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.

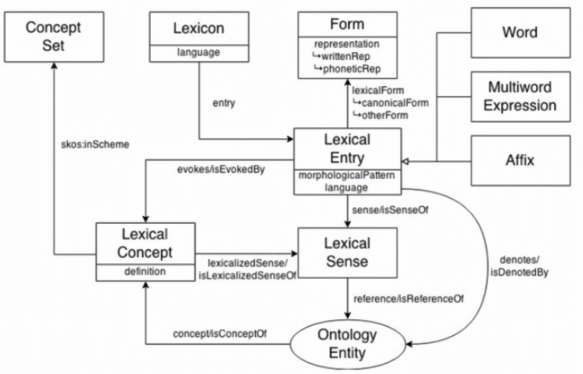

Figure 1: illustrates the core elements of the Ontolex-Lemon model (Cimiano et al., 2016).

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).

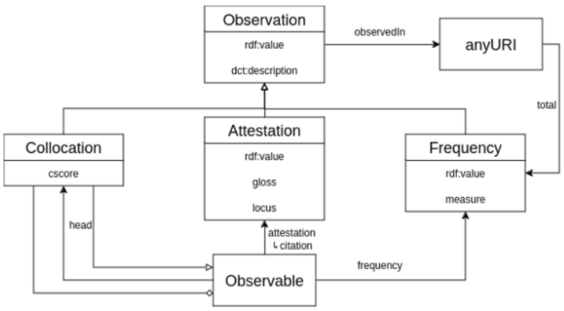

Figure 2: shows the OntoLex Module for Frequency, Attestation and Corpus Information (FrAC) (Chiarcos et al., 2025).

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.

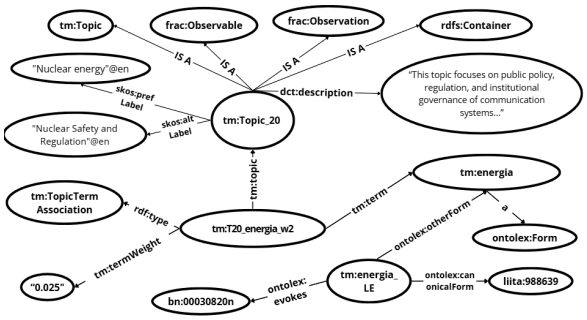

Figure 3: illustrates the modeling of Topic 20 as a `tm:Topic`, `frac:Observable`, `frac:Observation`, and `rdfs:Container`. The topic is associated with its preferred and alternative labels, as well as the LLM-generated description derived from the qualitative evaluation. Its lexical representation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

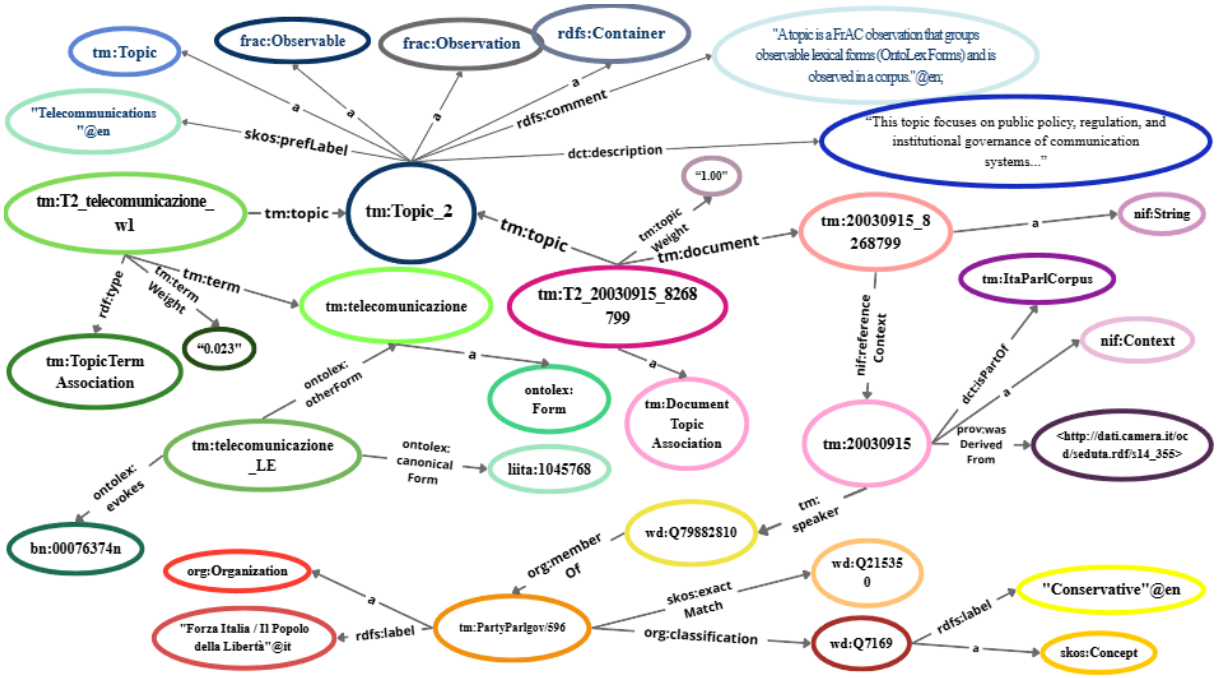

Figure 4: formalizes the representation of Topic 2, labeled as 'Telecommunications' and linked both to its lexical representation (here only the first most representative word is shown) and to one of the document that discuss this topic, together with information related to the speaker associated to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