Modeling linking between text and lexicon with OntoLex-Lemon: a case study of computational terminology for the Babylonian Talmud

Flavia Sciolette
Institute for Computational Linguistics "A.Zampolli"
Via Giuseppe Moruzzi, 1, 56124 Pisa PI, Italia
flavia.sciolette@ilc.cnr.it

Abstract

This paper illustrates the first steps in the creation of a computational terminology for the Babylonian Talmud. After introducing the motivation for this work and the state of the art, the paper exposes the choice of using OntoLex-Lemon and the new FrAC module for encoding the attestations and quantitative data of the terminology extraction. The Talmudic terminological base is introduced, with an example of an entry populated with the above-mentioned data. The choices for modeling are motivated by the rich representation the model allows and also for future needs for the management of the link between text and lexical entries.

Keywords: Computational Terminology, Linked Open Data, Talmud, OntoLex-Lemon, FrAC

1. Introduction

Over time, the gap between lexicographic and terminological practices has narrowed (Salgado et al., 2022) in terms of models and methodologies, thanks to the 'linguistic turn' of the 2000s (Bellandi et al., 2020), which is now well established in several studies. A term is the linguistic realisation of a domain concept (Buitelaar et al., 2005); in many contexts, however, the exact correspondence between term and concept - between actual use and norm - is not taken for granted (Soffritti, 2010). Therefore, for many resources, the text becomes a necessary starting point for the observation of the term, a manifestation of the word (Chiocchetti and Ralli, 2022), and consequently a basis for its extraction, in order also to subsequently build a knowledge base (Buitelaar et al., 2005). Although the debate on the creation of these resources is also inevitably experiencing the influence of Large Language Models (LLM)[1], we point out that the link between the text as source and term analysis is still fundamental when dealing with historical or highly specialised languages, less rich in resources suitable for training specific models. The preservation of this link allows to represent useful information, all the more so when considering quotations from corpora as examples of authentic linguistic usages (Klosa, 2015), which can convey many linguistic, historical, and cultural information.

1.1. Motivation

We choose to present the case study offered by the Babylonian Talmud - a fundamental text for Jewish religion and culture - and the creation of a terminological resource, currently under development, for the project of Italian translation for this text. A complex task like translation allows for an in-depth discussion on the need for resources based on state-of-the-art models and formats, as well as on shared standards that guarantee broad use and interoperability of data. The resource is indeed built according to the Linked Open Data (LOD) principles (Bizer et al., 2023) and recent good practices for modeling quotations and attestations. In the following sections, we consider the related work about the chosen case study, and consequently, a section is dedicated to the choice of adopting the OntoLex-Lemon model and the recent FrAC module for modeling attestations and frequency values; an example of an entry is provided, linked to contexts extracted from the treatises of which the Talmud is composed; finally, future developments are outlined in the conclusions. The resulting terminological resource is intended to be a useful tool to deepen the study of the languages used in the Talmud and to help translators in their choices.

2. Related Work

The Talmud represents a fundamental text for Judaism and constitutes a veritable mine of historical, cultural, social, legal, and scientific information. Among religious texts, it appears to be one of the

[1] For considerations on ontologies and ontology learning, see (Neuhaus, 2023) and (Babaei Giglou et al., 2023). For experiments on term and entity extraction, see (Meoni et al., 2023), (Liu et al., 2023), with general considerations on the use of these models in low-resourced contexts. For translation, similarly see (Robinson et al., 2023).

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

richest for cultural and linguistic information, as well as one of the most complex, considering also that it is multilingual, with a formulaic structure, and it is further enriched with several commentaries. Before the Italian translation project, there were no specific resources available for the Talmud in Italian, neither printed nor digital. About the latter in other languages, for an overview, see (Giovannetti et al., 2020) and (Saponaro et al., 2022), in particular for resources available as LOD. For the Italian language, the terminological extraction conducted on the translated treatises of the Talmud is also described in (Giovannetti et al., 2020). This extraction was also used in a terminology graph visualisation application (Marchi et al., 2022). Other experiments on the extraction of named terms and entities were mainly concerned with the creation of an ontology on master rabbis (Giovannetti et al., 2021). Also noteworthy is the study of the networks of relationships between master rabbis by (Satlow and Sperling, 2022).

3. Ontolex-Lemon

Linguistic Linked Open Data (LLOD) (Cimiano et al., 2020) constitute a subset of LODs and represent a set of best practices that facilitate the sharing and reuse of linguistic data in various applications and research domains, according to FAIR principles (Wilkinson et al., 2016). OntoLex-Lemon (McCrae et al., 2017) is the most well-known and widely used vocabulary for the creation, publication, and sharing of lexical and terminological resources such as LLOD. The model includes several extensions that have already been published or are currently under development; these include the module for representing frequency, attestation, and corpus information (FrAC).

