When OntoLex Meets Wikibase: Remodeling Use Cases

David Lindemann[♣] **Sina Ahmadi**[♠] **Anas Fahad Khan**[◊] **Francesco Mambrini**[♦]
Federica Iurescia[♦] **Marco Passarotti**[♦]
[♣]UPV/EHU University of the Basque Country, Vitoria-Gasteiz, Spain
[♠]Department of Computer Science, George Mason University, Fairfax, VA, USA
[◊] CNR-ILC, Italy
[♦] Università Cattolica del Sacro Cuore, Milan, Italy
david.lindemann@ehu.eus,sahmad46@gmu.edu, fahad.khan@ilc.cnr.it
francesco.mambrini@unicatt.it

Abstract

Wikibase is the software that powers Wikidata, but it can be also be used as a separate installation, to suit individual needs. This platform is ideal for creating data archives that can easily interact with the semantic web through the use of open standards; compared with other software solutions, it offers unique features, such as the option to manually edit every single semantic triple in a graphical interface. However, integrating various data models and vocabularies in Wikibase is a challenging task due to specialties in the data model. This study sheds light on modeling datasets in Ontolex-Lemon – the Lexicon Model for Ontologies, as one of the predominant and prevailing ontologies in lexicography – in Wikibase. We discuss some of the major issues that should be taken into account for remodeling Ontolex-Lemon on Wikibase, looking at two use cases dealing with Latin and Kurdish lexical data, respectively. We believe that our approach paves the way for further conversions in the future and towards a set of general guidelines.

1 Introduction

Wikibase[1] as an extension of MediaWiki is the software underlying Wikidata[2] (Vrandečić and Krötzsch, 2014), a very large knowledge graph maintained by the community of Wikidata users, and technically supported by Wikimedia Deutschland.[3] In addition to being an ontology of concepts with properties relating them to each other and pointing to typed literal values, or to external identifiers, Wikidata also contains descriptions of lexemes of multiple languages (for statistics as of 2020, see Nielsen, 2020).[4]

Although the data model for lexical data underlying Wikibase, illustrated in Fig. 1,[5] is based on the Ontolex-Lemon (McCrae et al., 2017; Declerck, 2018) core classes, i.e. `ontolex:LexicalEntry`, `ontolex:LexicalSense` and `ontolex:Form`, the fine-grained implementation of the former differs substantially from the latter. Thus, a thorough review and re-modeling of the existing datasets is required.

In this paper, we describe the data model used in Wikibase for the representation of lexemes along with two use cases that aim to adapt lexical resources modeled according to Ontolex-Lemon in a way that they can be uploaded to a Wikibase. Our use cases focus on data in Latin and Kurdish, respectively. While for the Latin data we have decided to interact with Wikidata directly, the latter case aims to use a separate Wikibase installation. Nevertheless, the final goal of the operation in both use cases is to enrich the Wikidata lexeme collection. Discussing the advantages and implications of each approach, we argue that the workflow descriptions are helpful for the creation of Wikidata-compatible lexical datasets.

2 Lexicographical Data on Wikibase

The Wikibase software comes with a default backbone ontology, the Wikibase Ontology,[6] for which widely used RDF vocabularies are deployed. For instance, the properties `rdfs:label` for entity labels, `prov:wasDerivedFrom` for references, the classes `ontolex:LexicalEntry` for lex-

[1] https://wikiba.se

[2] https://www.wikidata.org

[3] https://www.wikimedia.de

[4] For updated lexicographical coverage statistics, see https://www.wikidata.org/wiki/Wikidata:Lexicographical_coverage.

[5] See https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model and https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/RDF_mapping for a technical description.

[6] The canonical URI of the Wikibase RDF–resource description framework ontology is http://wikiba.se/ontology; the current version can be found at http://wikiba.se/ontology-1.0.owl.

emes and schema:Article for Wikipedia articles are used. Moreover, additional classes and properties are described. In the RDF representation produced by Wikibase, these classes and properties appear as such, hence called *passthrough properties*, and in the graphical interface for editing, their values have their fixed place in the page layout.

On the other hand, any additionally defined ontology concept or any additional property will be identified by a unique numeral. A Wikibase *item* with a value for rdfs:label is the minimum a user has to provide to create a concept. Such concepts appear in the canonical namespace for the corresponding Wikibase, such as http://www.wikidata.org/entity/ for Wikidata. The numeral is preceded by a capital letter: item identifiers are preceded by the letter Q, properties by a P, and as a third category, L-entities describe lexemes.

