–243. Springer.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Stuart Hall. 1973. Encoding and decoding in the television discourse. Discussion paper, ARRAY(0x56177e4471c0), Birmingham. This paper has been published in: CCCS selected working papers. Vol. 2 / edited by Ann Gray et al (Abingdon, 2007) pp. 386-398.

Felix Hamborg. 2023. *Revealing Media Bias in News Articles: NLP Techniques for Automated Frame Analysis*. Springer Nature, Cham, Switzerland.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer.

Tarek Mahmoud, Zhuohan Xie, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025. Entity framing and role portrayal in the news. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 302–326. Association for Computational Linguistics.

Giuliano Martinelli, Bruno Gatti, and Roberto Navigli. 2025. xcore: Cross-context coreference resolution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November*

4-9, 2025, pages 34264–34278. Association for Computational Linguistics.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

Enrico Motta, Enrico Daga, Aldo Gangemi, Maia Lunde Gjelsvik, Francesco Osborne, and Angelo A. Salatino. 2025. The epistemology of fine-grained news classification. *Semantic Web*, 16(3).

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

Louisa Parks and Wim Peters. 2023. Natural language processing in mixed-methods text analysis: A workflow approach. *International Journal of Social Research Methodology: Theory & Practice*, 26(4):377–389.

Marten Postma, Levi Remijnse, Filip Ilievski, Antske Fokkens, Sam Titarsolej, and Piek Vossen. 2020. Combining conceptual and referential annotation to study variation in framing. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 31–40, Marseille, France. European Language Resources Association.

Armin Pournaki and Tom Willaert. 2025. Extracting narrative signals from public discourse: a network-based approach. *Humanities and Social Sciences Communications*, 12(1):1774.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Marco Rovera, Federico Nanni, and Simone Paolo Ponzetto. 2021. Event-based access to historical italian war memoirs. *J. Comput. Cult. Herit.*, 14(1).

Elizabeth A Shanahan, Michael D Jones, Mark K McBeth, and Claudio M Radaelli. 2018. The narrative policy framework. In *Theories of the policy process*, pages 173–213. Routledge.

Philipp Wicke and Marianna M. Bolognesi. 2025. Red and blue language: Word choices in the trump and harris 2024 presidential debate. *PLOS ONE*, 20(6):1–30.

Ruth Wodak. 2001. The discourse-historical approach. In Ruth Wodak and Michael Meyer, editors, *Methods of Critical Discourse Analysis*, pages 63–94. Sage, London.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2024. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4):94:1–94:62.

Alexander Ziem. 2014. *Frames of Understanding in Text and Discourse: Theoretical Foundations and Descriptive Applications*. John Benjamins, Philadelphia.

7. Language Resource References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: A wide coverage linguistic linked data hub. In *European knowledge acquisition workshop*, pages 239–254. Springer.

Stefano De Giorgis, Aldo Gangemi, and Rossana Damiano. 2022. Basic human values and moral foundations theory in valuenet ontology. In *Knowledge Engineering and Knowledge Management - 23rd International Conference, EKAW 2022, Bolzano, Italy, September 26-29, 2022, Proceedings*, volume 13514 of *Lecture Notes in Computer Science*, pages 3–18. Springer.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. Newswire: A large-scale structured database of a century of historical news.

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

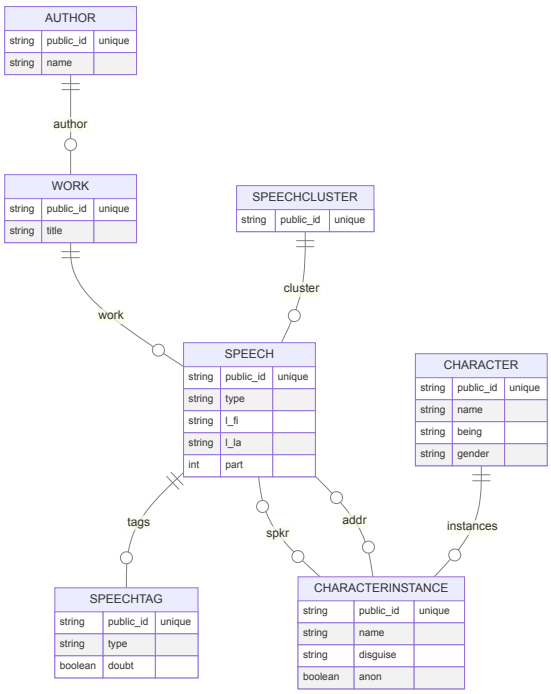

Figure 1: Simplified structure of the DICES database.

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

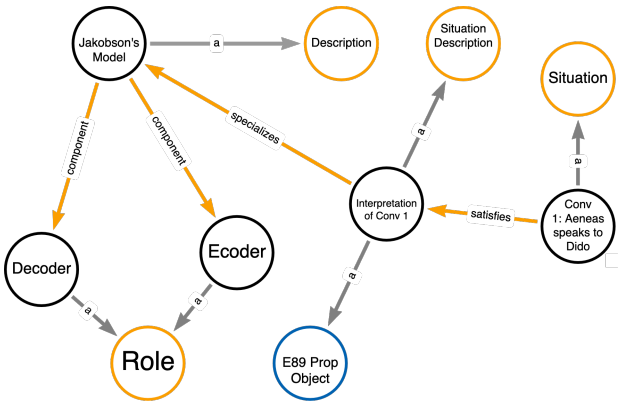

Figure 2: Model of a communication theory (simplified Jakobson model) as a Description and an application to a fictitious scene where Aeneas speaks to Dido (Conv 1), modeled as a Situation-Description. Yellow is used for classes and properties from DOLCE+DnS, blue for CIDOC CRM.

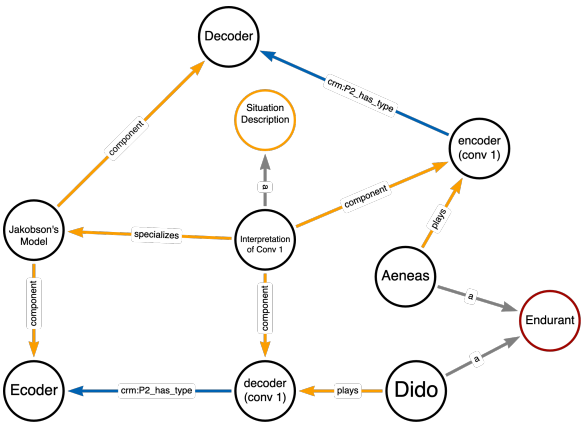

Figure 3: Character instances (endurant), local and general roles in the interpretative frame of the (fictitious) speech scene between Aeneas and Dido (same as in Fig. 2). Red is used for the core DOLCE classes.

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.

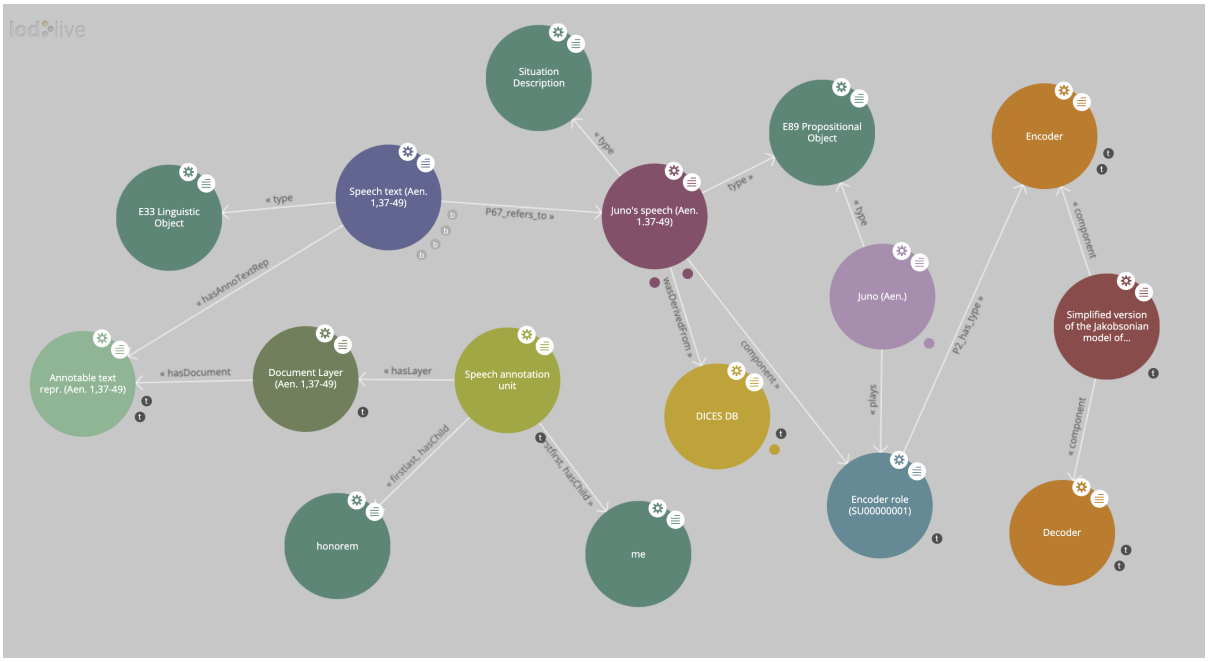

Figure 4: Overview of a speech (Juno's soliloquy in *Aeneid* 1.37-49).

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.

Open English NameNet: Extending English Wordnet with Names

John P. McCrae
Research Ireland Insight and ADAPT Centres, Data Science Institute, University of Galway
john@mccr.ae

Abstract

We present Open English NameNet, a new large-scale lexical resource that extends Open English Wordnet with named entities derived from Wikidata. While English Wordnet has historically included many proper nouns, its coverage has been incomplete and inconsistent, and encyclopedic knowledge sources have grown rapidly in parallel. To address this gap, we systematically extract and align named entities from Wikidata with the Open English Wordnet hierarchy, ensuring each entity is appropriately placed through instance hypernym relations. Our methodology combines existing WordNet–Wikipedia mappings with Wikidata information and applies domain-specific strategies for people, plants and animals, and languages, to account for structural and semantic differences between the resources. This approach results in the largest English lexical-semantic resource currently available, with extensive coverage and structured integration. We release the resource openly to support the development of lexically and encyclopedically informed language technologies.

Keywords: WordNet; Named Entities; Wikidata; Lexical Resources; Knowledge Graph

1. Introduction

Open English Wordnet (McCrae et al., 2019, 2020, OEWN) is an open-source fork of the original Princeton WordNet (Miller, 1995; Fellbaum, 2010)[1]. However, a major issue with the development of this resource has been the size of the wordnet and the large number of proper nouns and other named concepts inherited from the Princeton WordNet. These concepts range from widely known concepts that would be of value to lexicography, such as the names of countries and languages, to concepts that are much more obscure, such as the names of minor historical figures and obscure names for plants and animals. Meanwhile, large encyclopedic resources, such as Wikidata[2], have arisen that cover such concepts far more completely than could ever be achieved in the limitations of a lexicographic resource. As such, in this paper, we propose a new resource, Open English NameNet, built from the proper nouns of OEWN extended with concepts extracted from Wikidata, leaving the remaining section of Open English Wordnet to focus on the common nouns, verbs, adjectives and adverbs of the language. The two resources can follow different quality guidelines, with OEWN having higher quality guarantees, such as unique definitions for all concepts, while NameNet can be expanded to have a very wide coverage. The resources are designed to be combined and the combination of these is the largest lexicographic resource for English by a substantial margin.

The work of creating Open English NameNet is fundamentally based on the mapping between WordNet and Wikipedia and we rely on mapping from several sources, namely existing mappings in Open English Wordnet, created by McCrae and Cillessen (2021), Grammatical Framework (Angelov, 2020), BabelNet (Navigli and Ponzetto, 2010; Navigli et al., 2021) and mappings from YoVisto (Bergh et al., 2025). However, the mapping is not simply a conversion of the data in Wikidata to the wordnet format, as the new synset must be appropriately placed into the wordnet hypernym hierarchy, and for many concepts, the identification of an appropriate hypernym is a challenging task. Further, in many cases, the mapping of concepts requires further analysis. For example, for people, the superclass in Wikidata is generally human[Q5], but in wordnet the hypernyms are generally to the occupation of that person. Similarly, the organization of plants and animals in wordnet is very different from that of Wikidata and this further complicates the alignment. Finally, we also looked at languages specially as these are not instance hypernyms in wordnet and so the modelling of this also needed to be carefully considered.

The main contribution of this paper is Open English NameNet, a large-scale lexical resource that addresses the "coverage gap" between the high-precision linguistic data in Open English Wordnet (OEWN) and the vast encyclopedic breadth of Wikidata. By systematically extracting and aligning named entities, we extend the traditional wordnet hierarchy with millions of new synsets, creating the **largest English lexical-semantic resource currently available**. Our contribution lies in a novel methodology that maintains semantic coherence through instance hypernym relations while applying domain-specific strategies for complex categories like people, biological taxa, and languages. Unlike previous integrated resources, NameNet is not

[1] 'WordNet' is a trademark of Princeton University; we use 'wordnet' to describe resources following the structure of WordNet

[2] https://wikidata.org

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 60–68
12 May 2026.

merely a link between separate graphs but an expansion of the OEWN hierarchy itself, providing a structural bridge that allows language technologies to unify Word Sense Disambiguation (WSD) and Entity Linking (EL) within a single, consistent framework.

This paper is structured as follows. Section 2 reviews related work on integrating named entities and encyclopedic knowledge into lexical resources, situating our contribution within the broader research landscape and provides background on the differences between wordnets and encyclopedic resources. Section 3 outlines our methodology for constructing Open English NameNet, including the general strategy for hypernym assignment as well as domain-specific approaches for people, plants and animals, and languages. Section 4 describes the resulting dataset, including its size, structure, and modes of distribution. Section 5 provides some applications for the new resource and finally, Section 6 concludes with a discussion of the contributions of this work and future directions for extending and improving the resource.

2. Background

Efforts to extend and enrich lexical resources with encyclopedic knowledge and named entities have a long history. Toral et al. (2008) present a foundational approach, which describes the automatic extension of Princeton WordNet with named entities (NEs). Their method maps the WordNet noun hierarchy to Wikipedia categories in order to identify and extract named entities and their lexical and definitional information. This approach resulted in the enrichment of WordNet with over 300,000 named entities and hundreds of thousands of "instance of" relations, illustrating the potential of leveraging structured encyclopedic resources to complement lexicographic ones.

Wordnets, such as the original Princeton WordNet (PWN) and its successor, Open English Wordnet (OEWN), are primarily lexicographic resources. Their core purpose is to map the semantic relationships between "common" lexical concepts: nouns, verbs, adjectives, and adverbs. In contrast, resources like Wikidata are encyclopedic knowledge bases designed to store facts about specific entities. Wikidata covers nouns, mostly proper nouns and does not contain many verbs, adjectives or adverbs, and these are never marked as such. Further, Wikidata cannot provide the detailed syntactic annotation provided in a resource such as OEWN. As such, there is a substantial gap in the use cases between wordnets and encyclopedic resources like Wikidata. As such, this resource aims to close this "coverage gap" between OEWN and Wikidata. OEWN provides high-quality definitions and linguistic precision, it cannot realistically scale to cover the millions of named entities (NEs), due to the high level of quality and accuracy required in this resource. Conversely, while Wikidata provides vast coverage, it describes the concepts but does not describe the linguistic structures. By extending the wordnet format to include NameNet, this work provides a bridge that allows language technologies to access encyclopedic breadth without losing the structural coherence of a lexical hierarchy.

The most similar resource to Open English NameNet is BabelNet, which combines data from OEWN and Wikidata to create a multilingual dictionary. Navigli et al. (2021) provides an extensive overview of BabelNet, a multilingual lexical-semantic resource that merges heterogeneous sources and has been widely adopted in NLP and AI applications, demonstrating the advantages of multilingual and large-scale integration for both symbolic and statistical methods. In parallel, Rebele et al. (2016) introduces YAGO, an automatically constructed knowledge base that combines Wikipedia, WordNet, and GeoNames. YAGO emphasises high precision in its extraction process and enriches the integrated knowledge graph with temporal and spatial information, supporting complex and expressive queries. Its multilingual dimension further illustrates the benefits of linking lexical resources with broad-coverage encyclopedic data. While these resources integrate English Wordnet and Wikidata into single resources, they do so in a way that does not align with the structure of wordnets and, as such, are novel resources rather than extensions of the existing English Wordnet.

BabelNet (Navigli and Ponzetto, 2010) and YAGO (Rebele et al., 2016) are integrations of heterogeneous sources that retain the distinct structural logic of each. In contrast, NameNet is not a "link" between two separate graphs; it is an expansion of the OEWN hierarchy itself, where every entity is conformed into a wordnet-style synset structure. As such, NameNet takes only the entities that have semantic coherence (i.e., are of the same class) as concepts already in wordnet, and the mapping is non-trivial in that maps are made based on mapping multiple properties rather than the simple links of previous work, as described below for people, biological taxa and languages. NameNet is a resource that not only includes imported data from Wikidata, but also manual modifications developed by the OEWN team. As such, it should be noted that NameNet is an open-source resource to which anyone may contribute, unlike closed resources such as BabelNet and YAGO. Finally, Open English NameNet is, as its name suggests, an attempt to document the English language specifically and does not attempt a 'one-size-fits-all' multilingual approach, opening the door to the development of

namenets for other languages.