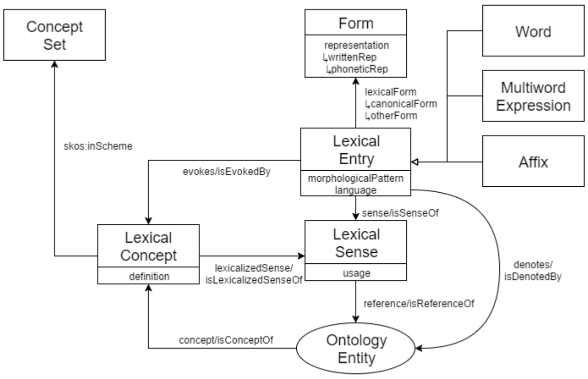

Figure 1: Module Core of OntoLex-Lemon.

3.1. FrAC

FrAC is currently at an advanced phase (Chiarcos et al., 2022)and undergoing final revision[2]. Here we mention only the part of module used for the examples in this paper: the class `Attestation`, which constitutes "a special form of citation that provides evidence for the existence of a certain lexical phenomenon"; the property `attestation`, which associates `Attestation` with an `Observable` of FrAC. To this is added the relation `quotation`, to insert the text of the quotation in natural language; the `Frequency` for the absolute number of times the term appears in an attestation; the property `measure` to indicate the term frequency-inverse document frequency (tf-idf[3]).

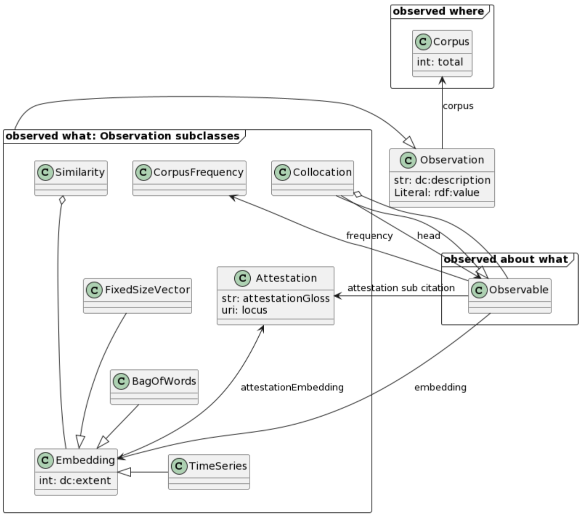

Figure 2: Diagram of FrAC.

4. Talmudic Terms

4.1. First steps of the terminological base

The starting point was the extraction of terms from the translated treatises of the Talmud. In this context, the considered definition was "a (candidate) term was defined as a simple (single-word) or complex (multi-words) nominal structure with modifiers" (Giovannetti et al., 2020). It is therefore a 'procedural' definition, functional for querying the text using regular expressions. The extraction was conducted with the Term To Knowledge tool (T2K) (Dell'Orletta et al., 2014), on the Italian translation[4]. For each

[2] https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md

[3] https://it.wikipedia.org/wiki/Tf-idf

[4] There are no tools for the automatic analysis of Biblical and Mishnaic Hebrew. For the use of tools for Modern Hebrew, see (Pecchioli et al., 2018).

term, the absolute frequency and tf-idf were provided. A high tf-idf value implies that the term frequently appears in a few documents and is therefore specific, whereas a low tf-idf value means that the term is distributed in many different documents (as is the case, for example, with ’rabbi’, which appears extensively throughout the Talmud treatises). Consequently, it was decided to model all terms with high tf-idf values, which were then manually checked by domain experts (approximately 4000 terms). These data were exported in a .csv format. From the .csv format, through a specially created Python script, the data structures for lexical entry were created, including language, canonical form, sense, absolute frequency, tf-idf, and the treatise to which the term belongs. The natural language definition and example quotation content were entered manually.

4.2. An example of entry

It follows an example of an entry modeled according to the OntoLex-Lemon model for Talmud terminology. The term is the Hebrew ’shemà’ which indicates one of the obligatory readings to be performed during the day. The main entry is the lexical entry `:shema`, which is associated with various types of morpho-syntactic information (part of speech, gender, number, etc.).

```
:shema_entry a ontolex:lexicalEntry;
dct:language <http://www.lexvo.org/page/
    iso639-3/ita>;

lexinfo:partOfSpeech lexinfo:commonNoun ;
lexinfo:gender lexinfo:masculine ;
lexinfo:number lexinfo:singular.
```

The lexical entry is associated with the sense, a canonical form corresponding to the lemma contained in the glossaries in use by the translators of the Talmud project, and the absolute frequency value. According to the description of the frequency class in FrAC, it is possible to associate the value with both the entry and the form; in this case, it was chosen to associate the value with the entry and to specify it further in the description of the individual forms if necessary.