The three kinds of Wikibase entities, i.e. Q, P, and L-entities, exist uniquely on one Wikibase instance. They can themselves be further described, and linked to equivalent entities in another Wikibase such as Wikidata, or enriched with external identifiers pointing to external entities declared as equivalent. These alignments can be used for federated querying, i.e., a query that would involve Wikidata and a custom Wikibase at the same time, and, as soon as Wikibase properties are mapped to W3C-recommended vocabularies, for exporting datasets in an RDF representation compatible to the LOD cloud.

Assertions made using Wikibase P-properties are called "statements". In the RDF representation of the entity data, e.g. 'lexicography' on Wikidata (Q184524), statements are blank nodes which allow attaching the property value along with qualifiers, ranks and references (see also Fig. 3). In the editing GUI, statements appear as a central section of the entity page, with no property or value preset. A Wikibase property used in statements, qualifiers, or references is by default restricted to one datatype.[7]

The modeling of lexemes in Wikibase which is described in the following subsections along with the three core classes and the links between them deploys the Ontolex-Lemon model. The predefined backbone ontology hardly specifies anymore, leaving much space for the user to define fine-grained details of the lexicographical data model.

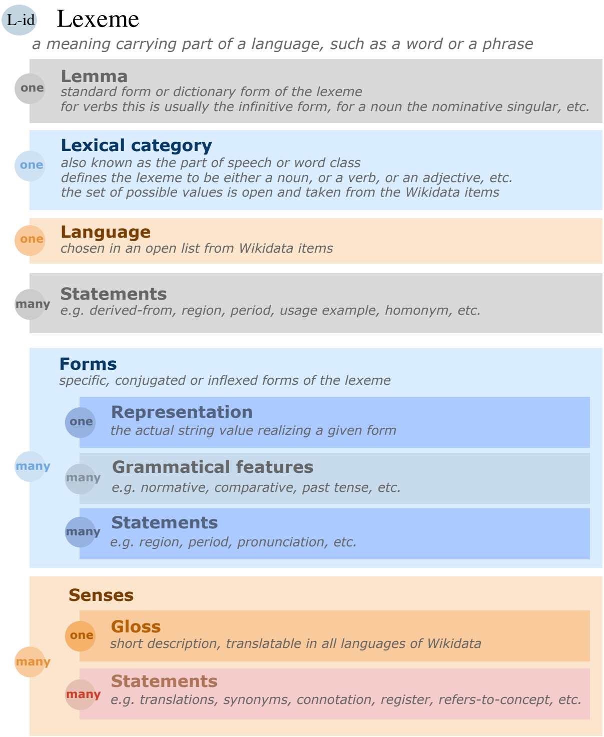
L-id
Lexeme
a meaning carrying part of a language, such as a word or a phrase
one
Lemma
standard form or dictionary form of the lexeme
for verbs this is usually the infinitive form, for a noun the nominative singular, etc.
one
Lexical category
also known as the part of speech or word class
defines the lexeme to be either a noun, or a verb, or an adjective, etc.
the set of possible values is open and taken from the Wikidata items
one
Language
chosen in an open list from Wikidata items
many
Statements
e.g. derived-from, region, period, usage example, homonym, etc.
many
Forms
specific, conjugated or inflexed forms of the lexeme
one
Representation
the actual string value realizing a given form
many
Grammatical features
e.g. normative, comparative, past tense, etc.
many
Statements
e.g. region, period, pronunciation, etc.
many
Senses
one
Gloss
short description, translatable in all languages of Wikidata
many
Statements
e.g. translations, synonyms, connotation, register, refers-to-concept, etc.

Figure 1: Wikidata Lexeme data model based on https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model

The path taken by the community around Wikidata lexemes can be conveniently explored using the ORDIA tool (Nielsen, 2019)[8]. In addition, the community also maintains user-oriented documentation with examples for the creation and querying of data.[9]

2.1 Lexical Entry

Each Lexeme entity in Wikibase is identified by a unique Lexeme ID. When users create lexemes, one or more lemmata have to be provided with a language code that specifies the language and script of each of the lemma strings (ku-arab, for example, for Kurdish in Arabic script, and ku-latn, for Kurdish in Latin script).[10] In the RDF representation of the lexeme, a lemma string appears as attached to the lexeme entity using a property called wikibase:lemma, and is additionally referred to by rdfs:label. The lemma is used for the indexation of Wikibase lexemes for ElasticSearch, akin to rdfs:label and skos:altLabel val-

[7] https://www.wikidata.org/wiki/Help:Data_type

[8] https://ordia.toolforge.org

[9] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

[10] For a list of available Wikimedia language codes, refer to https://www.wikidata.org/wiki/Help:Wikimedia_language_codes/lists/all.

ues used for that in the case of items and properties. That way, users can find lexemes performing a textual search in the GUI or via API without further descriptions.