A key motivation of Open English NameNet is to provide a clear separation between *instances* and *concepts*. This is a core distinction represented in language by the distinction between common nouns, which represent classes or categories of things (e.g., painter, river, language), and proper nouns, which refer to unique, specific entities that belong to those classes (e.g., Pablo Picasso, The Nile, English). This distinction is represented in nearly all conceptual modelling schemes, for example, OWL distinguishes between 'classes' and 'entities'. In Open English Wordnet, the distinction is formally maintained through the `instance_hyponym` relation for individuals and the `hyponym` relation for sub-categories. Similarly, Wikidata separates these concepts using the 'instance of'[P31] property for unique entities and the 'subclass of'[P279] property for taxonomic hierarchies.

3. Methodology

Our methodology follows a structured mapping process to align named entities from Wikidata to the OEWN hierarchy. In most cases, this involves assigning a suitable general hypernym to each new synset to ensure consistent integration into the lexicon, but specific strategies were followed for some domains such as people, plants and animals, and languages.

3.1. Linking Methodologies

The methodology in this paper is based on an existing bijective mapping between Wikidata and Open English Wordnet. The most recent and complete study is McCrae et al. (2026), which argues that while individual projects like Open English Wordnet (OEWN), Grammatical Framework (GF), BabelNet and yovisto have made significant strides, they remain largely complementary rather than overlapping. For instance, the complete intersection of all four resources covers only 3,017 synsets, whereas 49,219 synsets appear in at least one of these linkings. This discrepancy highlights a major opportunity for consolidation; while OEWN provides high-precision human-in-the-loop mappings (96.1% accuracy), it has the smallest coverage at 12,083 links. In contrast, BabelNet offers the largest dataset with 39,224 links but faces higher disagreement rates with other resources, such as a 14.3% disagreement rate with GF.

We rely on the unified resource they constructed that integrates these disparate mappings while leveraging similarity measures and human-in-the-loop validation. The original links were created by different strategies: GF projected Wikipedia links to Wikidata and merges them with community-contributed synset IDs; BabelNet uses probabilistic alignment algorithms based on contextual evidence like synonyms and gloss definitions ; and yovisto employs a dual-annotator system (DBpedia-Spotlight and yovisto-KEA) to map synsets via DBpedia URIs. The consolidation of these efforts, the proposed integrated resource aims to maximise coverage—which currently reaches approximately 60% of noun synsets—while resolving errors stemming from granularity mismatches and conceptual ambiguities.

The integration of Wikidata into the NameNet framework requires navigating fundamentally different structural philosophies: the Open English WordNet maintains a strictly acyclic taxonomy, permitting multiple inheritance and diamonds (multiple hypernyms which merge into a single concept higher in the taxonomy), but forbidding loops. Wikidata's hierarchy is substantially more detailed and contains directed cycles. As such, we only consider the immediate parent in our mapping, except where noted below, to avoid introducing cycles from Wikidata's hierarchy into NameNet. This approach enables multi-faceted encoding, where an entity such as a fictional character can be simultaneously represented through multiple ontological lenses, for instance, as a pig (biological species), a person (agentive role), and a fictional character (narrative status). The selection of the primary categories currently included (General, People, Species, Taxonomic Categories, and Languages) was motivated by the size of these categories in Wikidata and WordNet and the specific modelling challenges they present. This version focuses on the high-density domains of people, animals and languages and treats other classes only through a general mapping. However, we acknowledge that smaller, more specialised classes may still require more sophisticated, bespoke modelling strategies in future iterations to capture their unique relational nuances.

3.2. General Hypernyms

For general classes, that is instance hypernyms not requiring specific mapping strategies such as people, plants and animals, we primarily base hypernym selection on the Wikidata *superclass*[P31] relationships. The direct mappings were based on previous mappings established by other efforts; this included the mapping developed by McCrae and Cillessen (2021), which has been included in Open English Wordnet. However, we also used links that were in other resources, including Grammatical Framework (Angelov, 2020), BabelNet (Navigli and Ponzetto, 2010) and YoVisto (Bergh et al., 2025).

If no direct mapping exists, we follow the superclass chain transitively to identify a suitable hypernym within the existing WordNet hierarchy. When

Conflict case	Frequency	Examples
Both	43	**museum** and **castle**; **Hindu deity** and **king**
Accept One	78	**museum** not **organization**; **political party** not **political movement**
Alternative	16	**hill** and **mountain** → **elevation**[09389214-n]; **bay** and **lake** → **body of water**[09248053-n]
Neither	25	**government agency** or **region**

Table 1: Conflicts between different Wikidata Classes

multiple superclasses or inheritance paths yield conflicting hypernyms, we resolve the conflict by prioritizing the most specific existing hypernym in OEWN; otherwise, the conflict is resolved manually. As shown in Table 1, conflicts may result in accepting one option:

Both In many cases, it was acceptable to use both hypernyms as the classes were compatible, for example, a castle can also be a museum, as these are both locations in the world and so are semantically compatible.

Accept One The most frequent choice was to prefer one of the hypernyms over the other. This was motivated by examining similar cases that are already in OEWN and finding the most similar mapping. For example, a museum may also be an organization and this kind of systematic polysemy (Nunberg, 1992) is common; however museum is the primary function of the organisation, so this is the preferred hypernym. This is based on existing hypernyms of synsets that are organised this way.

Alternative In some cases, we see that the concepts are closely related and thus we can find an alternative concept that works as a hypernym that is consistent with both of the Wikidata superclasses, most frequently this means choosing a common hypernym of the two elements. For example, many entities in Wikidata are tagged as both *hills*[09325914-n] and *mountains*[09382700-n] so the common hypernym *elevation*[09389214-n] is used.

Neither We noted that for many concepts, it was not possible to make an effective rule as to which hypernym is preferred and these must be done on a case-by-case basis. As such, these concepts are not currently part of NameNet. For example, some concepts in Wikidata are classed as both *government agencies*[Q327333] and *region*[Q82794] and these are not really possible to map in a compatible manner.

Each resulting entry in Open English NameNet consists of the following elements that make it compatible with the rest of the structure. The definition of the entry is taken from the Wikidata definition, and unlike those in Open English Wordnet, we do not require that these are unique, and in fact, many of these are quite simplistic, such as 'city in the United States'. There is a requirement that each synset is linked into the wordnet graph by at least one link and this is achieved by an instance hypernym link, which connects the new synset to its appropriate position in the WordNet hierarchy, ensuring that the network remains semantically coherent. The entry also includes one or more lemmas, derived from the English label in Wikidata or from multilingual labels when no English equivalent exists, allowing for broad lexical coverage. In addition, each entry's part of speech is *noun* and a link to the corresponding Wikidata item is included, providing a stable reference for external alignment, future enrichment, and validation of the resource.

Linking Stage	Acceptance	Total Accepted
Occupations	90.1%	1132
Taxons	87.6%	4941
Species	99.5%	5076
Taxon/Common	88.1%	2087
Languages	68.4%	528

Table 2: Acceptance rate and total accepted from the manual linking

3.3. People

One of the key modelling differences between Wikidata and OEWN is the modelling of people, which in OEWN are characterised by their roles, whereas they are organised by their species in Wikidata. We align around the idea of personhood, which refers to entities that are treated as persons within the lexicon, including fictional characters. Philosophically, personhood is defined not by membership of a species but by the possession or attribution of qualities such as agency, intentionality, self-awareness, or social identity (e.g., *Peppa Pig* is a person).

In practice, this means that we consider not only the concepts in Wikidata that are instances

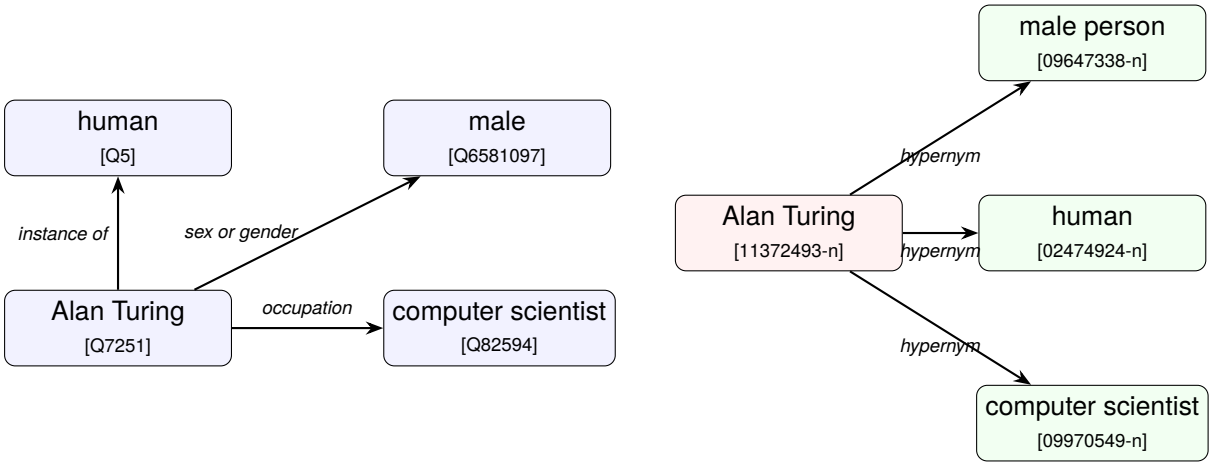

Figure 1: Example of the modelling of people in Wikidata (on the left) and OEWN (on the right). This illustrates the multiple properties used by Wikidata and the new hypernyms that can be inferred from Wikidata.

of *Human*[Q5], but also classes such as *Character*[Q95074]. Further, the modelling of roles in Wikidata is different in that the role is represented by the Occupation[P106] property. Further, we also decided to add specific gender information about the people in the resource, which has been identified as a weakness of OEWN previously (McCrae et al., 2025). In this way, a person in Open English NameNet would have at least three hypernyms: a species hypernym mostly referring to *human*[02474924-n], one or more occupation hypernyms, and a gender hypernym referring to either *male person*[09647338-n] or *female person*[09642198-n].

For example, in Figure 1, we see the modelling of the person 'Alan Turing' in Wikidata, where the modelling provides links that he is an instance of 'human'[H5] and the properties 'sex or gender'[P21] and 'occupation'[P106]. These are mapped onto hypernyms in OEWN. As 'Alan Turing' is already included in OEWN, we add these as extra facts alongside the current hypernym of the synset ('mathematician'[10320928-n]), deepening the representation of this concept in English Wordnet.

Many of the occupations listed in Wikidata were not linked to a synset in Open English Wordnet based on the existing mapping, so we examined the linkings that have been proposed in one of BabelNet, GF or YoVisto, and manually examined these linkings, accepting about 90.1% of these links into Open English Wordnet to improve the coverage of occupations. We also noted during this linking that 136 of the occupation values in Wikidata are invalid, for example, they refer to the organization, such as *police force*[Q35535], rather than the occupation, such as *police officer*[Q384593].

3.4. Plants and Animals

The plants and animals mapping in wordnet uses two linked hierarchies, differentiating between the common names and the scientific taxonomic names for plants and animals. For example, *bear*[02134305-n] is a distinct synset from *family Ursidae*[02134070-n], even though both refer to the same animals. *Bear* is then a hyponym of *carnivore*[02077948-n], while *family Ursidae* is a hyponym of *mammal family*[01865198-n] but a holonym[3] of *order Carnivora*[02077567-n]. However, actual species with binomial names such as *Ursus arctos* are part of the same synset as the common name *brown bear*[02134788-n]. Wikidata has a simpler modelling where common names and taxonomic names are labels of the same entity. To create a structure for plants and animals, that allows Wikidata information to be imported in a way that follows the wordnet hierarchy, it was necessary to convert to the wordnet hierarchy. The structure is shown for brown bears in Figure 2.

We first examined, the taxon names by identifying all the synsets that had a lemma consisting of the words "genus", "family", "order", "class", "phylum" or "kingdom" with a capitalised word in OEWN and looked up the matching name in Wikidata expressed by the property *taxon name*[P225]. These links were then manually validated with an 87.6% acceptance rate (see Table 2). We then also repeated this for binomial species names, where we achieved a much higher acceptance rate of 99.5%, and in fact, all errors were due to issues with OEWN rather than lexical ambiguity. In 15.0% of cases, the binomial species name did not match the corresponding genus linking, which is primarily due to changes in the organization of the species, for example OEWN lists *Felis pardalis*[02128146-n] as the

[3] part/whole relationship

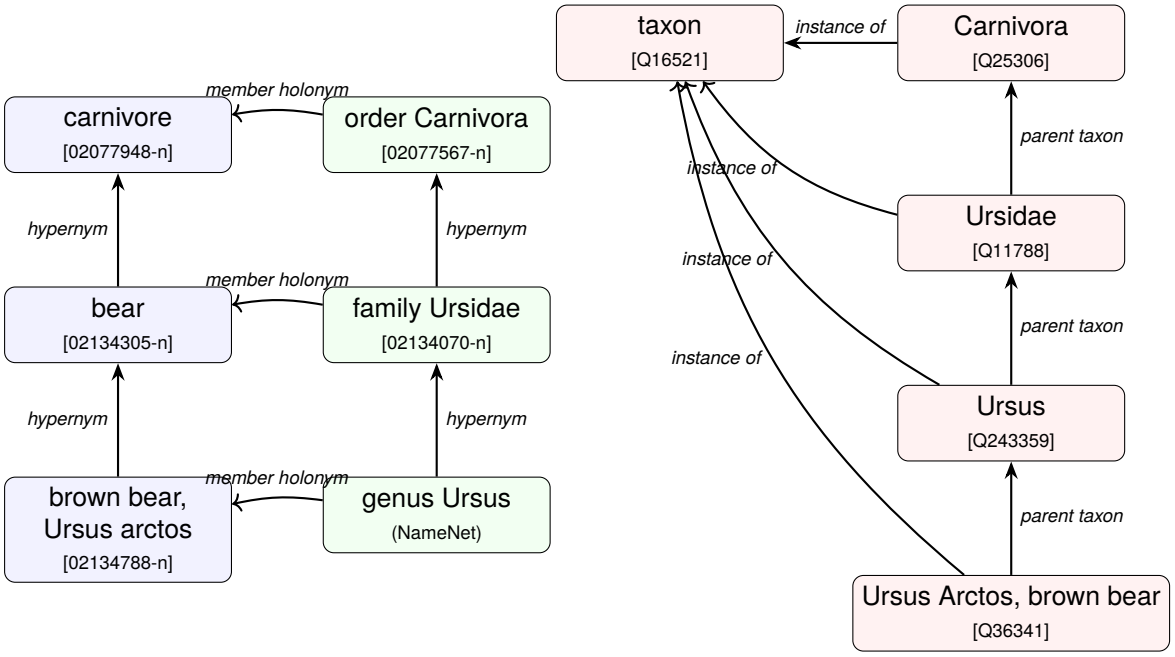

Figure 2: The representation of animals in Open English Wordnet and Wikidata as shown by the modelling of 'brown bear'. The blue represents the common name hierarchy in OEWN, the green the taxon hierarchy in OEWN and red the Wikidata hierarchy. Note the Wikidata hierarchy is simplified; the actual hierarchy includes several intermediate taxons (Ursinae, Ursoidea, Ursida, Arctoidea and Caniformia).

binomial name for the ocelot, this is now part of the genus *Leopardus* so it is known as *Leopardus pardalis*[Q33261] in Wikidata. Wikidata actually assigns unique identifiers for each binomial name, so *Felis pardalis* was first matched to the corresponding *Felis pardalis*[Q122170181] and then the *protonym of*[P12765] property was used to link it to the main entity. In retrospect, due to the high quality of the species linking, it would have been more effective to link species first and use this to confirm the taxonomic categories.

The process of linking thus involves the creation of two synsets in Open English NameNet for each taxonomic classes, such as *genus Ursus*, one using the common names, while the other has only the formulaic lemmas of the name (e.g., *Ursus*) and the name prefixed by the taxon rank (e.g., "genus Ursus"). As the organization of the taxonomic categories in OEWN is based on fairly dated material, it was decided to discard the hierarchy of OEWN and replace it entirely with the hierarchy of Wikidata. As the hypernym of these new concepts is a synset such as *mammal genus*, it was necessary to create a mapping between the common class and the taxon class in OEWN. While this internal link in OEWN is specified, it uses the meronym/holonym property that is also used to indicate the taxonomic hierarchy, i.e., *kingdom Plantae*[11550054-n] is a member meronym of both the common synset *plant*[00017402-n] but also divisions such as *division Bryophyta*[11557229-n]. These were manually disambiguated with an accuracy of 88.1% (see Table 2). The species were created in a manner similar to the general class, but they were given meronymy links to the taxonomic genus that was appropriate. If the species had a protonym, then this was given as a secondary Wikidata link in the data.

3.5. Languages

The modelling of languages in wordnet is also notably different to Wikidata in that the hypernyms are all instances, that is that *English*[06959794-n] is a hyponym, not an instance of *West Germanic Language*[06959585-n] and in fact has further hyponyms such as *American English*[06960241-n]. In order to extend these in NameNet, we then needed to manually extract the languages and create a link between Wikidata and the named languages in OEWN. We used only the name of the language to find candidates for this mapping; however, we found this to be a very high error rate linking with only 68.4% of named languages matching a similar named language in Wikidata. For example, "Manda" can refer to three unrelated languages spoken in Tanzania, India and Australia. This was mainly due to the relative outdated information in wordnet about languages, which often used incorrect spellings or grouped languages now considered to be distinct.