The `observedIn` relation clarifies the source of the data (in this case, the frequency) in the form of a URI. A second frequency information is modeled with the `frac:measure` relation to providing the value of tf-idf as an rdf:value. Finally, a certain number of individuals are associated to the entry with the `attestation` relation (three examples of segments are given in the modeled entry). Figure 3 provides a visual representation of the scheme for the entry.

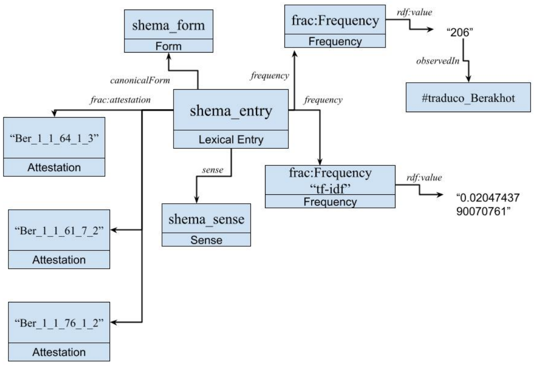

Figure 3: Encoding of the entry "shemà".

```
ontolex:sense :shema_sense;
ontolex:canonicalForm :shema_form;
frac:frequency[
a frac:Frequency;
rdf:value "206"^^xsd:int;
frac: observedIn <#traduco_berakhot>
      ];
frac:frequency[
a frac:Frequency; frac:measure "tf-idf" ;
rdf:value "0.0204743790070761"
      ].
frac:attestation :Ber_1_1_64_1_3,
    Ber1_1_61_7_2, Ber_1_1_76_1_2.
```

The definition is provided with the relation defined in SKOS[5] to provide a natural language description of sense, taken from the Talmudic glossaries compiled in the project. In this way, we can preserve the attestations and the definitions written by domain experts, as in the case of the shemà: "Three passages from the Pentateuch: ’Hear, O Israel’ (Deut. 6:4-9), ’And if you will listen’ (Deut. 11:13-21), ’And the Lord spoke to Moses’ (Num. 15:37-41), the reading of which is obligatory twice a day, in the morning and the evening." The selected example is contained in the first treatise of the Talmud, Berakhot: "One may stand and recite the Shemà."

```
:shema_sense a ontolex:LexicalSense;

skos:definition "Tre␣brani␣del␣Pentateuco
    :␣  Ascolta  ␣Israele ␣(Deut.␣6:4-9)
    ,␣  E  ␣se␣ascolterai ␣(Deut.␣
    11:13-21),␣  E  ␣il␣Signore␣disse␣a␣
    Moshè   ␣(Num.␣15:37-41),␣la␣cui␣
    lettura␣è␣obbligatoria␣due␣volte␣al␣
    giorno,␣alla␣mattina␣e␣alla␣sera."@it
    .

:Ber_1_1_64_1_3 frac:Attestation ;
frac:quotation "Si␣può␣stare␣in␣piedi␣e␣
    recitare␣lo␣Shemà".@it.
```

In this way, it is possible to link different sources of

[5] https://www.w3.org/TR/skos-reference/

data, even if not originally LLOD. Individuals can be described as an attestation; through the quotation, it is possible to provide the content of the attestation in natural language (a single segment as an example). The adoption of FrAC thus makes it possible to enrich the knowledge graph of lexical entries with quantitative information, potentially useful for various tasks such as topic modeling. In this way, other knowledge graphs outside the terminology or text annotations can also be linked to the entries (see next section).

4.3. Text management and lexical linking

The rich amount of information related to Jewish culture in the Talmudic text is systematised in the glossary entries prepared within the project. These glossaries have been produced in a translation-oriented manner and are therefore based on the annotation of the term in both the original text and its Italian counterpart. Maintaining this link is therefore fundamental in the elaboration of a terminological resource; the adoption of FrAC enables the linking of attestations to elements outside the graph, including annotations to the text itself. We call this task 'lexical linking', which consists, similarly to entity linking, in linking words in texts with linguistic entities (lemma, meanings, etc.) encoded in another resource. These annotations may include terms, but also information of a different nature, distributed on different annotation layers (e.g. to encode specific formulae in which terms are inserted). Currently, this phase is handled manually through an editor, named "Maia" prepared for this purpose, under development[6].