The form of the lexeme used as a lemma would typically be described as ontolex:Form attached to properties that describe morphological or other features. Also, values for wikibase:lexicalCategory and dct:language are required; both values are restricted to be of the same Wikibase instance and to describe parts of speech and natural languages, respectively. Other properties, such as those that describe pronunciation, etymology, or usage examples, are attached to the lexeme using Wikibase statements, i.e., using P-entities as properties. A lexeme is linked to senses using ontolex:sense, and to forms using ontolex:lexicalForm.

2.2 Lexical Sense

A Wikibase sense is identified by a numeral identifier preceded by the identifier of the corresponding lexeme, a dash, and the letter S. For example, wd:L3257-S1 for the sense of the English noun *apple* referring to “a fruit of a tree of the genus Malus”. Lexeme senses are by default described using textual sense glosses in any language. Those glosses appear attached to the Sense entity using the skos:definition property; the language of the gloss text is again specified by a language code. Any further description of the sense is done using Wikibase statements.

2.3 Lexical Form

The naming of Form URI is similar to that of sense entities, but using the letter F, as in wd:L3257-F2 for the plural form of ‘*apple*’ in English. The written representation is pointed to using ontolex:representation. Items describing grammatical features of a form are linked to using wikibase:grammaticalFeature. All other description is made through custom statements.

3 Ontolex-Lemon

The differences in ontolex:LexicalEntry, ontolex:LexicalSense and ontolex:Form in the OntoLex guidelines from those described above are not significant, but it is important to be aware of them to carry out mappings from OntoLex to Wikibase.

To start with, in OntoLex each entry must be associated with at least one instance of ontolex:Form via the property ontolex:lexicalForm, and can be associated with at most one lemma form via the functional property ontolex:canonicalForm. Each ontolex:LexicalEntry is effectively associated with a language via the rdf:langString language code tag on its form representations of which it must have one. Form representations are linked via the ontolex:writtenRep property (defined as a subproperty of ontolex:representation, which is used on Wikibase for that purpose). That said, it is not required that the language specification be done via the dct:language property. Moreover, there is no requirement for lexical category information to be associated with individual instances of ontolex:LexicalEntry; in such cases where this information is provided, the lexinfo vocabulary is recommended. There are no naming conventions presupposed by OntoLex for ontolex:LexicalSense and ontolex:Form URI.

Table 1 shows the mapping of the Wikibase *passthrough* properties by default used for lexemes and Ontolex-Lemon.

Ontolex-Lemon	**Wikibase**
rdfs:label	
ontolex:writtenRep	wikibase:lemma
ontolex:sense	ontolex:sense
(skos:definition)	skos:definition
ontolex:lexicalForm	ontolex:lexicalForm
ontolex:writtenRep	ontolex:representation
(lexinfo properties)	wikibase:grammaticalFeatures

Table 1: Wikibase default properties mapping

For other properties with no mapping, some particularities have to be considered. As explained above, an OntoLex entry is linked to *one* canonical form, the written representation of which is to be mapped to wikibase:lemma. On the other hand, it is common in OntoLex datasets that the lemma string is attached to the entry entity using rdfs:label; the same is always true for Wikibase lexeme RDF representations. A Wikibase lexeme can have multiple values for wikibase:lemma (see §2.1).

In Ontolex, sense-defining definitions are attached to the ontolex:LexicalConcept class, which is linked to the sense using ontolex:lexicalizedSense. OntoLex does not specify what property to use here. In Wikibase, sense short definitions, called *gloss*, are attached directly to the sense using skos:definition. Properties from the lexinfo vocabulary that describe morphosyntactic features of forms, i.e., sub-properties of lexinfo:morphosyntacticProperty like e.g. lexinfo:number, are all to be mapped to wikibase:grammaticalFeature, and their values, on Wikibase, need to be Wikibase items; as a consequence, lexinfo concepts need to be mapped to Wikibase items.