General	Synsets	Lemmas
General	9,673,356	37,678,495
- Star	3,754,413	11,290,295
- Chemical	1,274,034	4,020,428
- Substance	902,749	39,939,24
- Street	642,203	1,984,043
- Mountain	423,511	1,284,921
People	8,506,828	40,266,374
Species	2,757,140	2,876,804
Taxonomic Cat.	257,442	516,117
Languages	8,961	43,844
Total	21,203,727	81,381,634

Table 3: Total resource size in terms of the number of synsets (concepts) and lemmas (words).

4. Dataset

This resource is intended as an extension for Open English Wordnet, and as such, there are three main releases of OEWN available as part of the 2025 edition. Firstly, we provide a version of the OEWN that does not contain NameNet and provide NameNet as a separate version. We also provide a single file in the Global WordNet Format (McCrae et al., 2021), which contains the combined Open English Wordnet and NameNet data. Finally, we also create a legacy version that keeps all the named entities that were part of Princeton WordNet and previous versions of OEWN, but with extra hypernyms and taxon meronyms provided by NameNet.

The overall size of the resource is presented in Table 3, where we present the number of links created by each of the methods. For the general linking, we present the overall number of links created by this linking as well as the five largest subsets according to hypernym. For plants and animals, we distinguish between the species introduced and the taxonomic categories. In total, this introduces a large number of new synsets, making Open English NameNet by far the largest lexicographic resource available for English ever, while maintaining the quality and structure of a wordnet.

4.1. Linked Data Publishing

Open English NameNet as a resource is made available as RDF files through HuggingFace and the linked data interface is under-development to make the resource available through the Open English Wordnet portal. Open English NameNet is published in the Global WordNet Format, which is aligned with OntoLex-lemon and ensures interoperability with the broader ecosystem of linguistic linked data and existing wordnet infrastructures. By maintaining persistent links to Wikidata items, the resource serves as a stable bridge in the Linked Data cloud, allowing for the seamless integration of high-quality lexicographic definitions with vast, crowdsourced encyclopedic knowledge.

5. Applications

5.1. NameNets for Other Languages

While the resources described here is only for English, the methodology described here is designed to be highly replicable for other language-specific wordnet projects, and is designed such that our open-source system can be easily adopted by other langauges to create new namenets. Since Wikidata is inherently multilingual, the majority of the extraction and alignment process, such as following the 'instance of'[P31] and 'subclass of'[P279] chains, is language-independent and can be carried out automatically. As such, the process of creating namenets for new languages, where the wordnet is already aligned with either Princeton WordNet or Open English Wordnet, is quite trivial. However, the procedure for constructing the namenet explicitly excludes concepts with no lexicalisation in English, so some concepts may be excluded. As such, replication would involve:

Initial Alignment The alignment of concepts from the NameNet can be imported and verified in coverage

New Alignments Develop new alignments in line with Section 3.2, for language-specific concepts

Generate New Resource The open-source pipeline[4] can be used to create new synsets

While some manual validation is required to resolve language-specific nuances, the core framework provides a scalable and semi-automated pipeline for expanding any wordnet with encyclopedic named entities.

Further, this work can be of substantial help to the development of wordnets for other languages, as the links provided to these resources can provide lemmas or alternative synonyms for many synsets in these resources. As an example in Open Multilingual WordNet, a simple lemma such as 'Sudan'[09051827-n] has a lemma for only 16 out of 35 (45.7%) of languages, while this lemma is available for all of the languages in Wikidata, often with alternatives like 'Republic of Sudan.'

5.2. Bridging WSD and EL

A significant challenge in Natural Language Processing is the artificial divide between Word Sense

[4] https://github.com/globalwordnet/english-namenet

Disambiguation (Bevilacqua et al., 2021, WSD), which handles common nouns, and Entity Linking (Sevgili et al., 2022, EL), which identifies specific named entities. Traditionally, these tasks require separate models and distinct knowledge bases (e.g., WordNet for WSD and Wikipedia for EL). Open English NameNet acts as a structural bridge between these domains by integrating them into a single, semantically coherent hierarchy. As such, datasets cannot be constructed that deal with both WSD and EL in the same sentence. For example, in "the *mercury* reached record temperatures while the *Mercury* program was still in its infancy", we would see both word sense disambiguation of the word 'mercury' (a common noun referring to the metal) and the proper noun (referring to the NASA space programme). With NameNet, both concepts exist in the same graph, with the same schema. A system can verify the semantic coherence of the sentence by tracing both "senses" to their common ancestors in the OEWN hierarchy. Unlike raw Wikidata, NameNet entries include the linguistic metadata of a wordnet, such as part-of-speech tags and variant lemmas. This enables EL systems to handle morphological variations and syntactic constraints that are usually absent from purely encyclopedic resources.

6. Conclusion

In this paper, we have presented *Open English NameNet*, a large-scale extension of Open English Wordnet that systematically incorporates named entities derived from Wikidata. By decoupling the treatment of common nouns, verbs, adjectives, and adverbs from proper names, we enable different quality and coverage strategies, ensuring that NameNet can grow to cover a broad and dynamic range of concepts while OEWN maintains high lexicographic precision.

NameNet is based on Wikidata and while it is a massive and ever-evolving resource, we do not foresee significant efficiency issues regarding the scalability of NameNet in the near term. Given that Wikidata's growth follows a consistent trajectory, our current ingestion and cycle-breaking pipelines are well-equipped to handle future updates without requiring a fundamental architectural shift.

Our approach combines existing mappings between WordNet and Wikipedia with structured information from Wikidata, supported by tailored mapping strategies for specific domains, including people, plants and animals, and languages. Through this methodology, we achieve both large-scale coverage and a semantically coherent integration into the WordNet hierarchy. This results in the largest English lexical resource currently available, combining the rich linguistic structure of English Wordnet with the encyclopedic breadth of Wikidata, while bridging the gap between word senses and named entities. In future work, we aim to further refine hypernym selection, address ambiguous or unmapped entities through semi-automated alignment techniques, and extend the resource with multilingual correspondences. By making Open English NameNet openly available, we hope to support a wide range of applications in NLP, linguistics and AI that depend on high-quality lexical and encyclopedic knowledge.

Acknowledgements

We would like to thank Al Waskow for their comments on the draft.

John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

7. Bibliographical References

Krasimir Angelov. 2020. A parallel WordNet for English, Swedish and Bulgarian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3008–3015, Marseille, France. European Language Resources Association.

Johann Bergh, Jörg Waitelonis, and Melanie Siegel. 2025. Leveraging LLMs for Constructing WordNets Automatically as Bilingual Resources. In *Proceedings of the 2025 Global WordNet Conference*.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

John P. McCrae, Johann Bergh, Jörg Waitelonis, and Krasimir Angelov. 2026. Towards a comprehensive english wordnet-wikidata mapping. In *Proceedings of the Fifteenth Biennial Language Resources and Evaluation Conference (LREC)*.

John P McCrae and David Cillessen. 2021. Towards a Linking between WordNet and Wikidata. In *Proceedings of the 11th Global Wordnet Conference*, pages 252–257.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The globalwordnet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

John P. McCrae, Haotian Zhu, Fei Xia, Al Waskow, and Kexin Gao. 2025. Remedying Gender Bias in Open English Wordnet. In *Global WordNet Conference 2025*.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten Years of BabelNet: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. ijcai.org.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 216–225. The Association for Computer Linguistics.

Geoffrey Nunberg. 1992. Systematic polysemy in lexicology and lexicography. In *Proceedings of the 5th EURALEX International Congress*, pages 386–396, Tampere, Finland. Tampereen Yliopisto.

Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, volume 9982 of *Lecture Notes in Computer Science*, pages 177–185.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Antonio Toral, Rafael Muñoz, and Monica Monachini. 2008. Named Entity WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

8. Language Resource References

Christiane Fellbaum. 2010. *WordNet*, pages 231–243. Springer Netherlands, Dordrecht.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

John P. McCrae, Ewa Rudnicka, and Francis Bond. 2020. English wordnet: A new open-source wordnet for english. *K Lexical News*, (28):37–44.

George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten Years of BabelNet: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. ijcai.org.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 216–225. The Association for Computer Linguistics.

Bridging the Gap Between Ontologies and Dictionaries: Requirements and Implementation of a New Core for OntoLex-Lemon

John P. McCrae[1], Jorge Gracia[2], Fahad Khan[3], Philipp Cimiano[4]
[1] Research Ireland Insight and ADAPT Research Centres,
Data Science Institute, University of Galway, Ireland
[2] University of Zaragoza, Spain
[3] CNR-ILC, Italy
[4] Cognitive Interaction Technology Center (CITEC), Bielefeld University, Germany
john@mccr.ae, jogracia@unizar.es,
fahad.khan@ilc.cnr.it, cimiano@cit-ec.uni-bielefeld.de

Abstract

This paper presents the requirements and implementation details for a new core module of the OntoLex-Lemon model, representing the first major evolution of the de-facto standard since its 2016 release. While the original model successfully bridged ontologies and dictionaries through "semantics by reference," community adoption has identified critical gaps in handling lexicographic structures and retrodigitized resources. We detail a community-driven methodology that identified fifteen key requirements and we present a proposed architecture for the new OntoLex core, which integrates elements from the Lexicography module and addresses both semantic web and lexicography use cases. Further, we improve interoperability with standards like DMLex and TEI-Lex0 while maintaining strict backwards compatibility for existing users of the model.

Keywords: lexicography, linked data, ontologies, standardisation

1. Introduction

The OntoLex-Lemon Model (McCrae et al., 2017) has emerged as the de-facto standard for modelling dictionaries, lexical resources, terminologies and other resources as Linked Data. The official specification of the model was released in May 2016, almost 10 years ago at the time of writing, as a result of intense discussions of the W3C Ontology-Lexica Community[1] on "Lexicon Models for Ontologies". The model itself drew inspiration from previous models such as LMF (Francopoulo et al., 2006), LexOnto (Cimiano et al., 2007), LIR (Montiel-Ponsoda et al., 2011) and lemon (McCrae et al., 2011).

Since the emergence of the specification in 2016, the OntoLex-Lemon model has been widely used and adopted. As a result of this extensive usage, several gaps have been identified and new requirements have emerged. This paper summarises the key requirements that have been identified via community discussions. Current key limitations that have been identified include the inability of the current model to handle complex dictionary structures including nested entries, senses that can not be directly linked to one ontology element, as well as problems regarding the representation of retrodigitised dictionaries.

Building on top of this legacy, the W3C Ontology-Lexicon group is developing a new core module for the OntoLex-Lemon model that maintains backwards compatibility with existing resources published using OntoLex. This transition is driven by a community-led effort to bridge existing gaps, such as the need for more flexible dictionary structures and the representation of senses that do not align perfectly with ontological entities. It is important to note that the updates described in this paper, including the structural refinements and the integration of specialised modules into the core, represent a work-in-progress. These proposed changes are currently being evaluated through iterative discussions and remain subject to final approval by the OntoLex Community Group to ensure they meet the diverse needs of the Linguistic Linked Open Data community. The goal of this paper is twofold: to offer a comprehensive overview of the background and motivation of the new updated model and to serve as a reference to stimulate discussion in the Ontolex community.

The rest of this paper is structured as follows: Section 2 provides background on the existing OntoLex-Lemon model and its current ecosystem of modules. Section 3 details the community-driven methodology used to gather requirements and identify gaps in the current specification. Section 4 presents a comprehensive analysis of the fifteen key requirements identified during the consultation process, ranging from linking senses to forms to the representation of data quality. Section 5 describes the implementation of the proposed new OntoLex core, including the integration of lexicographic el-

[1] https://www.w3.org/community/ontolex/

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 69–79
12 May 2026.

ements and structural refinements. Section 6 discusses the evolution of the model in comparison to other standards. Finally, Section 7 concludes the paper with a summary of benefits for dictionary creators and an outlook into next steps for adoption.

2. Background: The OntoLex-Lemon Model

As mentioned in the introduction, the OntoLex-Lemon model (McCrae et al., 2017) has successfully established itself as the de facto standard for publishing lexical data as linked data. Its primary objective is to bridge the gap between ontologies and dictionaries by providing a formal framework for representing lexical entries, their forms, and their corresponding meanings. The model is architected around a core module that defines the fundamental relationships between a *Lexical Entry*, its morphological *Form*s, and its *Lexical Sense*s, which are linked to ontology entities in a process called "semantic by reference." This core is further supported by a robust ecosystem of modules. In addition to the core module, the proposed update of which is the focus of this paper, the OntoLex-Lemon was released including the following modules:

Syntax and Semantics (SynSem): This module manages the mapping between syntactic frames and semantic arguments.

Decomposition (Decomp): This module provides the structure for modelling multi-word expressions and the decomposition of lexical entries into their constituent components.

Variation and Alignment (VarTrans): This module is used to describe lexical and semantic relations, such as synonymy or translations between different languages.

Metadata (Lime): This module provides a vocabulary for expressing the linguistic metadata of a dataset, such as its language coverage and the density of links (Fiorelli et al., 2015).

After the initial 2016 release of the Ontolex-Lemon model, the following modules have been developed and are released or nearing completion:

Lexicography (lexicog): This module was introduced to address the complexities of dictionary structures, specifically handling the nesting of entries and the specific ordering requirements of senses and lexical entries (Bosque-Gil et al., 2017).

Frequency, Attestation and Corpus (FrAC): Developed to provide mechanisms for linking lexical data to corpus observations, this module supports the inclusion of frequency data and citation mechanisms for corpus-based lexicography (Chiarcos et al., 2022a).

Morphology (morph): This module provides a more granular framework for representing the internal structure of words and complex morphological patterns (Klimek et al., 2019; Chiarcos et al., 2022c,b).

In addition, the current OntoLex-Lemon ecosystem relies on LexInfo (Cimiano et al., 2011) to provide the necessary data categories, such as part-of-speech values (e.g., `lexinfo:noun`), which are typically mapped to standardised URIs. It is not a requirement for datasets to use LexInfo alongside OntoLex, and some implementations, such as the Global WordNet Association formats (McCrae et al., 2021; Bond et al., 2016) use different models. The LexInfo ontology is created as an open-source repository, where the main elements, such as catalogues of part-of-speech values, are editable as CSV files on the GitHub repository. This allows the community to manage and propose changes to extend the set of categories in this model easily.

The adoption of OntoLex-Lemon has moved beyond academic frameworks to become the backbone of several large-scale, widely-used lexical resources. Notably, **Wikidata** has integrated the core classes of the model into its Wikibase ontology (Lindemann et al., 2023) to represent its Lexeme entity type, making it perhaps the largest existing deployment of OntoLex-Lemon with millions of entries across hundreds of languages. Similarly, **BabelNet** (Navigli and Ponzetto, 2012) utilised the model to publish its multilingual semantic network as linked data (Ehrmann et al., 2014), effectively bridging the gap between encyclopedic and lexicographic knowledge. In the domain of computational lexicons, the **Open English WordNet** (McCrae et al., 2019, 2020b,a) and various initiatives within the **Open Multilingual Wordnet** (Bond and Foster, 2013) ecosystem have adopted the model to enrich traditional synset-based structures with granular morphological and syntactic descriptions.

Beyond general-purpose lexicons, OntoLex-Lemon has seen significant application in the field of terminology. There is an ongoing shift from traditional XML-based standards, such as **TBX** (Term Base eXchange), towards RDF representations to improve interoperability. This is exemplified by the **IATE** (Interactive Terminology for Europe) database, where research has focused on converting its complex terminological entries into OntoLex-Lemon to allow for better integration with other Linguistic Linked Open Data (LLOD) resources (Martín-Chozas et al., 2025; Ibarbia et al., 2025). These applications demonstrate the model's versatility in supporting both the "concept-centric" view of terminology and the "lemma-centric" view

of traditional lexicography.

3. Methodology

The development of the new OntoLex core followed a community-driven approach aimed at identifying the gaps in the existing model and gathering evolving requirements from various stakeholders. The process began with an open call for requirements distributed via the W3C Ontology-Lexicon Community Group mailing list. This initial consultation allowed the community to highlight specific limitations encountered during the implementation of the original model in diverse projects.

To ensure a comprehensive analysis, three specialised subgroups were formed to focus on key areas of lexical representation:

1. **Lexicography:** Focusing on the requirements of professional and historical dictionaries, specifically regarding entry nesting and the representation of complex senses.

2. **Terminology:** Addressing the needs of terminological resources and the alignment with standards such as TBX and IATE.

3. **Relation to other models:** Investigating the interoperability and alignment between OntoLex-Lemon and emerging standards, most notably the ISO/TC 37 LMF standard, the *DMLex* (Data Model for Lexicography) and TEI Lex-0.

The findings and requirements from these subgroups were systematically reported and aggregated. These requirements were then subjected to rigorous discussion during a series of regular teleconferences, where members of the community evaluated the proposed changes. This iterative process ensured that the resulting updates to the core model were both technically sound and representative of the needs of the broader Linguistic Linked Open Data (LLOD) community.

