LANGUAGE, DATA and KNOWLEDGE 2025

Proceedings of the 5th Conference on **Language, Data and Knowledge: Workshops**

UniorPress
Naples
2025

PROCEEDINGS OF THE 5TH CONFERENCE ON LANGUAGE, DATA AND KNOWLEDGE: WORKSHOPS

EDITORS:

Katerina Gkirtzou, ILSP “Athena” Research Center, Greece
Slavko Žitnik, University of Ljubljana, Slovenia
Jorge Gracia, University of Zaragoza, Spain
Dagmar Gromann, University of Vienna, Austria
Maria Pia di Buono, University of Naples “L’Orientale”, Italy
Johanna Monti, University of Naples “L’Orientale”, Italy
Maxim Ionov, University of Zaragoza, Spain

ISBN 978-88-6719-334-9

UniorPress — University of Naples “L’Orientale”
Via Nuova Marina 59 — 80133 Napoli (Italy)

UniorPress

Table of Contents

Fifth OntoLex Workshop

Fifth OntoLex Workshop

Fifth OntoLex Workshop

The Fifth OntoLex workshop, which takes place in Naples, Italy on September 9th, 2025, is the latest in a series of events dedicated to the OntoLex-Lemon vocabulary and which has previously been held in conjunction with the "Language, Data and Knowledge" conference series in Leipzig (2019), Zaragoza (2021), and Vienna (2023) and as a stand-alone event in Leiden (2018).

Having established itself as a pivotal standard for representing lexical data as linked data, the use of the OntoLex-Lemon model has recently expanded into diverse domains, including computational lexicography, terminology management, and AI-driven language technologies. However, as new linguistic challenges and new technological advancements emerge, it is imperative that the OntoLex model continue evolving to meet these needs. With this in mind, the aim of the OntoLex workshop is to provide a forum for sharing ideas and proposals in order to meet these new challenges and to ensure that OntoLex keeps pace with the latest developments. Overview: In the morning session, the workshop features a series of invited talks by experts in the area, this is followed by six contributed papers which have been peer-reviewed by the program committee (see below); these are as follows. The paper Inferring Adjective Hypernyms with Language Models"(Augello and McCrae) explores how masked language models can detect semantic relations like hypernymy between adjectives, leveraging Open English WordNet. Philosophising Lexical Meaning as an OntoLex-Lemon Meta-Ontology"(Zamborlini, Zhu, van Erp and Betti) introduces a conceptual framework to extend OntoLex-Lemon on a philosophically informed basis. Bringing IATE into the Semantic Web Family"(Diez-Ibarbia, Martín-Chozas and Montiel-Ponsoda) reports on ongoing efforts to convert the EU's Interactive Terminology for Europe (IATE) into an RDF-compliant resource, contributing structured terminology to the Linked Open Data cloud. OntoLex-Lemon in Wikidata and other Wikibase instances"(Lindemann) discusses the mapping of lexical data to Wikidata entities using OntoLex-Lemon, describing methods for aligning linguistic annotations and glosses with multilingual data stores, including practical challenges in using Wikibase software. The article A Lightweight String Based Method of Encoding Etymological Information in RDF"(Khan, Ionov, Marongiu and Salgado) presents a strategy for representing etymological content as RDF string literals using a well-defined regular grammar. This method supports a shallow description of lexical histories that can be queried using the SPARQL regular expression function. Finally Ontologies for historical languages: using the LiLa and OntoLex-Lemon framework to build a Lemma Bank for Old Irish"(Fransen) focuses on historical language data, showing the use of OntoLex in developing a lemma bank for Old Irish, similar to the one created in the Linking Latin project.

The afternoon session of the workshop, which is also open to remote participation by the members of the W3C Ontology-Lexicon group, provides an overview of the latest advancements in the OntoLex model. It features an introduction to preparatory work on a forthcoming 2.0 version of OntoLex, status updates on the development of the FrAC and Morph modules, software demos and proposals for new modules, e.g. for etymology and conceptual modelling.

Fahad Khan, John McCrae, Matteo Pellegrini and Philipp Cimiano
The OntoLex 2025 Organising Committee

Organizing Committee

Anas Fahad Khan, Istituto di Linguistica Computazionale A. Zampolli, Consiglio Nazionale delle Ricerche, Pisa, Italy
John Philip McCrae, Data Science Institute, National University of Ireland Galway, Galway, Ireland
Matteo Pellegrini, Università Cattolica del Sacro Cuore, Milan, Italy
Philipp Cimiano, Bielefeld University, Bielefeld, Germany

Program Committee

Maxim Ionov, University of Zaragoza
Sina Ahmadi, University of Zurich
Patricia Martín-Chozas, Universidad Politécnica de Madrid
Katerina Gkirtzou, ILSP “Athena” Research Center, Greece
Gilles Sérasset, LIG - Université Grenoble Alpes
Francis Bond, Palacký University
Christian Chiarcos, University of Augsburg
Ana Salgado, NOVA Lisbon
Elena Montiel-Ponsoda, Universidad Politécnica de Madrid
Adrian Doyle, University of Galway
Paul Buitelaar, University of Galway
Sara Carvalho, University of Aveiro
Jorge Gracia, University of Zaragoza
Manuel Fiorelli, University of Rome, Tor Vergata

Inferring Adjective Hypernyms with Language Models to Increase the Connectivity of Open English Wordnet

Lorenzo Augello[*]
Università Cattolica del Sacro Cuore,
Milan, Italy.
lorenzo.augello01@icatt.it

John P. McCrae
Insight Research Ireland
Centre for Data Analytics,
University of Galway,
Galway, Ireland.
john@mccr.ae

Abstract

Open English Wordnet is a key resource published in OntoLex-lemon as part of the linguistic linked open data cloud. There are, however, many links missing in the resource, and in this paper, we look at how we can establish hypernymy between adjectives. We present a theoretical discussion of the hypernymy relation and how it differs for adjectives in contrast to nouns and verbs. We develop a new resource for adjective hypernymy and fine-tune large language models to predict adjective hypernymy, showing that the methodology of TaxoLLaMa can be adapted to this task.