4 Use Case 1: Kurdî Wikibase

As a low-resourced and under-represented language, Kurdish faces many challenges in language technology due to a paucity of data. To remedy this, Azin and Ahmadi (2021) and Ahmadi et al. (2019) address the creation of lexicographical resources compatible with semantic technologies, particularly by relying on Ontolex-Lemon. An entry in Ontolex-Lemon in these resources is provided in Figure 2. As such, there are four resources freely available under an open source license for Kurdish varieties.[11] The resources are described as follows:

- Northern Kurdish (also known as Kurmanji, kmr): Over 4,000 headwords are provided in Northern Kurmanji in the Latin-based script. Headwords are defined with part-of-speech tags, grammatical gender, and glosses based on distinct senses in Northern Kurdish and English. Usage examples are also provided in some cases.
- Central Kurdish (also referred to as Sorani, ckb): Over 5,000 headwords are provided in Central Kurdish written in the Latin-based script. This script, unlike Northern Kurmanji, is not much used by Central Kurdish speakers; the Perso-Arabic-based script is mostly used for this variant. Entries are described with part-of-speech tags, glosses in English and, sometimes, usage examples. Grammatical gender is not present in Central Kurdish.
- Southern Kurdish (sdh): This resource contains over 11,000 headwords, the highest number among the selected resources. The headwords are written in both Perso-Arabic and Latin-based scripts and are described with glosses in Persian and other varieties of Kurdish. Such varieties include words from Kurdish varieties along with Laki and Luri languages. That said, the distinction of the varieties is not explicit in the resource. Therefore, Kurdish glosses in this resource are specified with the ku code as an umbrella code to refer to all varieties of Kurdish. It should be noted that Luri (ldd) is a distinct language from Kurdish.
- Gorani (also known as Hawrami, hac): In comparison to the other resources, this resource is the smallest one containing around 1,000 headwords written in the Latin-based script and described with part-of-speech tags, grammatical gender, glosses in English and a few usage examples. Similar to Central Kurdish, this language is mostly written in the Perso-Arabic-based script of Kurdish.

Among the selected resources, those of Southern Kurdish and Gorani are especially important as they are relatively more under-resourced than Northern and Central Kurdish. While the two latter are also available on Wiktionary[12] facilitating community support, Southern Kurdish and Gorani are barely available for such initiatives.

Nevertheless, these resources have not been much used by the native speakers' community, due to chiefly what we believe is the lack of familiarity with LOD. Furthermore, any static resource compatible with LOD requires a SPARQL endpoint to be queried and effectively integrated into other applications. This further hinders the interoperability of the resources as individual efforts do not necessarily adapt to a larger scale usage.

To remedy this, we create an instance of the Wikibase software where the existing Kurdish resources are made available. To that end, we convert the resources into a Wikibase-compatible format requiring further modeling and modification of the lexical data. The Wikibase is accessible at https://kurdi.wikibase.cloud. In addition to the modeling described in §2, we provide information on the modeling and conversion of the selected resources as follows.

4.1 Data Remodelling

One of the major ongoing issues in creating Kurdî Wikibase is the lack of language codes sdh and hac respectively for Southern Kurdish and Gorani

[11] https://github.com/sinaahmadi/KurdishLexicography

[12] https://ku.wiktionary.org

```
:lex_kmr_7802180323 a
    ontolex:LexicalEntry, ontolex:Word ;
 ontolex:canonicalForm
     :form_kmr_7802180323 ;
 dct:language
     <www.lexvo.org/page/iso639-3/kmr> ;
 rdfs:label "partî"@kmr-latn ;
 lexinfo:partOfSpeech lexinfo:noun ;
 ontolex:sense :kmr_8301494711_sense ;
 lexinfo:gender lexinfo:feminine .

:form_kmr_7802180323 a ontolex:Form ;
 ontolex:writtenRep "partî"@kmr-latn ;
 lexinfo:number lexinfo:singular .

:kmr_8301494711_sense a
    ontolex:LexicalSense;
 skos:definition "political party"@en.

:kmr_8301494711_sense ontolex:usage [
 rdf:value "partiyên kurd yên
     siyasî"@kmr-latn ;
 rdf:value "Kurdish political
     parties"@en ] .
```

Figure 2: An example entry from the Northern Kurdish data in Ontolex-Lemon

on Wikibase. To tackle this, we use `ku` as the language code for all the varieties, but point to an item describing the variety using `dct:language`. This can be further refined once language codes are added to Wikibase.[13]

In addition to language codes, we also face specific challenges related to the lexicographical data of the sources, particularly in orthographic normalization using Latin vs. Perso-Arabic scripts and spelling variations. This is of importance to LOD technologies given that duplicated entries, i.e. several entries that describe the same lexeme, should be avoided. Therefore, we verify and unify scripts among the resources to conform with the orthographies that are widely used, e.g. ë [ʕ] for a glottal stop, is replaced with ‘‘’. Moreover, some of the headwords in the selected resources contained punctuation marks, which are removed.