4. Requirement Analysis

The main requirements that were obtained from the community consultation procedure are summarised along with the actions proposed in Table 1 and are described in more detail as follows:

Sec.	Requirement	Status
1	Linking Senses to Forms	*Import*
2	Multiple POS Values	*New Modelling*
3	Usage Examples	*Import*
4	Diachronic/Diatopic Links	*No Change*
5	Ordering	*Import*
6	Senses w/o Ontologies	*Axiomatic Change*
7	Definitions	*New Modelling*
8	Cross-references	*No Change*
9	Literal POS Values	*New Modelling*
10	Usage Notes	*Axiomatic Change*
11	Sources	*No Change*
12	Reliability and Status	*New Modelling*
13	POS Property	*New Modelling*
14	Inflected Form	*New Modelling*
15	Module Integration	*Partial Move*

Table 1: Summary of the requirements and the proposed solution. Solutions involved either introducing new modelling, importing modelling from a module, changing the axioms of existing concepts, partial move or no changes.

4.1. Linking Senses to Forms

The first requirement addresses the necessity of associating a lexical sense with a specific grammatical form or colligation of a lexical entry rather than the entry as a whole. While this need is present in general computational lexicons, it is particularly prevalent in lexicographic resources where certain meanings are restricted to specific inflections, such as the plural form “airs” signifying a condescending manner or “games” in specific athletic contexts. Historically, this has been addressed in the OntoLex Lexicography (`lexicog`) module through the `FormRestriction` class, which allows for the explicit narrowing of a sense to a particular form. Current discussions for the new OntoLex core centre on whether to migrate this functionality into the core module to provide a more direct link, similar to the “subject lexeme form” property used in Wikidata or to continue relying on property-based restrictions. This requirement is fundamental for accurately representing “pluralia tantum” or senses tied to suppletive forms (e.g., *cow* vs *cattle*) where the semantic value is inseparable from the morphological realization.

4.2. Entries with Multiple Part-of-Speech Values

This requirement was identified by multiple initiatives and addresses a structural limitation in current lexical models where a single entry is often restricted to a single part-of-speech (POS). In traditional and retrodigitised dictionaries, it is common for one headword to encompass multiple grammat-

ical roles[2]; forcing the creation of separate lexical entries for each POS creates a disparity with the original source and complicates the modelling of languages like Basque, where nominals may function as both nouns and adjectives with identical morphological behaviour. However, interaction with other modules around syntax and morphology requires that a lexical entry has a single part-of-speech value, as the morphology or frame semantics of a word are usually different if it is a member of a different part of speech. To resolve this, the proposed OntoLex core architecture will introduce an `Entry` superclass, previously proposed as part of the OntoLex lexicography module, that can act as a container for multiple `LexicalEntry` components or support more generalised grammatical categories. This solution effectively bridges the gap between the computational need for strict POS tagging and the lexicographic reality of multi-functional headwords, while also clarifying the distinction between a high-level dictionary `Entry` and a specific `LexicalEntry`.

4.3. Usage Examples

Lexicographic models have a fundamental need to include usage examples, which are ubiquitous in both modern and legacy, retrodigitised lexicographic resources. While current implementations often attach examples exclusively to a specific sense, the discussion for the new OntoLex core has brought up the need for a more flexible approach where examples can be associated at both the entry and sense levels. Furthermore, this model supports many-to-many relationships, enabling a single example to illustrate multiple senses or even different entries simultaneously through linking properties similar to Wikidata's "subject sense"[3].

4.4. Diachronic and Diatopic Links

A requirement on the necessity of representing regional (diatopic) and historical (diachronic) variations, such as the differing definitions of *fanny* in UK versus US English, was raised. While this is a common feature in lexicographic resources, the consensus, which has arisen for discussion of the issues through teleconferences and on GitHub, is that the existing OntoLex-Lemon model already provides the necessary mechanisms for this[4].

4.5. Sense Ordering and Lexical Entry Ordering

Many dictionaries and lexical resources order senses by frequency, historical precedence, or other criteria. An important requirement for the OntoLex-Lemon model is thus to allow for ordering senses and lexical entries[5], a feature which was in large part implemented in the `lexicog` module. The community has proposed to move this feature from the `lexicog` module into the core module.

4.6. Senses without Ontologies

"Semantics by reference" was raised as a limitation in traditional and retrodigitised lexicography, where a single dictionary sense may not correspond to a single, clearly defined ontological concept[6]. To resolve this, the proposed evolution of the new OntoLex core involves relaxing the strict modelling constraints, specifically by removing the axiom that requires every `LexicalSense` to have exactly one reference to an `rdfs:Resource`. In particular, it has been suggested that the following axiom could be removed from the core model:

$$\text{LexicalSense} \sqsubseteq 1\ \text{reference.Resource}$$

4.7. Definitions

A need for more robust representation of definitions in lexicographic and terminological resources has been identified. The current OntoLex-Lemon specifications only give explicit guidance in cases where a dictionary sense corresponds to a lexical concept (the latter being a kind of `skos:Concept`): in which case, the use of `skos:definition` to link a lexical concept to a gloss is proposed. However, as we saw in the discussion of sense by reference in Section 4.6, this may not always be the case. Furthermore, a single string literal is often insufficient to capture complex metadata such as definition references, provenance, or internal notes. To resolve this, the new OntoLex core module proposes the use of reified definitions, treating a definition as a resource rather than a simple literal . This allows the textual content to be stored in an `rdf:value` property while supporting additional properties for source citations or semantic links, ensuring that the model can handle the high granularity required for

[2] One obvious example here is the word *youth* which in many languages is frequently listed both as noun and adjective under one common dictionary entry, e.g., this is often the case for Romance languages: *giovane*/*jovem*/*joven*.

[3] https://www.wikidata.org/wiki/Property:P6072

[4] However, there is a necessity for better documentation here to help users understand how they can do this via e.g., language tags.

[5] This might be as simple as a metadata statement which gives the kind of ordering which has been adopted for the dictionary.

[6] For instance, this is very clearly the case with conjunctions such as *that*, but it also causes problems for retrodigitised or philological dictionaries where what is listed as a single sense might not correspond to a single neatly defined concept.

professional lexicography and large-scale terminological databases such as IATE[7].

4.8. Cross-references

The community has raised the need to cross-reference other entries in a dictionary entry. This is particularly the case for entries that primarily serve to point to other headwords. The current consensus suggests that these relationships can be effectively modelled using existing properties like `rdfs:seeAlso` or by extending LexInfo with specific lexicographic properties together with better documentation, rather than requiring structural changes to the OntoLex core.

4.9. Literal Part-of-Speech Values

This requirement highlights a critical need in the digitization of historical and printed dictionaries: the ability to preserve the exact wording or visual representation of grammatical information as it appears in the original source. While computational models typically map parts of speech to standardised categories (such as `lexinfo:noun`), retrodigitised resources often require the retention of the specific string used by the original lexicographer, such as "Substantiv" or "n. f.". To address this requirement, the new OntoLex core will introduce a property like `ontolex:partOfSpeechString`, which allows for a literal representation of grammatical data alongside the standardised URI. However, it is unclear if this need is general enough to justify an extension to the core model or whether this should be handled by introducing a specific property into the LexInfo model.

4.10. Usage Notes

This requirement, which was obtained from multiple sources, addresses the necessity of including usage notes, recommendations, and domain-specific data, which are essential components of authoritative terminology records and traditional dictionaries. Such notes often exceed simple literal descriptions, requiring a combination of textual recommendations and links to external resources or provenance information via standard vocabularies such as PROV-O (the Provenance Ontology)[8]. Current modelling in OntoLex-Lemon is limited by the fact that the `ontolex:usage` property is restricted to the domain of `LexicalSense`. To provide the flexibility required for professional lexicography and terminology, such as applying "archaic" or "dialectal" labels to specific forms, morphemes, or entire entries, the proposed update for OntoLex involves removing these domain restrictions. This allows `ontolex:usage` to serve as a versatile property for reified usage nodes that can capture complex metadata, ensuring that language professionals can model nuanced usage recommendations across all levels of a lexical resource.

4.11. Sources

One requirement underscores the necessity of maintaining traceability across lexical and terminological resources, particularly when automated processes or multiple contributors are involved. OntoLex will emphasise distinguishing between "Original Sources" (such as corpora or specific individuals) and "Intermediate Sources" (such as information providers like IATE). While existing modules, such as FraC, provide citation mechanisms for corpus observations, they do not fully cover the metadata needs of notes and definitions. The current consensus suggests that PROV-O is sufficient for this purpose, provided that the elements, such as a `ConceptDefinition`, are also typed as `prov:Entity`.

4.12. Reliability and Record Status

Two requirements address the qualitative metadata of an entry, specifically its *Reliability* and *Record Status*. *Reliability* refers to the confidence rating terminologists assign to a term, often following a standardised system like the IATE four-star scale. While `lexinfo:confidence` currently exists, there is a proposed move toward a more standardised `ontolex:confidence` property, potentially supported by a "proxy" module for catalogues of values to avoid overloading the core ontology. The closely related requirement on *record status* indicates whether an entry is "finalized", "provisional", or "superseded". Although some aspects of status, such as being "superseded", can be handled via `dc:isReplacedBy` or PROV-O, the community is exploring a unified categorization that aligns reliability and status to distinguish between validated data and data that is work-in-progress.

4.13. Part-of-Speech Property

This requirement addresses the absence of a core part-of-speech property in the current OntoLex model, a feature that is already present in related standards such as DMLex. To enhance the model's utility for defining lexical entries, there is a proposal to migrate the widely-used `lexinfo:partOfSpeech` property directly into the OntoLex core as `ontolex:partOfSpeech`, potentially maintaining compatibility through an `owl:equivalentProperty` axiom. This move would support the introduction of formal axioms,

[7] https://iate.europa.eu/home
[8] https://www.w3.org/TR/prov-o/

e.g., requiring a `LexicalEntry` to have exactly one part-of-speech, and opens the door to incorporating standardised values like Universal POS Tags to improve cross-linguistic interoperability.

4.14. Inflected Form

This requirement regards the necessity of introducing a specific `inflectedForm` property to distinguish morphological inflections from other non-lemma forms. While the current OntoLex model provides `ontolex:otherForm`, this property is often viewed as having a broader interpretation that encompasses various types of non-canonical forms.

Take, for example, the Old English verb *bregdan* meaning, among other things, 'to move quickly' and 'to pull'. This lemma form (the infinitive) has a variant form *brægdan*. It also has inflected forms such as *bregdeþ* (third person singular indicative present) and *brægd* (third person singular indicative preterite). Interestingly, both of these two forms also have variants that occupy the same cell in the inflectional table for the verb (i.e., *brēt*, *brīt* for the former and *brǣd* for the latter). The definition of form in the current guidelines is as follows[9]:

> A form represents one grammatical realization of a lexical entry.

However in our example these forms encode both separate grammatical *and* phonetic variations (the result of different processes) as well as different corpus distributions. It isn't clear how we could model this properly, taking into consideration the correspondences between grammatical, phonetic and other properties in the case of different forms, using the current guidelines.

Among other things, the discussion for the new version weighs the benefit of adding a dedicated `inflectedForm` subproperty (something which could assist in clarifying situations such as those described in the preceding paragraph) against the potential increase in modelling complexity. A key consideration for this requirement is identifying specific use cases where a formal distinction between inflections and other lexical variations is essential for computational or lexicographic accuracy.

4.15. Integration of Modules and the Evolution of Lexicography Module

The original release of the OntoLex-Lemon specification introduced a modular architecture where a central core was accompanied by four initial modules (SynSem, Decomp, VarTrans, and Lime). Subsequent developments to the ecosystem, such as the *lexicog* and *FrAC* modules, were released as independent modules with a separate documentation to address specialised needs. However, the community consultation for the new OntoLex core module revealed that the current fragmentation makes it hard for newcomers to understand the model, leading to a requirement for a more integrated core.

The main focus here is the **lexicog** module, which was originally designed to handle the structural complexities of lexicography. To determine how key features of the **lexicog** module should be integrated into the core module, three distinct strategies were considered:

Option 1: Full Absorption: This approach would involve migrating all classes and properties from the `lexicog` namespace to the `ontolex` namespace, effectively merging the two specifications into a single document.

Option 2: Lexicog as a Core Module: Following this approach, *lexicog* would be maintained as a distinct section within the main specification, but all elements would retain their original URL in the `lexicog` namespace to maintain backward compatibility.

Option 3: Partial Move: This strategy involves a selective promotion of elements. High-utility components currently in the module would move to the `ontolex` core, while highly specialised lexicographic structures would remain in a separate module.

The community has expressed a strong preference for **Option 3 (Partial Move)**. This approach allows for an extended, more powerful core by promoting elements that have proven universally useful, such as `UsageExample` and `FormRestriction`, while keeping the core lean by leaving niche lexicographic structures (like complex entry nesting) within a dedicated module. This ensures that the model remains accessible to general users while providing the necessary depth for professional lexicographers.

5. Implementation

The implementation of the new OntoLex Core Module involves an iterative improvement of the model in a fully backwards-compatible manner to ensure that the model satisfies the requirements discovered in over ten years of deployment of the model across a wide range of applications. These changes are depicted in Figure 1 and are proposed changes subject to approval by the community.

[9] https://www.w3.org/2016/05/ontolex/#forms

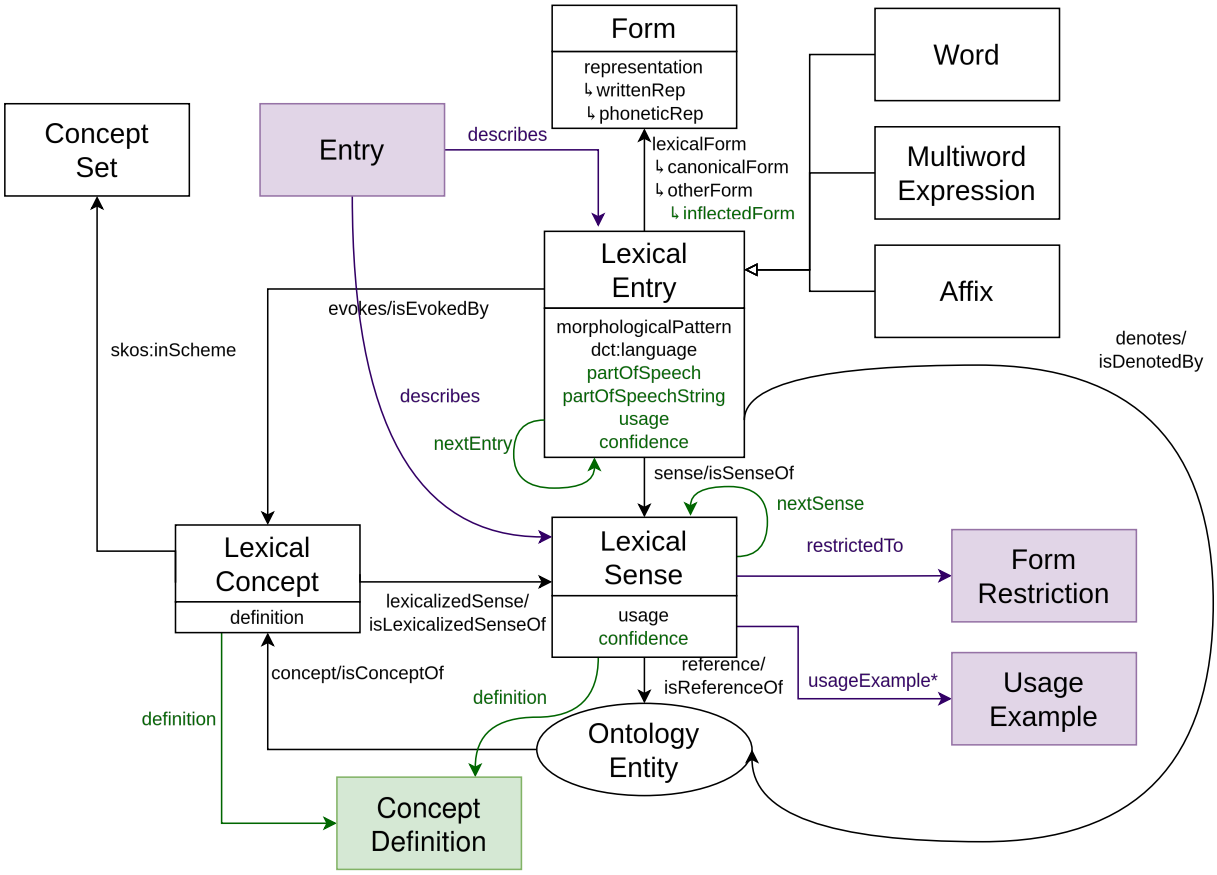

Figure 1: The proposed updated core diagram of OntoLex-Lemon. The green elements indicate new modelling to be added to the core model. The purple elements indicate modelling from modules to be promoted into the core.

5.1. Integration of Lexicographic Elements

Following the decision to adopt a hybrid integration strategy, several key classes and properties have been promoted from the Lexicography module into the OntoLex Core Module. This move enables the representation of complex dictionary structures without requiring the overhead of external modules.