5. Conclusion and future works

In this paper, we presented a case study, offered by the creation of Talmudic terminology, for the encoding of dictionary attestations and quantitative data. The starting point was offered by an extraction of terminology for the Italian translation of the Talmud. We, therefore, presented an example of an entry using the new FrAC module, currently undergoing final revision, to show its productivity, also with a view to future linking with the annotated text, thanks also to a specific editor currently being developed. Future work includes linking terminology entries and specific senses to an Italian reference lexicon in LLOD, Compl-It[7] currently available on CLARIN; linking to ontological references of individual terms and occurrences of Hebrew terms in the untranslated text.

6. Acknowledgments

This work was conducted in the context of the TALMUD project, the scientific cooperation between S.c.a r.l. PTTB and CNR-ILC, and the contribution of a short-term mobility (STSM) grant within the COST Action “Nexus Linguarum”, at the University of Augsburg.

7. Bibliographical References

Hamed Babaei Giglou, Jennifer D’Souza, and Sören Auer. 2023. LLMs4OL: Large language models for ontology learning. page 408–427, Berlin, Heidelberg. Springer-Verlag.

Andrea Bellandi, Emiliano Giovannetti, Simone Marchi, Silvia Piccini, and Flavia Sciolette. 2020. Come dare senso a un termine? caratteristiche, potenzialità e opportunità dello strumento Lexo. In *Comunicazione al XXX Convegno Ass.I.Term.*, Eurac Research, Bolzano.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2023. Linked data – the story so far. In *Linking the World’s Information: Essays on Tim Berners-Lee’s Invention of the World Wide Web*, pages 115–143. Association for Computing Machinery, New York, NY, United States.

Paul Buitelaar, Philippe Cimiano, and Bernardo Magnini. 2005. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian 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, Gyeongju, Republic of Korea.

Elena Chiocchetti and Natascia Ralli. 2022. Introduzione. In *Risorse e strumenti per l’elaborazione e la diffusione della terminologia in Italia*. Eurac Research, Bolzano.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. Linguistic linked data in digital humanities. In *Linguistic Linked Data in Digital Humanities*, pages 229–262. Springer International Publishing, Cham.

Felice Dell’Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2K^2: a system for automatically extracting and organizing

[6] https://github.com/klab-ilc-cnr/Maia

[7] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle /20.500.11752/ILC-1007

knowledge from texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation*, pages 178–190, Reykjavik. ACL.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, David Dattilo, Angelo Maria Del Grosso, and Simone Marchi. 2021. An ontology of masters of the Babylonian Talmud. *Digital Scholarship in the Humanities*, fqab043.

Emiliano Giovannetti, Andrea Bellandi, David Dattilo, Angelo Maria Del Grosso, Simone Marchi, Alessandra Pecchioli, and Silvia Piccini. 2020. The terminology of the Babylonian Talmud: Extraction, representation and use in the context of computational linguistics. *Materia Giudaica*, XXV.

Annette Klosa. 2015. On corpus citations in monolingual general dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 36(1):72–87.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Simone. Marchi, Marianna Colombo, David Dattilo, and Emiliano Giovannetti. 2022. Un esperimento di visualizzazione grafica della terminologia del Talmud Babilonese. In *AIUCD 2022 - Proceedings*, Lecce, Italy.

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 conference*, pages 19–21.

Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.

Fabian Neuhaus. 2023. Ontologies in the era of large language models: a perspective. *Applied Ontology*, 18(4):399–407.

Alessandra Pecchioli, Davide Albanesi, Andrea Bellandi, Emiliano Giovannetti, and Simone Marchi. 2018. Annotazione linguistica automatica dell'ebraico mishnaico: Esperimenti sul Talmud Babilonese. *Materia Giudaica*, XXIII.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Ana Salgado, Rute Costa, and Toma Tasovac. 2022. Applying Terminological Methods to Lexicographic Work: Terms and Their Domains. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pages 181–195, Mannheim. IDS-Verlag.

Davide Saponaro, Emiliano Giovannetti, and Flavia Sciolette. 2022. From religious sources to computational resources: Approach and case study on hebrew terms and concepts. *Materia Giudaica*, 27.

Michael L. Satlow and Michael Sperling. 2022. The rabbinic citation network. *AJS Review: The Journal of the Association for Jewish Studies*, 46(2):291–319.

Marcello Soffritti. 2010. Termontografia e innovazione della terminologia plurilingue. In Franco Bertaccini, Sara Castagnoli, and Francesco La Forgia, editors, *Terminologia a colori*, pages 31–5. Bononia University Press, Forlì.

Mark D. Wilkinson, Michel Dumontier, I Jsbrand Jan Aalbersberg, Gabriel Appleton, Myles Axton, Arie Baak, and Niklas Blomberg. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.