Usage examples on Kurdî Wikibase are attached to a sense, and described with their English translation, while on Wikidata, usage examples are attached to a lexeme, qualified with their subject sense. On the one hand, that modeling corresponds to the OntoLex source where senses point to usage examples via `ontolex:usage` and, on the other, this made the upload process more convenient, since a usage example attached to a lexeme could not be qualified with the URI of its subject sense until that sense would get an identifier, which doesn't happen until the item data is written on the Wikibase. When transferring to Wikidata, we attach usage examples to lexemes using `wd:P5831`.

4.2 Transfer to Wikibase

While Wikibase data output is available in RDF and in a JSON format[14], the format required for upload is the latter. We have used python modules for parsing the source RDF Turtle files,[15] and for producing and uploading data in the required JSON representation.[16] The mapping of source URI to Wikibase URI is hardcoded in the script, but could also be taken from Wikibase, since URI alignments are also represented there.[17]

4.3 Federation with Wikidata

Since our focus on this project is on creating a Wikibase for Kurdish with minimal manual manipulation and verification of the data, we aim to raise awareness in the related community to contribute to Kurdî Wikibase. This way, not only the existing data can be checked, but also further completed by missing senses and headwords. This is, in fact, a chance for such under-represented communities to promote their language on their own Wikibase. Ultimately, this results in cleaner data without creating uncertain or noisy material on Wikidata. After the community is trained on their own Wikibase, members would most probably go on editing Wikidata, when the data is transferred to that global platform.

Once curation tasks are done, transferring the Kurdish lexical data from Kurdî Wikibase to Wikidata will be trivial, as long as all implied URI are aligned; that is done using a Wikibase property of type *external identifier*, which points to the equivalent Wikidata entity, `kdb:P1` in our case. Some specialties of our Wikibase data have to be taken into account, e.g. the different locations of usage examples. Open issues to be solved are only two: How to model the English translations of usage examples, and the already mentioned lack of certain language codes. Both issues are already addressed

[13] We have filed a request for inclusion of these codes in subsequent releases of the Wikibase software.

[14] https://www.mediawiki.org/wiki/Wikibase/DataModel/JSON

[15] RDFLib: https://rdflib.readthedocs.io

[16] WikibaseIntegrator: https://github.com/LeMyst/WikibaseIntegrator

[17] Details at https://kurdi.wikibase.cloud/wiki/Project_Log

in Wikidata lexemes community discussions. Fig. 3 shows a modeling proposal for an example Kurmanji entry on Wikidata.

```
wd:L1083983 a ontolex:LexicalEntry;
wikibase:lemma "partî"@kmr-latn ;
rdfs:label "partî"@kmr-latn ;
wikibase:lexicalCategory Q1084 ; # noun
dct:language wd:Q36163 ; # Kurmanji
ontolex:sense wd:L1083983-S1 ;
ontolex:lexicalForm wd:L1083983-F1 ;
wdt:P5185 wd:Q1775415 ; # gender fem.
p:P5831 [ps:P5831 "partiyên kurd yên siyasî"@kmr-latn ;
pq:P2441 "Kurdish political parties"@en ;
pq:P6072 wd:L1083983-S1] . # usage example

wd:L1083983-S1 a ontolex:LexicalSense ;
skos:definition "political party"@en ;
wdt:P5137 wd:Q7278 . # item-for-this-sense political party

wd:L1083983-F1 a ontolex:Form ;
ontolex:writtenRep "partî"@kmr-latn ;
wikibase:grammaticalFeature wd:Q110786 . # singular
```

Figure 3: An example entry from the Northern Kurdish data in Wikidata (proposal)

5 Use Case 2: Latin and Wikidata

5.1 The LiLa Project

Latin is a widely attested and studied language: its written attestations go from the earliest inscriptions (variously dated from the 7th to the 5th century BCE) until nowadays (as Latin is the official language of the Vatican State). In the span of more than two millennia, Latin has been spoken by several peoples in Europe; it outlived the political domination of Rome over the Mediterranean and part of Central Europe as the main international language of culture for centuries. This resulted in a corpus of texts that show a large diatopic, diastratic, and diaphasic variation, as well as a covering of a wide range of genres. Notably, Latin, a part of the Indo-European family thus historically related to many widely attested idioms of the world, is also the direct ancestor of the Romance languages, spoken in many countries of Europe and of the world.

In the last decades, several lexical and textual digital resources have been individually developed for Latin, often as retro-digitization of earlier printed sources. Most of these resources are the product of separate efforts by the research community across multiple decades and waves of digitization campaigns. While their existence is of great importance, they do not rely on common vocabularies and ontologies, nor do they provide a standard query language to access the data. Such a situation undermines their effectiveness as tools for the researchers and the general public.