Entry The introduction of the `Entry` class as a broader concept alongside the existing `LexicalEntry` allows for a more flexible grouping of lexical data, particularly for retrodigitised dictionaries where a single headword may describe multiple `LexicalEntry` instances (e.g., across different parts of speech). The `LexicalEntry` class is kept with a narrower interpretation in order to maintain interoperability as well as backwards compatibility with other modules.

Form Restrictions These elements allow a `LexicalSense` to be explicitly linked to a specific `Form`. This is essential for modelling cases where a meaning is only valid for a specific inflection (e.g., the plural “airs”).

Usage Examples Formerly part of the lexicography module, this class has been moved into the core in order to allow for the reified representation of usage examples and citations. The original usage example property was called `usageExample`, however, we will rename this to just `example` to avoid confusion between property names and support the specification in line with W3C guidelines.

To ensure backwards compatibility, previous URLs will still exist, but will have axioms such as `owl:sameAs` or `owl:equivalentProperty` to their new URLs.

5.2. New Properties and Structural Refinements

To enhance the granularity of the model and its alignment with standards like DMLex, the following properties and classes have been proposed:

Linguistic Metadata: New properties, including `partOfSpeech` and `partOfSpeechString` allow for both URI-based (LexInfo) and literal-based grammatical tagging. The

`inflectedForm` property provides a direct way to identify non-canonical forms.

Ordering: To support the sequential nature of dictionaries, `nextEntry` and `nextSense` have been introduced.

Data Quality: A `confidence` property allows for the representation of uncertainty, which is frequent in automatically generated or OCR-derived lexicons.

Definitions: A new class, `ConceptDefinition`, has been introduced to allow definitions to be treated as first-class objects, enabling them to be shared across multiple senses or concepts.

As these changes are all additive, they will not affect the backwards compatibility of the model.

5.3. Axiomatic and Domain Changes

In addition to new entities, the new core relaxes several constraints of the original model. The domain of the `usage` property has been extended to include `LexicalEntry`, allowing for broader labels (e.g., register or dialect) to be applied at the entry level. Most significantly, the strict axiom on `LexicalSense`, which previously required a reference to an external ontology entity, has been removed. This allows for "non-ontological" senses, enabling the publication of dictionaries where meanings are expressed solely through definitions or examples without the need for a formal URI reference. As these changes only relax existing constraints, this is a non-breaking change. Existing data that does use references remains perfectly valid under the new, more permissive logic.

6. Discussion

6.1. Comparison with Existing Standards

The development of the new OntoLex core has been heavily informed by a comparative analysis with other prominent lexical and lexicographic standards, ensuring that the model remains a robust bridge between the Semantic Web and traditional linguistics.

- **DMLex (Data Model for Lexicography):** As an OASIS standard, *DMLex* (Filip et al., 2024) provides a functional, abstract model for dictionary data. While DMLex is focused on the data structures required for dictionary management systems, the new OntoLex core acts as its realization in the Linked Data space. The introduction of more general entries, specific part-of-speech properties and definitions brings OntoLex closer to the DMLex core model, allowing resources to be more easily converted between these two formats. Noticeably, DMLex has a module for the modelling of etymology and this spurs the development of a potential new module for OntoLex to enable further integration of these two models.
- **TEI-Lex0:** Traditionally, the TEI (Text Encoding Initiative) guidelines, specifically the *TEI-Lex0* customization, have been the standard for the digital representation of dictionaries as documents. While TEI-Lex0 excels at capturing the layout and textual details of a source, OntoLex-Lemon focuses on the data's semantic interoperability. The new core module will reduce the friction involved in converting TEI-encoded resources into RDF by promoting features like `UsageExample` and `partOfSpeechString`, which are more directly compatible with the string-heavy metadata found in TEI headers.
- **LMF (Lexical Markup Framework):** Compared to the closed ISO (LMF) standard[10] (Romary et al., 2019), OntoLex maintains an open web-centric approach. While LMF utilises a data-category-based model that often relies on complex XML structures[11], OntoLex leverages the inherent graph nature of RDF to provide more flexible linking between senses and specific forms, a requirement that has historically been difficult to model succinctly in LMF without the need for ad hoc extensions.

6.2. Evolution of the Specification

The development of this new core module reflects a shift in the community's priorities from ontology-based modelling, as exemplified by the "semantics by reference" principle, toward "lexicographer-friendly" modelling. The original specification was highly successful in linking lexicons to formal ontologies, but it created a high entry barrier for publishers of legacy dictionaries who did not have (or need) a corresponding OWL ontology for every lexical sense. The process of evolving this specification

[10] Originally published as a single standard, ISO 24613, LMF was subsequently released as a multipart standard, including a core model ISO 24613-1:2024; a machine-readable dictionary module ISO 24613-2:2020; an etymological extension, ISO 24613-3:2021, a TEI serialisation, ISO 24613-4:2021; another serialisation in the obscure Lexical base exchange (LBX) format ISO 24613-5:2022; and finally a Syntax-Semantics module, ISO 24613-6:2024. While the original LMF standard was adopted in a number of projects and for encoding several lexicons, we are not aware of any lexicons that have been developed using the new version of LMF.

[11] The 'official' XML serialisation of LMF is in TEI-XML.

through the formation of specialised subgroups allowed us to collect feedback from a wide range of users of the models. By identifying which elements were widely used, such as reified definitions, and integrating the lexicography module more tightly with the core, we have created a model that is both more powerful and easier to adopt. This evolution demonstrates how a standard can remain viable and balance the formal world of ontologies with the needs of lexicographers. The community is discussing using SHACL (Knublauch and Kontokostas, 2017) to further improve the validation of implementations of the OntoLex-Lemon model.

7. Conclusion

The development of the new core module will further extend on the success of OntoLex as a bridge between formal ontologies and the practical needs of lexicographers. Our community-driven approach allows us to address long-standing limitations such as the handling of complex dictionary structures, multi-functional headwords, and the necessity of preserving original source metadata. Key refinements in this module promote highly useful modelling into the core, while more flexible semantics will further increase the number of use cases that OntoLex can satisfy. However, mostly, new modelling, such as for part-of-speech strings, inflected forms, and reified definitions ensure that OntoLex can capture the nuanced data required for authoritative terminological records and retrodigitised resources. Ultimately, the new OntoLex core balances the formal rigour of the Semantic Web with a "lexicographer-friendly" architecture.

This new core module will be published in accordance with the public review procedures of the OntoLex W3C Community Group and is subject to approval and comments by the full community. This acts as a basis to allow OntoLex to develop further, in particular through the development of new modules and the completion of modules to enable OntoLex to be a flexible and forward-looking standard.

Acknowledgements

This publication is based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

Bibliographical References

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1352–1362. The Association for Computer Linguistics.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the Global WordNet Conference 2016*.

Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. Towards a module for lexicography in OntoLex. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017*, volume 1899 of *CEUR Workshop Proceedings*, pages 74–84. CEUR-WS.org.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Unifying morphology resources with OntoLex-morph. a case study in German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille, France. European Language Resources Association.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational morphology with OntoLex-morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86, Marseille, France. European Language Resources Association.

P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. 2007. LexOnto: A model for ontology lexicons for ontology-based NLP. In *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07*.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 401–408. European Language Resources Association (ELRA).

David Filip, Miloš Jakubíček, Simon Krek, John McCrae, and Michal Měchura. 2024. Data Model for Lexicography (DMLex) Version 1.0. OASIS committee specification draft 02, OASIS.

Manuel Fiorelli, Armando Stellato, John P. McCrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. LIME: the metadata module for OntoLex. In *Proceedings of 12th Extended Semantic Web Conference*.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Paula Diez Ibarbia, Patricia Martín-Chozas, and Elena Montiel-Ponsoda. 2025. Bringing IATE into the semantic web family. In *Proceedings of the 5th Conference on Language, Data and Knowledge: The 5th OntoLex Workshop*, pages 12–17.

Bettina Klimek, John McCrae, Maxim Ionov, James K. Tauber, Christian Chiarcos, Julia Bosque-Gil, and Paul Buitelaar. 2019. The OntoLex-lemon morphology module. In *Proceedings of the Sixth Biennial Conference on Electronic Lexicography (eLex 2019)*, Sintra, Portugal.

Holger Knublauch and Dimitris Kontokostas. 2017. Shapes Constraint Language (SHACL). W3C recommendation, W3C.

David Lindemann, Sina Ahmadi, Anas Fahad Khan, Francesco Mambrini, Federica Iurescia, and Marco Carlo Passarotti. 2023. When OntoLex meets Wikibase: Remodeling use cases. In *Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 13, 2023*, volume 3640 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Patricia Martín-Chozas, Thierry Declerck, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. 2025. Representing terminological data in the semantic web: A proposal based on OntoLex-lemon. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 31(2):171–207.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The GlobalWordNet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

John P. McCrae, Ewa Rudnicka, and Francis Bond. 2020a. English WordNet: A new open-source WordNet for English. *K Lexical News*, (28):37–44.

John P. McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020b. English WordNet 2020: Improving and extending a wordnet for english using an open-source methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*, pages 14–19.

Elena Montiel-Ponsoda, Guadalupe Aguado de Cea, Asunción Gómez-Pérez, and Wim Peters. 2011. Enriching ontologies with multilingual information. *Nat. Lang. Eng.*, 17(3):283–309.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. LMF reloaded. In *Proceedings of the 13th Conference of the Asian Association for Lexicography*.

Linked Open Data for West Nilotic Languages: The NILOMORPH project

Matteo Pellegrini, Matthew Baerman, Oliver Bond

University of Surrey
Surrey Morphology Group
{matteo.pellegrini,m.baerman,o.bond}@surrey.ac.uk

Abstract

In this paper, we present the NILOMORPH project, that aims at describing the complex non-concatenative morphology of West Nilotic languages and reconstructing the dynamics of its evolution from a more straightforward concatenative system. The project adopts techniques from several methodologies and draws on many kinds of data displaying different formats, tagsets and conventions. Data are also multilingual, documenting different West Nilotic varieties, and multimodal, including also audio and video recordings. This makes the process of integration of these data particularly challenging. We first describe how the data can be converted to standard formats such as CLDF and Paralex, to achieve interoperability between resources of the same kind. We then discuss how they can be modelled as Linguistic Linked Open Data in the Resource Description Framework, reusing already existing vocabularies and defining new classes and properties to meet the needs of the project, to also achieve interoperability between resources of different kinds.

Keywords: West Nilotic, morphology, Linguistic Linked Open Data

1. Introduction and Motivation

The West Nilotic languages are spoken in an area that includes South Sudan, south-western Ethiopia, the north-east of the Democratic Republic of Congo, northern Uganda and south-western Kenya. A remarkable feature of some of those languages is that they mark multiple morphological distinctions simultaneously through modifications of phonological features of the stem vowel. For instance, in the Nuer lexeme for 'rain', case and number distinctions are signalled by contrasts in tone, length, phonation type and height, as in NOM.SG /bɛ̰̀ː/ vs. GEN.SG /bêl/ vs. NOM.PL /bɛ̌ː/ vs.GEN.PL /bɛ̂ː/ (Baerman and Monich, 2021). The NILOMORPH project aims to reconstruct in detail how such a complex and cross-linguistically rare non-concatenative system might have emerged from the more straightforward concatenative one still preserved in other West Nilotic varieties. To do that, it applies techniques taken from many different methodologies, including field linguistics, acoustic analysis, experimental linguistics, computational simulations, typology, and the historical-comparative method.

Such methodological variety poses the challenge of how to deal with a wide range of data types (Section 2): lexical resources such as lexicons of inflected forms; textual resources, such as raw texts (sometimes with translations), glossed examples, or more structured tagged corpora; as well as material derived from language experiments. Data are necessarily multilingual (to allow for the application of the comparative method), and multimodal (since audio recordings and visual documentation may provide useful information unavailable elsewhere). It is therefore crucial that these multifarious data types be interoperable with each other, allowing for the extraction of information from different sources in a unified fashion.

A natural solution to this challenge is offered by the framework of Linguistic Linked Open Data (LLOD; Cimiano et al., 2020), that aims to make data FAIR – Findable, Accessible, Interoperable, Reusable (Wilkinson et al., 2016) by leveraging Semantic Web technologies. The backbone of this enterprise is the Resource Description Framework (RDF) data model (Lassila and Swick, 1999), where all items are treated as resources with their own Unique Resource Identifier (URI), information is expressed through triples that connect a subject (a resource) to an object (a resource, or some data about it) through a property (itself a resource), and the relation between items is reflected in a hierarchical structure of sub-classes and sub-properties. RDF data can be retrieved with the SPARQL Protocol and RDF Query Language (Prud'hommeaux and Seaborne, 2008), that allows for queries that efficiently extract variegated information from different sources. The use of standard vocabularies and ontologies also makes it possible to have interoperability between datasets created by different people for different projects.

To date, RDF LLOD have been released for an increasing number of languages, as documented in the LLOD cloud.[1] Of particular note are a handful of larger-scale projects that have been launched for the integration of resources of different kinds available for a single language, starting with LiLa for Latin (Passarotti et al., 2020), later followed by LiITA for Italian (Litta et al., 2024) and MOLOR for Old Irish (Fransen et al., 2024). Efforts have been made also to achieve interoperability between languages of the same family – e.g. Romance in the ALMA project (Tittel,

[1] https://linguistic-lod.org/.

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 80–86
12 May 2026.

2023) – and between data in different modalities (Menke et al., 2013).

We illustrate here our proposal for how to model the variegated data of the NILOMORPH project as RDF LLOD. Doing so will contribute to this framework by increasing its coverage to include several under-resourced languages that were previously not represented, including Nuer, Dinka, Shilluk and Thok Reel. It will also provide the community with a discussion of the specific characteristics of the project data, including multilinguality, multimodality, and coding of cognacy between elements in different related languages.

The paper is structured as follows. Section 2 details and exemplifies the different kinds of data. Section 3 presents the model that we propose to handle the data, discussing existing standards and vocabularies relevant to our needs. Section 4 concludes and highlights the next steps.

2. The Data

Firstly, there are lexical resources. These include both traditional bilingual dictionaries available from earlier documentation efforts, and more structured resources produced more recently. An example of the former is Kiggen's (1948) Nuer-English dictionary, with information on lexical category and other morphosyntactically relevant features (e.g. transitivity-based inflection classes), giving English translation(s) and translated usage examples, and possibly additional notes and comments. Entries also list principal parts (the set of inflected forms from which other paradigm cells can be inferred; Stump and Finkel, 2013). All this information has undergone OCR and is available in machine-readable format as an Excel spreadsheet. This has been supplemented with other classifications relevant for the project goals, e.g. flagging stem-final consonants relevant for identifying suspected cognate entries. As an example of the latter type of resource, for Nuer there is also a more structured paradigmatic lexicon which provides a larger list of inflected forms in IPA transcription, and other pieces of information, such as paradigmatic patterns of vowel length and quality alternations, along with audio recordings of different forms in context (Bond et al., 2020). All this information feeds an interactive website, and it is also available as an Excel spreadsheet. For lexical resources, a crucial aim of the project is the development of a database that consist of several paradigmatic lexicons for different varieties of West Nilotic languages, and the identification of cognate lexical items across languages, to enable the application of the comparative method for the reconstruction of Proto-West Nilotic morphology.

Secondly, there are textual resources. These include both written material[2] and audio recordings of spontaneous speech (e.g. Remijsen et al., 2014-2024). In many cases, only the raw text is available, with no linguistic annotation. In some cases, we also have English translations of those texts. In other cases, richer linguistic information is provided. For written material, these may be in the form of interlinear glosses added to the text. For audio recordings, phonological transcription, morpheme segmentation, translation and comments can be provided in .eaf format using the ELAN software. For textual resources, the project aims at enriching the information provided on audio recordings with annotations in .textgrid format using the speech analysis software Praat, and at building tokenised, lemmatised, PoS-tagged and possibly morphologically analysed corpora from the texts available.

In addition, data may also consist of materials prepared for language experiments of different kinds, for instance, production and perception experiments to determine the exact phonetic realisation of sound contrasts in the languages under investigation. One example is Remijsen et al. (2022), that provides .wav audio files of forms elicited from speakers of the Bor dialect of Dinka to investigate contrasts of voice quality (modal vs. breathy) and tone (low vs. high), together with .textgrid files of annotations made with Praat.

The project will make this wealth of information publicly accessible[3] on the web in machine-readable open formats (e.g. converting Excel spreadsheets to .csv tables), and connect the different datasets with each other, and possibly with other datasets available on the web for other languages, using standard vocabularies and ontologies in the RDF data model, thus achieving the five stars in the grading system of open data recommended by Tim Berners-Lee.[4]

3. The Model

Standard formats have been proposed for many of the kinds of resources we are dealing with. The Cross Linguistic Data Formats (CLDF, Forkel et al., 2018) initiative offers standardised ways to represent in tabular format the data most often gathered while documenting and describing languages. To do that, it builds on the recommendations of the CSV on the Web W3C Working Group, namely the Model for Tabular Data and Metadata on the Web[5] and the Metadata Vocabulary for Tabular Data.[6] CLDF datasets

[2] E.g. the Nuer storybooks at https://www.nuerlexicon.com/books.php.

[3] Material protected by copyright will be standardised and converted to RDF but not released openly.

[4] https://www.w3.org/DesignIssues/LinkedData.html.