The *LiLa:Linking Latin* project[18] was started precisely to address this lack of interoperability with the help of Semantic Web technologies. The goal of the project is to connect the Latin lexical and textual resources currently available on the web by describing them with a common set of vocabularies and a shared data model. A parallel aim is also to foster, by adopting a common model for data representation, interoperability with the resources for the other languages of the Indo-European and of the Romance family. Particularly in the field of etymology and language contact, LiLa aims at representing phenomena like inheritance (Mambrini and Passarotti, 2020) and borrowing (Franzini et al., 2020).

A key process that, for a morphologically rich language like Latin, concerns both lexical and textual resources is lemmatization. Lexicons are indexed using canonical forms of citation for lexemes; lexical research in Latin corpora is only possible with lemmatized text. Therefore, the core of the LiLa is represented by the LiLa Knowledge Base (LKB), a collection of Latin word forms that are conventionally used to lemmatize corpora and index lexica (Passarotti et al., 2020). The LKB is based on Ontolex-Lemon: lemmas are defined as instances of the class `ontolex:Form` and are described according to a series of morphological features, like part of speech, using the list of the "upos" of the Universal Dependencies project (de Marneffe et al., 2021), and grammatical gender (used for nouns only). Nouns, verbs, pronouns and adjectives are also classified for their inflection type, adopting the traditional classes used in Latin grammars. Finally, about 47,000 lemmas of Classical Latin also register information on the prefixes and suffixes that can be identified in their formation, derived from the data compiled for the *Word Formation Latin* lexical resource (Pellegrini et al., 2022).

The list of lemmas, as well as the morphological analyses of each of them, is derived from the morphological analyzer Lemlat (Passarotti et al., 2017). Currently, the LKB includes ca. 200,000 lemmas.

[18] https://lila-erc.eu

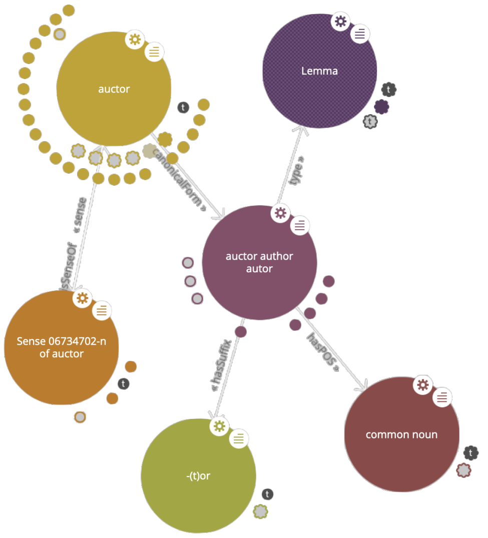

Figure 4: The lemma '*auctor*' ('maker, author') in the LKB

Starting from the LKB, LiLa is now connecting a growing collection of lexical resources (Passarotti and Mambrini, 2021) and textual corpora (Mambrini et al., 2022b; Fantoli et al., 2022). The former, all described with the Ontolex-Lemon model, collect a series of instances of `ontolex:LexicalEntry` to the lemmas in the LKB via the `ontolex:canonicalForm` property. These lexica include a manually revised version of the Latin WordNet, a valency lexicon (Mambrini et al., 2021), and a Latin-English dictionary (Mambrini et al., 2022a).

Figure 4 visualizes one lemma (*auctor* 'maker, author') in the LKB. The figure represents only a limited set of connections: the lemma (in the center) is linked via `ontolex:canonicalForm` to a lexical entry of the Latin WordNet,[19] which is in its turn connected to a sense. The lemma is then joined to the nodes representing the suffix *-(t)or* and the part of speech "common noun" (NOUN in the UD POS tagset).

5.2 Data Remodelling

Before we began our work to integrate information from the LKB, the Wikidata lexeme collections included 32,183 entries in Latin. Most of them originated from the word list used in William Whitaker's WORDS software for morphological analysis of Latin forms.[20] While there were overlaps with this preexisting collection, the LKB held a substantial amount of additional materials. Our work was split into two subtasks: 1. aligning the two resources for the preexisting words; 2. setting up a workflow for the integration of new Wikidata Latin lexemes from the data in the LKB.

The 32,183 preexisting lexemes were variously distributed across 25 values of `wikibase:lexicalCategory` some of which – like "hapax legomenon" (`wd:Q168417`), or "letter" (`wd:Q9788`) – had no correspondence to LiLa's parts of speech. In particular, phrases, idioms, and other multi-word expressions did not have any equivalent in the LKB. Prefixes and affixes used in word formation processes were also among the Wikidata Latin lexemes. Although, as said, the LKB includes affixes, they are currently not aligned with the OntoLex ontology (Passarotti et al., 2020, 191); we decided therefore not to consider them in our first alignment.