[5] https://www.w3.org/TR/tabular-data-model/.

[6] https://www.w3.org/TR/tabular-metadata/.

consist of several .csv files that refer to each other by means of unique identifiers, following the best practices of relational database management. This allows users to provide different pieces of information, avoiding redundancy. Different modules are introduced for specific kinds of dataset, namely wordlists, dictionaries, structure datasets (such as the World Atlas of Language Structures, Haspelmath et al., 2005), parallel texts in different languages, and corpora. Different components (tables) are defined to express information on different items, such as forms, lexical entries, senses, and languages. Different columns are defined to express different pieces of information on those items, e.g. the language, part-of-speech and headword of lexical entries.

Among the resources that can be represented as CLDF lexicons, there are many of the ones relevant to our project (such as dictionaries, parallel texts and corpora), but a very important one is missing, namely, paradigmatic lexicons that document inflected forms that appear in different cells of lexemes. The Paralex initiative[7] fills this gap, offering a standard format for such resources. To do that, it draws on many of the fundamental principles of CLDF, such as the use of multiple tables with a relational structure and the coding of metadata in a machine-readable format, although for the latter it relies on the frictionless[8] framework for data packages, rather than on CSVW. Consequently, different tables are defined for items on which information is provided, such as forms, lexemes, cells; and different columns are defined for the pieces of information that are provided, such as the orthographic and phonetic/phonological transcription, cell and lexeme of a form. The Paralex standard format is also well equipped for the coding of phonetic and phonological aspects that are crucial for the aims of NILOMORPH. Those can be expressed in a separate table whose lines contain segments and whose columns contain either the phonological features that define them or links to standardised repositories – such as CLTS (Anderson et al., 2018) and PHOIBLE (Moran and McCloy, 2019).

Adopting standard formats such as CLDF and Paralex introduces quite strict requirements that allow for a great degree of interoperability between resources of the same kind, e.g. between CLDF datasets of different languages but with the same module and components. Such interoperability is not only structural (pertaining to the data formats and languages), but also semantic (pertaining to the actual categories and values used). This makes it possible to develop tools that can be seamlessly applied to resources complying with the standards in order to perform fully comparable linguistic analysis, e.g. entropy-based measurements of predictability with the Qumin toolkit (Beniamine, 2018) on Paralex lexicons. Consequently, such standards have been increasingly adopted by the research communities working on these kinds of data.[9]

To achieve interoperability between resources of the same kind available for the different languages involved in NILOMORPH, and with other resources of that kind available for other languages, we start from the legacy data mentioned in Section 2. These come in various formats and use resource-specific conventions and tagsets. We convert them into the standard formats of CLDF (for traditional dictionaries and texts, glosses and additional annotations), and Paralex (for paradigmatic lexicons). In doing so, there is the potential to enrich the vocabularies of these formats to account for the complexity of West Nilotic data, e.g. the different phonological dimensions involved in vowel alternations.

However, *per se* this does not address another project requirement, i.e. interoperability between resources of different kinds, which may include ones that have been created for different purposes. RDF LLOD technology is the natural way to satisfy this requirement. Indeed, both CLDF and Paralex provide built-in ontologies that introduce RDF classes and properties for the tables and columns of the formats, and define them as sub-classes and sub-properties of resources in existing ontologies for the relevant domains. This allows seamless conversion to RDF LLOD, ensuring interoperability with a wider scope. For additional tables and properties defined for our data, it will be necessary to define additional mappings to RDF classes and properties, either by reusing existing vocabularies directly, or by extending the ontologies.

We now turn to the LLOD vocabularies of interest for the data of our project. Crucial to any work on language data are repositories of linguistic terminology, both general-purpose ones, such as GOLD (Farrar and Langendoen, 2003), and others for specific use cases, e.g. Lexinfo for lexical resources (Cimiano et al., 2011), or OLiA for annotations (Chiarcos and Sukhareva, 2015).

For lexical resources, currently the *de facto* standard is the OntoLex vocabulary (McCrae et al., 2017), which provides classes for lexical entries, their form, and meaning (senses and concepts), as well as properties that relate those items to each other. Additional modules are provided for more specific aspects, such as *decomp* for the decomposition of lexical entries, and *lime* for metadata of lexical resources. Other

[7] https://www.paralex-standard.org/.

[8] https://framework.frictionlessdata.io/.

[9] See the CLDF and Paralex communities on Zenodo (https://zenodo.org/communities/paralex, https://zenodo.org/communities/cldf-datasets, respectively) for a list of the datasets released in the two formats.

modules are also being developed, such as *morph* for morphological information (Chiarcos et al., 2022c) and *FrAC* for frequency and corpus attestations (Chiarcos et al., 2022a). Given NILOMORPH's focus, representations of morphological data using *morph* are particularly relevant, such as work by Chiarcos et al. (2022b) on German and Ionov and Rosner (2023) on Maltese.

For textual resources, the NLP Interchange Format (NIF, Hellmann et al., 2013) provides a way to unambiguously identify portions of texts on the web, while POWLA (Chiarcos, 2012) allows for the addition of separate layers on which different annotation levels can be operated on the same text. Furthermore, the Open Annotation vocabulary reifies annotations with their own class, introducing properties relating them to the annotated item, and to the content of the annotation. The latter vocabulary is of particular interest for this project because it supports annotation of different media types. This allows us to handle audio recordings and their annotation. For interlinear glossed text, the ligt framework (Chiarcos and Ionov, 2019) is explicitly designed to model this kind of data, and its tools are available for an automatic conversion from glossed examples in CLDF format (Ionov, 2025).

The Paralex ontology has been designed to allow for conversion to OntoLex compliant lexical resources, providing mappings to classes and properties of the OntoLex model and its modules where relevant, alongside categories from lexinfo and recommendations to reuse the Open Annotation model for the coding of morphological features, and mapping to other standard vocabularies such as the one of the Unimorph project (Kirov et al., 2018), via an OLiA annotation model (Chiarcos et al., 2020). The parallel release of data as both Paralex and OntoLex lexicons has already been tested on existing resources (Pellegrini et al., 2025). Consequently, its application to NILOMORPH data should not require any further modelling effort, except for resource-specific tables and columns introduced for the purposes of the project.

However, the CLDF ontology only defines mappings to the GOLD ontology for data.[10] Because of the status of OntoLex as a *de facto* standard for the release of lexical resources, at the stage of conversion to RDF we plan to also introduce an OntoLex-compliant modelling, defining entries, forms, senses and concepts as belonging to the corresponding OntoLex classes, and using OntoLex properties to relate them. Similarly, for textual resources we will supplement the classes and properties of the CLDF ontology with mappings to the vocabularies mentioned above for different pieces of information – NIF and POWLA for texts, ligt for glossed examples.

Furthermore, more elaborate resources do not lend themselves easily to release as either CLDF or Paralex datasets. For instance, this is the case of richer and more structured corpora that can be produced by tokenising and annotating the available texts. Such resources can be released as RDF LLOD directly, following best practices defined by Cimiano et al., (2020) and applied, for instance, to Latin corpora in the LiLa knowledge base (Mambrini and Passarotti, 2019).

Finally, once all resources are available in standardised formats and as RDF LLOD, they need to be connected with each other. Projects such as LiLa, LiITA and MOLOR follow a common strategy to achieve this, by connecting entries and lexical resources and tokens of textual resources to the respective lemma, which is defined as a sub-class of forms in the OntoLex vocabulary. Such a strategy can be applied also to our project, but it only covers the case where we have different resources for the same language. However, given the scope of the project, it is necessary to also reach interoperability between different West Nilotic languages. For this, cognacy relations between items will be identified. These can be coded using the cognates and cognate sets components of CLDF on the one hand, and through the lemonEty vocabulary for etymological information in OntoLex lexicons (Khan, 2018) as RDF LLOD on the other. Etymological information will play a pivotal role in allowing for interoperability between all the datasets relevant to the NILOMORPH project.

4. Conclusions and Future Work

We have presented here the NILOMORPH project, which aims to describe and reconstruct the complex non-concatenative morphology of the West Nilotic family, thus elaborating and creating data on under-resourced languages. We have described the multifarious data handled by the project and discussed how existing standards and ontologies are reused to model them as RDF LLOD, also showing how specific aspects of the data require us to enrich and extend these vocabularies.

We plan next to move on to other steps of the LLOD generation process as outlined by Cimiano et al. (2020), namely, i) data generation, ii) linking, iii) publication and iv) exploitation. We will i) convert the heterogeneous legacy formats into CLDF and Paralex datasets, and then convert them to RDF, ii) connect the different resources together through their citation forms and etymons and add links to other resources, iii) publish the data and metadata with open licences in

[10] For metadata, links to dcat and dcterms are also provided.

repositories that assign them a DOI, and provide access to them in a triplestore, also offering tabular and graph visualisations with lodview and lodlive,[11] and finally iv) develop tools to exploit all these interconnected resources. All these steps will ultimately be helpful to achieve the goals of the project concerning linguistic analysis of the languages involved and the reconstruction of the proto-West Nilotic system.

5. Bibliographical References

Cormac Anderson, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, 4(1):21–53.

Matthew Baerman and Irina Monich. 2021. Paradigmatic saturation in Nuer. *Language*, 97(3):e257–e275.

Sacha Beniamine. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. PhD Thesis, Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7).

Christian Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012, Proceedings*, pages 225–239. Springer, Dordrecht.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju. International Committee on Computational Linguistics.

Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2020. Annotation Interoperability for the Post-ISOCat Era. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5668–5677. European Language Resources Association, Marseille.

Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2022b. Unifying Morphology Resources with OntoLex-morph. A Case Study in German. In Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille. European Language Resources Association.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022c. Computational Morphology with OntoLex-Morph. In Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, and Maxim Ionov, editors, *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86, Marseille. European Language Resources Association.

Christian Chiarcos and Maxim Ionov. 2019. Ligt: An LLOD-native vocabulary for representing interlinear glossed text as RDF. In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019). LDK 2019, May 20–23, 2019, Leipzig, Germany*, page 3:1-3:15. Dagstuhl, Wadern.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic linked data*. Springer, Dordrecht.

[11] https://github.com/LodLive/LodView.

Scott Farrar and D. Terence Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT international*, 7(3):97–100.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR Lemma Bank: A New LLOD Resource for Old Irish. In Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, Patricia Martin Chozas, editors, *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 37–43, Torino. ELRA and ICCL.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using linked data. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lola Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web - ISWC 2013. 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings*, pages 98–113. Springer, DordrechtNLP.

Maxim Ionov. 2025. Ligt: Towards an Ecosystem for Managing Interlinear Glossed Texts with Linguistic Linked Data. In Mehwish Alam, Andon Tchechmedjiev, Jorge Gracia, Dagmar Gromann, Maria Pia di Buono, Johanna Monti, and Maxim Ionov, editors, *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 100–105. UniorPress, Napoli.

Maxim Ionov and Michael Rosner. 2023. Beyond Concatenative Morphology: Applying OntoLex-Morph to Maltese. In Sara Carvalho, Anas Fahd Khan, Ana Ostroški Anić, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch, and Ana Salgado, editors, *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 385–391, Vienna. NOVA CLUNL.

Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12):304.

Jan Kiggen. 1948. *Nuer-English Dictionary*. Missiehuis, Steyl bij Tegelen.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan.

Ora Lassila and Ralph R. Swick. 1999. Resource Description Framework (RDF) Model and Syntax Specification.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio, and Cristina Bosco. 2024. The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

Francesco Mambrini and Marco Passarotti. 2019. Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa *Proceedings of eLex 2017 conference*, pages 19–21.

Peter Menke, John Philip McCrae, and Philipp Cimiano. 2013. Releasing Multimodal Data as Linguistic Linked Open Data: An Experience Report. In Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P. McCrae *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other*

language data, pages 44–52, Pisa, Italy. Association for Computational Linguistics.

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(1):177–212.

Matteo Pellegrini, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2025. PrinParLat: a Lexicon of Principal Parts of Latin Verbs Linked to the LiLa Knowledge Base. *Language Resources and Evaluation*:1–41.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL Query Language for RDF.

Gregory T. Stump and Raphael A. Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge.

Sabine Tittel. 2023. Ceci n'est pas un dictionnaire. Adding and Extending Lexicographical Data of Medieval Romance Languages to and through a Multilingual Lexico-Ontological Project. In Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček, and Simon Krek *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Lexical Computing CZ sro, Brno.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E. Bourne. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

6. Language Resource References

Oliver Bond, Tatiana Reid, Irina Monich, and Matthew Baerman. 2020. Nuer Lexicon. www.nuerlexicon.com. Accessed 6 February 2026.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0.

Bert Remijsen, Mirella L. Blum, and Jon Pen de Ngong. 2022. A dataset on Voice Quality, Tone, and Vowel Quality in the Bor South dialect of Dinka.

Bert Remijsen, Otto Gwado Ayoker, and Maria Bocay Onak. 2014-2024. Collection of Shilluk narratives and songs.

A linguistic ontology for constructicography: the Research Constructicon and its ontology modules

Elodie Winckel, Peter Uhrig, Stephanie Evert
Friedrich-Alexander-Universität Erlangen-Nürnberg
Research Training Group *Dimensions of Constructional Space*
elodie.winckel@fau.de, peter.uhrig@fau.de, stephanie.evert@fau.de

Abstract

This paper introduces the Research Constructicon (RCxn), a project developed within the Research Training Group *Dimensions of Constructional Space*. The training group finances PhD projects in the framework of Construction Grammar (CxG), which views language as a network of form-meaning pairings. The RCxn is designed as a dynamic, community-driven resource that documents linguistic constructions while also capturing the research processes and findings associated with them. The project addresses three core dimensions: (1) the development of a modular ontology to represent constructions, their relationships, and the research surrounding them; (2) the implementation of a database populated by researchers' contributions; and (3) the creation of a web application to visualize and interact with the data. This paper focuses on our work to implement a rich ontology for the RCxn, which has to accommodate diverse research needs, from cross-linguistic comparisons to multimodal analyses, while ensuring flexibility and interoperability. We detail the modular design of the ontology, its alignment with semantic web standards (RDF/OWL), and the integration of existing ontologies (e.g., OLiA, FOAF). The RCxn's development is iterative, driven by feedback from our diverse group of PhD researchers.

Keywords: Constructicography, Semantic Web Technologies, Ontology

1. Introduction

This paper presents the Research Constructicon (hereafter: RCxn), a project funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Research Training Group *Dimensions of Constructional Space*. The latter is dedicated to research in Construction Grammar (CxG), a theoretical framework that views language as a network of form-meaning pairings referred to as "constructions". According to Construction Grammar, constructions range from words and morphemes to complex syntactic patterns and discourse structures. The RCxn project aims to create a dynamic, community-driven resource that not only documents linguistic constructions but also captures the research processes and findings associated with them.

The project was divided into three interconnected dimensions. First, we developed an ontology for all the terms that will allow us build a database that documents individual constructions, their interconnections, and the research conducted on them. Accordingly, the ontology needs to cover a wide range of aspects, from the internal features of constructions to their relationships with other constructions and the scientific discourse surrounding them. Second, we used the ontology to create an RDF database that currently includes contributions from 13 researchers, with plans to expand to 30 or more by the end of 2027. Finally, we developed a web application to visualize and interact with the database content, which is now operational and accessible (https://bdlweb.phil.uni-erlangen.de/RCxn/).

While these three dimensions (ontology, database, and web application) can be conceptualized as sequential steps, each one requiring the previous one to be implemented, their development has been iterative. The ontology continues to evolve as new needs emerge, influencing both the database structure and the web application's frontend.

In this paper, we focus on the first of these dimensions: the ontology that underpins the RCxn. We begin by outlining the theoretical motivations for adopting a network-based approach to constructionist research, emphasizing the advantages of Semantic Web Technologies. We then discuss the technical implementation of this ontology. Finally, we present the modular ontology we developed, where each module is tailored to address specific aspects of constructionist research, from documenting linguistic resources and their relationships to capturing the research process itself.

2. Motivation

A **reference constructicon** (hereafter: Cxn) is a collection of linguistic construction descriptions rooted in the framework of Construction Grammar, usually in the form of a databases accessible via a web application. A Cxn captures the full spectrum of linguistic units, including multi-word expressions, idiomatic phrases, and grammatical patterns. By contrast, a dictionary might catalog focus particles like *only* or *even* but not cleft constructions (e.g., *It was John who left*), which serve the same focalizing function. Cxns embrace the notion of a continuum of linguistic units (from words to phrases, sen-

Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026, pages 87–96
12 May 2026.

tences, and even discourse patterns). This holistic approach aligns with the core tenets of Construction Grammar (Goldberg, 1995; Croft, 2001), which posits that language is organized as a network of constructions at all levels of abstraction.

In Germany, a German Cxn (FKD) has been developed at Heinrich-Heine-Universität Düsseldorf (FKD), and an English Cxn at Friedrich-Alexander-Universität Erlangen-Nürnberg (Herbst et al. 2023–; Herbst and Hoffmann 2024). The scope of FKD, for example, ranges from (semi-)lexicalized constructions like "NP pur" ('NP in its purest form'; Babal 2026) to fully schematic constructions like nominal compounds[1].