The preexisting lexemes were matched both by lemma string (comparing the `ontolex:writtenRep` of the LKB lemmas and the lemma string stored as the data property `wikibase:lemma`), and the POS (manually mapping Universal Dependencies POS to the corresponding Wikidata item to use as the value for `wikibase:lexicalCategory`, whenever possible). 26,379 lexemes (81.97% of the preexisting Latin words in Wikidata) were matched univocally to one lemma in the LKB; 1,356 (4.21%) were ambiguous (matching more than one lemma in the LKB), while 4,448 lexemes (13.82%) were not found.

Of the latter, more than 1,000 entries could be further matched using simple rules of alignment for POS and lexical categories (e.g. nouns of Wikidata and proper nouns of the LKB). In total, we aligned 27,486 entries; these entries in Wikidata are now linked to the corresponding lemmas in the LKB via the special property "LiLa Linking Latin URI" (`wdt:P11033`, see §5.3). The work of semi-automatic disambiguation and matching of the remaining lexemes is still in progress.

The lexical database of the Lemlat analyzer, from which, as said, the LKB was originally derived, aggregated lemmas from three classes of sources: a dictionary for Classical Latin, an Onomasticon of proper nouns, and a dictionary of Medieval Latin (*cf.* Passarotti et al., 2017). In our first experiment on expanding the Wikidata Latin

[19] `http://lila-erc.eu/data/lexicalResources/LatinWordNet/id/LexicalEntry/l_90615`.

[20] `https://latin-words.com`

lexeme inventory, we decided to concentrate on the lexicon of Classical Latin (which was already covered, though not exhaustively, by the collection from Whitaker's WORDS).

We identified 24,007 additional lemmas belonging to the Classical base of the LKB that were not present in Wikidata. We proceeded to create the new lexemes, with the mandatory information of the lemma string (mapped onto the `rdfs:label` of the LKB lemma), the lexical category and the grammatical gender for nouns.

Only a handful (49) of the preexisting Latin lexemes in Wikidata had information on the inflection type, as a paradigm class (`wdt:P5911`) or conjugation class (`wdt:P5186`). These properties link verbs and nouns to, respectively, the 5 declensions and 4 conjugations of traditional grammars (e.g. `wd:Q3921592` for the Latin first declension of a-stems). As said, LiLa includes comprehensive information on inflection types. We consider these data particularly relevant for enriching the Latin lexemes, as the classification of the lexemes into morphological types can potentially be used to support disambiguation of homographs (e.g. *dico* "proclaim, dedicate", 1st conjugation, vs *dico* "say", 3rd conjugation), or automatic form generation.

We proceeded to align the 53 inflection types used in LiLa to the relevant entities in Wikidata. In 5 cases, we have created new Wikidata items, since a few classes in LiLa (primarily for uninflected or invariable words) did not have any match in Wikidata, or we have added English labels to the preexisting entries (which, as in the case of `wd:Q3606519`, had only an Italian one). Some examples of this alignment are reported in Table 2. The classification in the LKB is more fine-grained, as LiLa distinguishes several sub-classes of the traditional declensions and conjugations; for instance, LiLa includes a subdivision of irregular nouns of the 2nd declension, or the impersonal verbs of the 3rd conjugation, which are mapped to the general Wikidata categories for 2nd declension and 3rd conjugation respectively (see the ex. in Table 2).

The enrichment of the Latin lexemes with the information on inflection taken from LiLa is currently ongoing; as well as the alignment of Latin lexeme senses with Wikidata ontology items using `wd:P5137`, using Princeton Wordnet synset ID as pivot.[21]

5.3 Transfer to Wikidata

Since, in this case, we are interacting directly with Wikidata, we have gone through the process established by the community to create a new *external identifier* property for the linking to the LKB, and for the batch writing bot permission request. We have uploaded the data using python modules (see §4.2), and have documented the upload process.[22]

6 Conclusions

The published datasets for Gorani and Southern Kurdish varieties are unique as those are rarely represented on the web, Together with the Sorani and Kurmanji data, they are now accessible and reusable as part of the Wikibase ecosystem, and, after completing the addressed curation tasks, prepared for transfer to Wikidata.

As for Latin, Wikidata now includes 56,202 lexemes, 51,492 of which now provide a link to the LKB via the property `wdt:P11033`. For these lexemes, a federated query over the LiLa SPARQL endpoint[23] already gives access to the wealth of information provided in the LiLa resources. Starting from a lexeme in Wikidata, for instance, it would be possible to retrieve all the occurrences of the words lemmatized under the connected lemma in the collection of LiLa corpora.

Lexemes in Wikidata are typically enriched with their inflected forms (which are available, for instance, for those obtained from Withaker's WORDS) and their senses. Currently, LiLa does not link the canonical forms in the LKB to any other forms of the same word; in other words, it will not be possible to use LiLa to retrieve an exhaustive list of the forms of a lexeme (unless the scope is limited to the forms attested in the LiLa corpora and lemmatized under any given lemma). Plans to create a lexical resource linked to LiLa with all possible inflected forms are, however, under development. In the near future, therefore, it will be possible to use LiLa to enhance the catalog of forms for the Latin lexemes.

The repertoire of senses, on the contrary, is already available for a limited number of Latin words. The 53,437 lexical entries from the Lewis and Short Latin-English dictionary (*cf.* Mambrini et al., 2022a) and the 6,269 entries of the Latin Word-

[21] This is mentioned in section 6; In the camera-ready version of this paper, we will most probably be able to report on that as finished tasks.

[22] See `https://www.wikidata.org/wiki/Property_talk:P11033`.

[23] `https://lila-erc.eu/sparql`

LiLa		Wikidata	
URI	Label	URI	Label
lila:n2	2nd declension (m/f) nouns	wd:Q3953983	2nd decl.
lila:n2e	2nd decl irregular nouns	wd:Q3953983	2nd decl.
lila:n6	1st class adj.	wd:Q3606519	Adj. of 1st class
lila:v3r	3rd conj. verb	wd:Q54295441	Latin 3rd conj.
lila:v3e	3rd conj. impersonal verb	wd:Q54295441	Latin 3rd conj.
lila:n	uninflected noun/adj.	—	—

Table 2: Mapping of LiLa and Wikidata Latin inflection classes

Net are all provided with definitions and senses. In particular, for the latter, links to the synsets in the Princeton Wordnet are available. Since Wikidata concepts are linked to Princeton Wordnet using wd:P8814, for the intersection of these with the Princeton Wordnet ID in Latin Wordnet, we will use these existing alignments for setting wd:P5137 relations between Latin lexeme senses and Wikidata concepts.

We are recording all direct alignments between OntoLex and lexinfo URI to Wikidata defined in the use cases presented in this paper, as well as indirect correspondences (those that imply some re-modeling), and plan to expand these towards all entities in the OntoLex and lexinfo ontologies, to provide general guidelines for transferring OntoLex datasets to Wikidata.

References

Sina Ahmadi, Hossein Hassani, and John P. McCrae. 2019. Towards electronic lexicography for the Kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*, pages 881–906, Sintra, Portugal.

Zahra Azin and Sina Ahmadi. 2021. Creating an Electronic Lexicon for the Under-resourced Southern Varieties of Kurdish Language. *Proceedings of Seventh Biennial Conference on Electronic Lexicography (eLex 2021)*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Thierry Declerck. 2018. Towards a Linked Lexical Data Cloud based on OntoLex-Lemon. In *Proceedings of the LREC 2018 Workshop "6th Workshop on Linked Data in Linguistics LDL-2018"*, Miyazaki, Japan.

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.

Greta Franzini, Federica Zampedri, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2020. Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Seventh Italian Conference on Computational Linguistics. Bologna, Italy, March 1-3, 2021*, pages 1–6, Bologna. CEUR-WS.org.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2022a. Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In Elisabetta Fersini, Marco Passarotti, and Viviana Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021*, pages 214–220. Accademia University Press.

Francesco Mambrini and Marco Passarotti. 2020. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, pages 20–28, Paris. European Language Resources Association (ELRA).

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Further with Knowledge Graphs. Studies on the Semantic Web 53*, Amsterdam. IOS Press.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022b. The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4022–4029, Marseille, France. European Language Resources Association.

John 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: Lexicography from scratch. Proceedings of eLex 2017*, pages 587–597, Brno. Lexical Computing CZ s.r.o.

Finn Nielsen. 2020. Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 82–86, Marseille, France. European Language Resources Association.

Finn Årup Nielsen. 2019. Ordia: A Web Application for Wikidata Lexemes. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events*, volume 11762, pages 141–146. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, volume 133, pages 24–31, Gothenburg. Linköping University Electronic Press.

Marco Passarotti and Francesco Mambrini. 2021. Linking latin: Interoperable lexical resources in the lila project. In *Building new resources for historical linguistics*, pages 103–124, Pavia. Pavia University Press.

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, 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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57:78–85.