While existing Cxns primarily serve as descriptive or reference resources, our project introduces a novel perspective: a **research constructicon** (RCxn) designed as a platform for documenting ongoing research in Construction Grammar. Currently, linguistic findings in this framework are predominantly published in research papers, lacking a centralized repository. Although initiatives like the DELPH-IN Consortium attempt to combine linguistic descriptions of constructions with varying levels of schematicity in a single implementation, their scope remains limited to specific communities, particularly those focused on formal syntax and semantics.

The RCxn aims to address this fragmentation by providing a community-driven tool for gathering, documenting, and sharing construction-based research. This platform will not only preserve the diversity of constructionist approaches but also enable researchers to build on existing work in a structured and transparent manner.

Building a RCxn presents unique challenges due to the inherently interconnected and dynamic nature of linguistic constructions, as well as the diversity of research in linguistics. One key complexity lies in capturing the relationships between constructions, such as inheritance hierarchies or sister constructions. These relationships are not merely descriptive but Construction Grammar posits that they have a cognitive reality. But comparative linguistics adds another layer of complexity, as the RCxn must facilitate cross-linguistic analyses of similar constructions. Finally, the diachronic dimension demands attention to how construction networks evolve over time, as historical shifts reshape inheritance links, functional roles, and formal properties. The network perspective is at the core of this kind of research: constructions are not isolated entities but part of a larger, dynamic system where their relationships define their behavior, variation, and evolution. The RCxn thus aspires to reflect this interconnectedness.

[1] https://framenet-constructicon.hhu.de/constructicon/constructionfamily?id=14

The RCxn is not only focused on the network that constructions form with each other, though. One of its main goals is to provide a rich description of individual constructions and the research findings associated with them. For example, we aim to document phraseological research, which examines the interplay between compositionality and idiomaticity, showing how constructions range from fully transparent to highly conventionalized units. The challenges extend further into multimodality, where constructions span multiple formal levels (e.g., syntax, prosody and gesture).

3. Methodology

Development of the RCxn is guided by a considerable number of PhD projects associated with the Research Training Group (close to 30 completed and ongoing projects in total so far). Their work spans a wide range of phenomena—from cross-linguistic and diachronic comparisons to multimodal analyses of speech and gesture, each requiring different levels of formalism and granularity. This diversity mirrors the broader Construction Grammar community, making our group an ideal microcosm for developing a resource that can scale to wider use. To ensure the RCxn meets real-world research needs, we started by collecting construction descriptions from these researchers, analyzing their requirements for both the database structure and the frontend interface. This bottom-up approach is central to the project because it ensures that the project is grounded in the practical needs of its contributors.

One illustrative example is a PhD project in the domain of **phraseology**, which investigates constructions with a high degree of idiomaticity (idioms, or even proverbs) across languages (Rastegar, 2026). This works explores the factors impacting the acquisition of such idiomatic expressions in a second (or foreign) language. Especially, it aims to identify the impact of similarities of idioms (both formal and functional) in the first and second language. The idiom *to take the bread out of someone's mouth* in English has formal as well as functional similarities with *nān-e kasi rā ājor kardan* (lit. 'make someone's bread a brick', meaning 'to deprive someone of their means of survival') in Persian. On the other hand, the latter is formally similar but functionaly different from *das ist ein hartes Brot* in German (lit. 'it is a hard bread', meaning 'this situation is difficult'). For this research, the ability to link constructions across languages based on varying degrees of formal and functional similarity was critical.

Another **multimodality** project investigated whether gesture is an inherent component of certain constructions or whether phonology and ges-

ture operate as separate, simultaneously activated streams (Iabdounane, 2025). This work identified inherently multimodal constructions such as "This close to V-ing". Its accompanying gesture (holding two fingers close together) appears to be entrenched alongside the verbal form. This finding suggests that the construction is truly multimodal, integrating both linguistic and gestural elements. For the RCxn, this posed a specific challenge: while it was necessary to include features for describing gestures, gesture could not be treated as a typical construction element. Unlike lexical or syntactic components, the duration and timing of gestures are more flexible.

A third project, situated in the domain of **computational linguistics**, examined how quantitative data from large corpora can be used to identify constructions in a (semi-)automatic manner (Patel, in prep). Central to this research is the use of colloprofiles, which are quantitative representations of a construction's collocational behavior. A colloprofile captures the statistical distribution of lexical items or grammatical patterns that co-occur with a construction, providing insights into its formal and functional properties. For example, the construction "come + [noise/manner of movement] V-ing" is particularly frequent with verbs such as running, tumbling, hurting, and bursting, which, in turn, confer a connotation of sudden, uncontrolled action upon the construction. It was essential for the RCxn to integrate colloprofiles into its database and ensure they are accessible and visualizable on the web interface.

Some challenges in designing the RCxn were about the representation of constructions in a general manner. For example, we needed to address how to capture the compositional aspects of a construction's meaning, how to model the nesting of construction elements (e.g., specific properties of a preposition within a prepositional phrase), and how to represent the relationship between a construction A occurring within a construction B and construction B itself. Other challenges, however, were specific to individual projects.

4. Technical Background

To the best of our knowledge, all existing reference constructicons rely on relational databases for their implementation. In contrast, our RCxn adopts Semantic Web Technologies, whose graph-based format is appealing for both theoretical and practical reasons. Theoretically, a **knowledge graph representation** aligns seamlessly with the principles of Construction Grammar, particularly as articulated by Diessel (2019, 2023).

Diessel's model conceptualizes linguistic knowledge as a complex network comprising two interrelated dimensions. First, each individual construction (or linguistic sign) forms a network in its own right, as functional features are linked to semantic features, and construction elements are interconnected through sequential dependencies or valency relationships. Diessel refers to this internal organization as the "sign as a network". Second, constructions themselves are embedded within a broader network, where schematic constructions serve as nodes in an inheritance hierarchy, and others are linked through derivational processes (e.g., a construction with a literal meaning with a formally similar construction expressing the metaphorical meaning) or shared similarities (e.g., phonetic similarities). This external organization is described as the "network of signs". Knowledge graphs are uniquely suited to model both the internal structure of a construction (the "sign as a network") and the relations between constructions (the "network of signs"), offering a theoretically grounded representation.

Our approach is further inspired by formalisms such as Head-Driven Phrase Structure Grammar (HPSG) and Sign-Based Construction Grammar (SBCG), where the internal structure of signs (e.g., a word, phrase, or construction) interconnects syntactic, semantic, and phonological properties within a single unit. For example, a verb's description might link its argument structure (e.g., subject and object requirements) to the respective semantic roles of said arguments (e.g., agent, patient), capturing how form and meaning are integrated within the sign itself. This corresponds very closely to Diessel's "sign as a network". At the same time, HPSG and SBCG also model "networks of signs" through hierarchical relationships, such as inheritance hierarchies or type hierarchies. For example, the head-modifier construction inherit properties from a more general headed construction. Formally, these framework use attribute-value matrices, which can very straightforwardly be conceived as graphs (with features as edges and values as nodes).

Practically, the use of **RDF (Resource Description Framework)** offers several advantages. Its flexible schema allows for heterogeneous construction descriptions, where certain features can be omitted without implying absence or default values. This can be seen as a drawback to those interested in highly formalized representations. However, in the reality of capturing linguistic research, this flexibility is essential. Researchers bring diverse interests and analytical depths to their work: some may prioritize fine-grained syntactic or semantic details, while others focus on other aspects (like interconnection with other constructions). A rigid schema would risk excluding valuable contributions.

Additionally **Linked Open Data** principles allows

for interoperability and community collaboration. By developing and sharing this ontology, we enable other researchers to build upon our work, while also ensuring compatibility with related projects, such as the ACoLi Dictionary Graph (Chiarcos et al., 2020). Our approach builds on pioneering efforts in linguistic ontology development, including GOLD (General Ontology for Linguistic Description; Farrar and Langendoen 2010), OLiA (Ontologies of Linguistic Annotation; Chiarcos and Sukhareva 2015; Chiarcos et al.), and lemon (W3C OntoLex Community Group), which have demonstrated the value of Semantic Web Technologies for linguistic resource representation.

The ontology serves as terminological baseline for the database containing construction descriptions, which we refer to as the A-box.[2] Figures 1, 2 and 3 illustrate the representation of a construction description in our database, showing a simplified version of the English "come + [noise/manner of movement] V-ing" construction. This construction is used to express that someone or something moves towards a reference point in a specific manner.[3] An example of this construction is *Sally came running excitedly into the room*, where someone (identified by the proper name *Sally*) moves toward the reference point (the inside of the room) in a very fast manner (denoted by the verb *run*). Figure 1 offers a visualization of the construction level. The construction consists of six elements: the (optional) subject, the verb *come*, a V-ing form of a verb expressing noise or manner of movement, an (optional) source, an (optional) trajectory and an (optional) goal. The IRI (Internationalized Resource Identifier) for the construction, `cx:comePLUSnoisemannerofmovementVing` (shortened in Figures 1 and 2 as `cx:CNM`) is related to its title, metadata, meaning and a sequence of its construction elements (see section 5.3.1). The IRI for the metadata (shortened to `cx:CNM_MD`) links to information about the creation date of the entry and the doctoral researcher who contributed this construction (`membr:Patel`). The researcher is in turn linked to their PhD project (`membr:Project_Patel`).

Figure 2 describes further two of the construction elements, the mandatory elements 2 and 3, and should therefore be seen as an expansion of two nodes of the previous graph. Both are linked to their formal description, a node of type `rcxn:SlotForm`, that links to other constructions. Element 2 is indeed an instance of the "come" construction and Element 3 of the "Verb+ing" construction.

The simplified description of this construction uses classes and properties defined by the Constructicography module (prefix `rcxn:`; section 5.3.1) of our ontology, by the RDF ontology (prefix `rdf:`; http://www.w3.org/1999/02/22-rdf-syntax-ns#), by the FOAF ontology (prefix `foaf:`; http://xmlns.com/foaf/0.1/) and by the DCMI Metadata Terms ontology (prefix `dcterm:`; http://purl.org/dc/terms/). The prefixes `cx:` and `membr:` refer to the A-box of the RDF database, where the IRIs for constructions (and their features) and for researchers (and their research questions) are stored, respectively.

To keep the visualization of Figures 1 and 2 comprehensible, we omitted many details of the construction description. The complete A-box contains the description of the other construction elements and their word order, the examples that illustrate the construction, research data such as colloprofile for Element 3, and further description of the researcher and their research question. Figure 3 displays fragments of the code in the A-box that correspond to what is shown in the visualization.

5. The Ontology

Having outlined the theoretical motivations and technological foundations of our project, we now turn to the core contribution of this paper: the modular ontology designed to be used in the RCxn. These modules share a common purpose (the RCxn's database) and are designed to document both linguistic resources and linguistic research.

5.1. Modular Ontology

Some of the modules in our ontology document **linguistic resources** while addressing foundational questions about the formal and functional properties of constructions. For instance:

- How many construction elements (e.g., slots, lexical items, or morphological markers) comprise a given construction?
- What is the meaning of the construction?
- Under which pragmatic conditions can this construction be used?

Complementing this, some of the modules are targeted at documenting **linguistic research**, while focussing on the scientific process and findings associated with each construction. Key questions here include:

- What is the colloprofile of the construction in a particular corpus?

[2] The A-box (where A stands for assertional knowledge), is stored separately and therefore uses different prefixes than the ontology, also called T-Box (where T stands for terminological knowledge).

[3] The full entry can be viewed online at https://bdlweb.phil.uni-erlangen.de/RCxn/app_entries/construction/comePLUSnoisemannerofmovementVing.

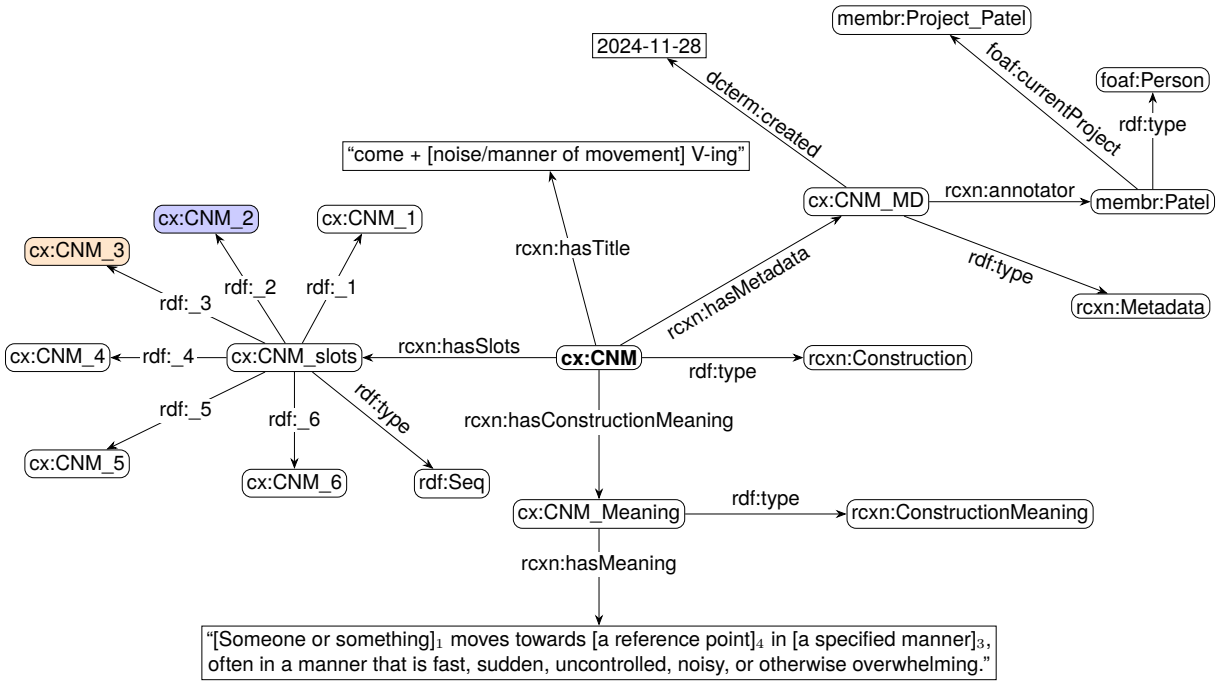

Figure 1: A simplified graph representation of the "come + [noise/manner of movement] V-ing" Construction in the Research Constructicon. Nodes with round edges denote IRI objects, while nodes with sharp edges denote literal objects.

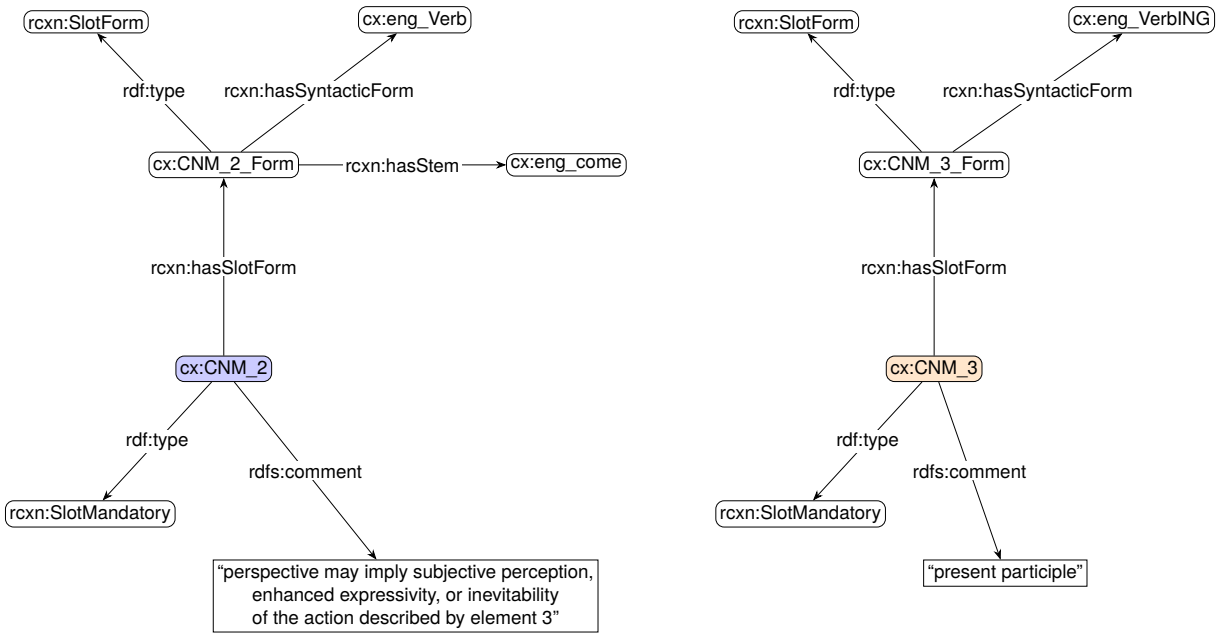

Figure 2: A simplified graph representation of the second and third construction elements of the "come + [noise/manner of movement] V-ing" Construction. Nodes with round edges denote IRI objects, while nodes with sharp edges denote literal objects.

```
membr:Patel a foaf:Person ;
    foaf:currentProject membr:Project_Patel ;
    [...] .

cx:comePLUSnoisemannerofmovementVing a rcxn:Construction ;
    [...]
    rcxn:hasConstructionMeaning cx:comePLUSnoisemannerofmovementVing_Meaning ;
    rcxn:hasMetadata cx:comePLUSnoisemannerofmovementVing_MD ;
    rcxn:hasSlots cx:comePLUSnoisemannerofmovementVing_slots ;
    rcxn:hasTitle "come + [noise/manner of movement] V-ing" .

cx:comePLUSnoisemannerofmovementVing_MD a rcxn:Metadata ;
    [...]
    rcxn:annotator membr:Patel ;
    dcterm:created "2024-11-28"^^xsd:date .

cx:comePLUSnoisemannerofmovementVing_Meaning a rcxn:ConstructionMeaning ;
    rcxn:hasMeaning "[Someone or something]1 moves towards [a refer-
ence point]4 in [a specified manner]3, often in a manner that is fast, sudden, un-
controlled, noisy, or otherwise overwhelming." .

cx:comePLUSnoisemannerofmovementVing_slots a rdf:Seq ;
    rdf:_1 cx:comePLUSnoisemannerofmovementVing_1 ;
    rdf:_2 cx:comePLUSnoisemannerofmovementVing_2 ;
    rdf:_3 cx:comePLUSnoisemannerofmovementVing_3 ;
    rdf:_4 cx:comePLUSnoisemannerofmovementVing_4 ;
    rdf:_5 cx:comePLUSnoisemannerofmovementVing_5 ;
    rdf:_6 cx:comePLUSnoisemannerofmovementVing_6 .

cx:comePLUSnoisemannerofmovementVing_2 a rcxn:SlotMandatory ;
    [...]
    rdfs:comment "perspective may imply subjective perception, enhanced expressiv-
ity, or inevitability of the action decribed by element 3" ;
    rcxn:hasSlotForm cx:comePLUSnoisemannerofmovementVing_2_Form .

cx:comePLUSnoisemannerofmovementVing_2_Form a rcxn:SlotForm ;
    rcxn:hasStem cx:eng_come ;
    rcxn:hasSyntacticForm cx:eng_Verb .

cx:comePLUSnoisemannerofmovementVing_3 a rcxn:SlotMandatory ;
    [...]
    rdfs:comment "present participle" ;
    rcxn:hasSlotForm cx:comePLUSnoisemannerofmovementVing_3_Form .

cx:comePLUSnoisemannerofmovementVing_3_Form a rcxn:SlotForm ;
    rcxn:hasSyntacticForm cx:eng_VerbING .
```

Figure 3: Fragments of the A-box of the RCxn: RDF-Turtle code for the “come + [noise/manner of movement] V-ing” Construction. Some triples have been omitted to enhance readability.

- What research questions motivated the investigation of this construction?
- What empirical findings or theoretical insights have researchers reported about the construction, and how do these contribute to broader debates in Construction Grammar?

Together, this modular ontology ensure that the RCxn not only documents the *what* of linguistic constructions but also the *how* and *why* of the scientific research.

Rooted in the theoretical framework of Construction Grammar, we did our best to make this ontology inclusive to all currents in Construction Grammar. While their immediate application is tailored to the needs of our RCxn, their development reflects a broader ambition: to provide a shared resource for the the community. By adopting open standards and modular design principles, we aim to facilitate interoperability with other constructicography projects, enabling a community-driven ecosystem.

5.2. Technical Implementation and Standards

The development of our ontology builds upon established semantic web standards. We used existing ontologies such as RDFS (Resource Description Framework Schema) for foundational modeling, FOAF (Friend of a Friend) for describing researchers and their contributions, SKOS (Simple Knowledge Organization System) for conceptual organization, and especially OLiA (Ontologies of Linguistic Annotation) for linguistic features in describing constructions.

The ontology is formalized in OWL (Web Ontology Language), a standard for defining and instantiating ontologies on the Semantic Web. We used the ontology editor Protégé to write the different modules (except the `comcon` module, as explained below). In the OWL standard, classes are defined as `owl:Class` and properties as `owl:ObjectProperty`. Classes represent the categories of linguistic and research entities—such as constructions (`rcxn:Construction`), construction elements (`rcxn:Slot`), research questions (`rsrch:Project`), or language variety (`lg:variety`), while properties describe the relationships between these entities, such as a construction belonging to a specific language (`lg:partOfLanguage`), having specific construction elements (`rcxn:hasSlots`), or a researcher working on a specific research question (`rsrch:hasResearchQuestion`).

Each class and property is assigned a unique IRI (Internationalized Resource Identifier) for global identification, along with human-readable labels (`rdfs:label`) and, where applicable, alternative labels (`skos:altLabel`). Definitions are provided using `rdfs:comment`. Hierarchical relationships between classes are expressed using `rdfs:subClassOf`. Some OWL features can be used to enrich the description, such as `owl:disjointWith`, which states that two class are not mutually compatible (a single IRI cannot belong to both classes simultaneously). For example, the class `rcxn:SlotMandatory` is defined as a subclass of `rcxn:Slot`, disjoint from the class `rcxn:SlotOptional` (i.e., a construction element is either optional or non-optional); its label is “Non-optional element”, and its definition (rdfs:comment) is “Non-optional construction elements that need to be realized.”.

Hierarchical relationships between properties are expressed using `rdfs:subPropertyOf`. For properties, we employ OWL features such as `owl:SymmetricProperty`, `owl:TransitiveProperty` or `owl:equivalentProperty` where relevant.

For illustrative purpose, Figure 4 displays the code for the property `rcxn:hasSemanticContribution`, which links a construction element to the semantic meaning it brings to the construction. Our contributors can use OLiA’s rich vocabulary for semantic roles to describe this semantic contribution, in which case the subproperty `rcxn:hasSemanticRole` is used. This property ranges over an object of the class `oliatop:SemanticRole` and is indicated as being an equivalent to `olia:hasSemanticRole`, thus ensuring as many interoperability with OLiA as possible.[4] However, contributors who are not willing to use the terminology of semantic role can use the other subproperty `rcxn:hasOtherSemanticContribution` whose object is a literal object.

5.3. Brief Overview of the Modules

The RCxn is supported by a suite of ontology modules. A comprehensive technical documentation of the ontology and its modules is available on the project’s GitHub repository.[5] Below, we provide a concise overview of each module and its role within the RCxn.

5.3.1. The Constructicography Module (`rcxn.rdf`)

The Constructicography module, stored in `rcxn.rdf` and accessible via the prefix `rcxn:` (for https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#), forms the core of our framework for representing constructions in a Cxn. The prominent class of this module is `rcxn:Construction`, which serves to identify an IRI as a construction. Each construction is linked to a human-readable title (as a literal object) and associated metadata (an object of class `rcxn:Metadata`).

In line with the principles of Construction Grammar, the construction is also linked to a form and a meaning. Meaning is represented as an object of the class `rcxn:Meaning`, and can be further described with the help of associated classes and properties. Form, however, is not treated as a monolithic entity, but rather decomposed into a sequence of construction elements (objects of type `rcxn:Slot`). This design applies uniformly, whether the construction consists of a single element (e.g., morphemes or monomorphemic lexemes) or multiple elements. Each slot can then be further characterized by a series of features,

[4] It is technically not possible to use the property `olia:hasSemanticRole` directly because it links a linguistic annotation to a semantic role, whereas we need a property that takes a construction element as subject.

[5] https://github.com/ElodieWinckel/RCxn/blob/master/ontologies/ontology_doc.md

```
<owl:ObjectProperty rdf:about="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasSemanticContribution">
    <rdfs:domain rdf:resource="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#Slot"/>
    <rdfs:comment xml:lang="en">Describes the meaning (e.g., semantic role) of the construction element in the construction.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasSemanticRole">
    <rdfs:subPropertyOf rdf:resource="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasSemanticContribution"/>
    <rdfs:range rdf:resource="http://purl.org/olia/olia-top.owl#SemanticRole"/>
    <rdfs:comment xml:lang="en">Describes the semantic role of the construction element in the construction. The semantic roles are de-
fined by the OLiA ontology.</rdfs:comment>
    <rdfs:label xml:lang="en">Semantic role</rdfs:label>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="http://purl.org/olia/olia.owl#hasSemanticRole">
    <owl:equivalentProperty rdf:resource="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasSemanticRole"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasOtherSemanticContribution">
    <rdfs:subPropertyOf rdf:resource="https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rcxn#hasSemanticContribution"/>
    <rdfs:range>
        <owl:Restriction>
            <owl:onProperty rdf:resource="http://www.w3.org/2002/07/owl#topDataProperty"/>
            <owl:someValuesFrom rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
        </owl:Restriction>
    </rdfs:range>
    <rdfs:comment xml:lang="en">Describes the meaning of the construction element in the construction that cannot be captured by the seman-
tic roles of the OLiA ontology.</rdfs:comment>
    <rdfs:label xml:lang="en">Semantic contribution</rdfs:label>
</owl:ObjectProperty>
```

Figure 4: Fragment of the ontology definition of rcxn: RDF-XML code for the property rcxn:hasSemanticContribution and its two suproperties rcxn:hasSemanticRole and rcxn:hasOtherSemanticContribution

many of which are drawn from the OLiA ontology (e.g., `olia:hasAnimacy`, with possible values like `olia:Inanimate`). The module defines however a series of features, for example `rcxn:hasPhonology` (which takes as object a literal object).

Some classes and properties are furthermore dedicated to the bibliographical documentation of the construction, and to other aspects of its metadata (for example `rcxn:annotator`).

5.3.2. The Research Module (`rsrch.rdf`)

The Research module, stored in `rsrch.rdf` and accessible via the prefix `rsrch:` (https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/rsrch#), defines classes and properties that document the research process itself, including research questions, findings, and the researchers involved. For example, the class `rsrch:Project` represents individual research initiatives, and `rsrch:Finding` encapsulates empirical results, theoretical insights, or methodological contributions derived from these projects and from the study of individual constructions. To represent the researchers and their contributions, the ontology uses the FOAF (Friend of a Friend) vocabulary, enabling detailed descriptions of individuals.

This module is essential to the RCxn project that aims not only at documenting constructions but also the evolving scientific discourse surrounding them. Constructicography projects that do not concentrate on this aspect might have no need of this module.

5.3.3. The Links Module (`links-1.0.rdf` and beyond)

The Links module is dedicated to modeling the relationships between constructions. It is currently available in two versions (`links-1.0.rdf` and `links-1.1.rdf`) under the prefix `links:`. The initial version (`links-1.0.rdf`) was the topic of a previous publication (Winckel, 2025) aimed at a non-computational audience, focusing on lexicographic and theoretical issues in Construction Grammar. Evolutions in the project led to the development of `links-1.1.rdf`, which remains fully backward-compatible with the 1.0 version. This is a commitment to compatibility we intend to maintain in all future updates.

This ontology is grounded in the theoretical framework proposed by Diessel (2019, 2023), capturing both vertical relationships, that is inheritance links (e.g., a specific construction inheriting properties from a more general schema), and horizontal relationships, also called sister constructions, that share functional or formal similarities. Additionally, the ontology accommodates a cross-linguistic dimension, enabling comparative analyses of constructions across languages for projects in contrastive linguistics. Beyond these core relationships, the ontology also supports other types of links, such as metaphorical relationships, which highlight semantic extensions or conceptual mappings between constructions.

5.3.4. The Comparative Concepts Module (`compcon.ttl`)

The Comparative Concepts module, stored in `compcon.ttl` and accessible via the prefix `compcon:` (https://bdlweb.phil.uni-erlangen.de/RCxn/ontologies/compcon#), is dedicated to facilitating

cross-linguistic comparisons of constructions. Croft (2022) proposed a series of Comparative Concepts, which are abstract, typologically grounded categories (e.g., the construction category "free relative clause construction", or the semantic category "agent") that serve to compare constructions across Cxn, and therefore languages. This module is part of the broader MoCCA initiative (Lorenzi et al., 2024), a collaborative effort among Cxn projects to leverage the typological power of Comparative Concepts in order to align construction descriptions.

While the RCxn actively participates in the MoCCA enterprise, the `compcon.ttl` ontology itself is not an original development of our project. Instead, it represents an RDF translation of a pre-existing database of comparative concepts, originally developed by the MoCCA team and based on Croft (2022). This database is publicly available on their GitHub repository as a YAML database[6], and the aim of this module is to ensure interoperability of our RCxn with their framework.

5.3.5. Additional Modules: Language and Project-Specific Resources

In addition to the core modules, the RCxn includes a few auxiliary modules designed for specific purposes. The Language module (`lg.rdf`) serves as a repository for language instances used within the RCxn, providing definitions for the languages documented in our RCxn. While functional for our internal workflows, this module is highly project-specific. The CASA module (`casa.rdf`) is tied to a research project in collaboration with the English CASA Constructicon. Given its project-specific focus, this module does not need be detailed further here. It is available for reference in the project's GitHub repository.

6. Conclusion

6.1. Summary

This paper has presented the theoretical motivations, technical implementation, and modular design of the ontology developed for our RCxn project. We began by outlining the need for a dynamic, community-driven resource that documents linguistic research. We showed that a network-based approach is particularly suited to model descriptions rooted in Construction Grammar. Semantic Web Technologies have the further advantage of being flexible and the interoperability with other linguistic ontologies is an important asset.

We presented a modular ontology, with each module tailored to address specific aspects of constructionist research. The Constructicography module captures the core structure of constructions, modeling their form and meaning. The Research module documents the scientific process, linking research questions, findings, and contributors to the constructions they investigate. The Links module formalizes the relationships between constructions, while the Comparative Concepts module aligns with the MoCCA initiative to enable cross-linguistic comparisons.

6.2. Future of the Project

In the near future, we plan to identify and implement additional classes and properties to further refine the ontology. We are also exploring the development of a dedicated module to document empirical evidence for constructions, as well as an additional module to model examples of constructions, which require specific features such as glosses or translations. However, we remain open to using existing ontologies for this purpose if they meet our requirements.

Beyond its immediate goals, the RCxn project offers several broader benefits for the field of Construction Grammar. It contributes indeed to linking resources in this field. This makes future connections with other Cxn projects possible and enable a more holistic perspective on construction knowledge. The RCxn serves as a model for future projects, demonstrating how constructicographic work can be integrated with research documentation. The project also provides a robust software infrastructure and a well-designed ontology that can be adopted by other projects in Construction Grammar.

The project will remain accessible on the FAU server, ensuring that all current IRIs and resources stay available. The ontology will continue to be open for use and expansion by the community, even if no further development occurs. We aim, however, to secure additional funding to further enhance the project.

This work is very much a work in progress, and we welcome feedback and collaboration from the Construction Grammar community. With this an open and adaptive resource, we hope the RCxn will serve as a valuable tool in constructicography.

7. Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of the Research Training *Group Dimensions of Constructional Space* (project no. 468527017).

We are deeply grateful to the members of the Research Training Group for their invaluable contributions throughout this project. We also extend

[6] https://github.com/comparative-concepts/cc-database

our sincere thanks to three anonymous reviewers for their insightful feedback.

8. Bibliographical References

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press, Cambridge.

Holger Diessel. 2019. *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press, Cambridge.

Holger Diessel. 2023. *The Constructicon: Taxonomies and Networks*. Elements in Construction Grammar.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Thomas Herbst and Thomas Hoffmann. 2024. *A Construction Grammar of the English Language: CASA – a Constructionist Approach to Syntactic Analysis*. John Benjamins Publishing Company.

Yassine Iabdounane. 2025. *Multimodal Constructional Space*. Ph.D. Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig, and Ely E Matos. 2024. MoCCA: A Model of Comparative Concepts for Aligning Constructicons. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 93–98, Torino, Italia. ELRA and ICCL.

Malin Patel. in prep. *Corpus Evidence for Delineating Constructions*. Ph.D. Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Aria Rastegar. 2026. *Understanding idioms across languages*. Ph.D. Thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Elodie Winckel. 2025. Defining relationships in the constructional network: A Semantic Web ontology for Construction Grammar. *Lexicographica*, 41(1):299–317.

9. Language Resource References

Kardelen Babal. 2026. *'Konstruktion: Intensivierung_Attribut_nachgestellt:NP_pur' in FrameNet-Konstruktikon des Deutschen*. https://framenet-constructicon.hhu.de/constructicon/construction?id=1323 (accessed 19.02.2026).

Chiarcos, Christian and Adam, Angelika and Hellmann, Sebastian and Sukhareva, Maria and Fäth, Christian and Abromeit, Frank and Ionov, Maxim and Dimitrova, Vanya. *Ontologies of Linguistic Annotation (OLiA)*. Applied Computational Linguistics (ACoLi) Lab at the Goethe University Frankfurt, Germany.

Chiarcos, Christian and Fäth, Christian and Ionov, Maxim. 2020. *The ACoLi Dictionary Graph*. European Language Resources Association.

DELPH-IN Consortium. DELPH-IN: Deep Linguistic Processing with HPSG.

Scott Farrar and D. Terence Langendoen. 2010. *General Ontology for Linguistic Description*. Institute for Language Information and Technology.

FKD. *FrameNet-Konstruktikon des Deutschen*. https://framenet-constructicon.hhu.de (accessed 19.02.2026).

Herbst, Thomas and Hoffmann, Thomas and Uhrig, Peter and Garibyan, Armine and Evert, Stephanie. 2023–. *CASA ConstructiCon of the English Language*.

W3C OntoLex Community Group. *lemon - The Lexicon Model for Ontologies*.

Author Index