1 Introduction

Open English Wordnet (McCrae et al., 2019a, OEWN) is an open-source fork of the Princeton WordNet (Fellbaum, 1998), which is modelled as and released as OntoLex data (McCrae et al., 2017). This work aims to continue the work of maintaining the resource, as well as providing a central resource for linked data resources to connect to. However, OEWN is not itself a completely connected graph, as recent work on the verb hierarchy has shown (McCrae, 2025). For adjectives and adverbs, there are a very large number of synsets[1] that are not linked and this reduces the effectiveness of this resource as a central resource in Linguistic Linked Open Data (LLOD) cloud (McCrae et al., 2016).

Recent advances in large language models have significantly reshaped the field of lexical semantics (Moskvoretskii et al., 2024b). These models, trained on huge corpora, have demonstrated an emerging ability to infer nuanced lexical relations, and in tasks of taxonomy extraction and hypernymy discovery (Bordea et al., 2016). Our goal is to use this to fully connect the graph of OEWN 2024, however, the performance of these models has been predominantly evaluated on nouns and verbs, leaving open the question of whether they can meaningfully capture a complex relation such as hypernymy among adjectives, which is underrepresented in existing datasets and lexical resources.

Hypernymy has been extracted from large text corpora and modelled in widely used benchmarks for lexical relations classification such as BLESS (Baroni and Lenci, 2011), EvaLution (Santus et al., 2015) and HyperLex (Vulić et al., 2017). Yet, its application to adjectives remains considerably underexplored, despite its potential theoretical and practical relevance. Unlike noun and verb hypernymy, which often relies on well-defined hierarchical taxonomies, hypernymy between adjectives is inherently more fluid and context-dependent, influenced by issues such as polysemy, gradability, and contextual ambiguity (Kennedy, 2007), which are often investigated by existing resources only from other relations' points of view (Murphy, 2003).

Understanding hypernymy among adjectives is then critical and in order to explore its issues and despite their acknowledged complexity, our aim is to establish a clearer theory of adjective hypernymy that can offer insights into the broader organization of lexical meaning and support the development of more semantically aware language models.

In this work, we try to address the theoretical and practical gaps surrounding hypernymy between adjectives, with two main objectives. First, we propose a theoretical framework for interpreting hypernymy among adjectives and construct a gold-standard English dataset of hyponym-hypernym pairs reflecting this relation, which is released in RDF using the OntoLex model.[2] Starting from existing lexical resources for other languages such as Polish WordNet 3.0 (Maziarz et al., 2016) and

[*] Work completed at Insight Research Ireland Centre for Data Analytics, University of Galway.

[1] 'synonym set' or 'synset' for short are equivalent to lexical concepts in the OntoLex model

[2] https://github.com/lorenzoaugello/adjective-hypernymy

Open Dutch WordNet (Postma et al., 2016), we perform careful human annotation and validation to address issues of semantic ambiguity and ensure the reliability of the dataset for English by leveraging the OEWN. Second, we explore the ability of language models to recognise and interpret adjective hypernymy. We evaluate their performance both before and after training them on our gold standard, analysing how well they can capture this complex lexical relation[3].

Examining models' performance on fine-grained and subjective semantic relations can explore and unveil their underlying linguistic competence and biases, contributing to the ongoing effort to interpret and refine their behaviour. In order for the models to be better in understanding language dynamics, we think that the development of specialized evaluation datasets is a foundational step towards benchmarking and improving their ability to disentangle complex semantic phenomena (Putra et al., 2024).

In addition to its theoretical contribution that tries to address the lack of hierarchical structure of adjectives in English, our dataset offers several practical benefits: it fills a structural gap in lexical resources, enhancing their utility in knowledge representation and reasoning tasks; it supports improved performances in NLP tasks such as lexical entailment and word sense disambiguation; it serves as a novel benchmark for evaluating LLMs understanding of adjective semantics; it creates a foundation for possible multilingual extensions, towards better integration in the LLOD cloud.

The rest of this paper is structured as follows. In Section 2, we discuss related work on hypernymy, adjective semantics, and language model evaluation, and in Section 3, we introduce the Open English Wordnet. Section 4 presents our theoretical framework, while Section 5 describes the construction of our adjective hypernymy dataset. Section 6 focuses on the evaluation of language models on this task, including both zero-shot testing and finetuning experiments. Section 7 presents the results and a summary of the findings. Finally, Section 8 concludes the paper and suggests directions for future research, followed by the limitations of this study.

2 Background and Related Work

Theorizing the nature of adjectives, their function, taxonomy, and interrelations has traditionally begun with categorizing how they modify nouns (Raskin and Nirenburg, 1995), reflecting the general agreement around the definition for which adjectives serve as "modifiers of the nouns with which they are combined" (Lyons, 1977). However, the semantic categorization of adjectives proves considerably more complex than that of nouns and verbs, as their behaviour often eludes straightforward ontological modelling (McCrae et al., 2014). While several classification attempts have been made, adjectives remain understudied in lexical semantics and lack a universal agreement on a theoretical framework. Existing classifications typically remain at a high level of general categorization, grouping adjective synsets into a few broad semantic classes known as "supersenses" to create taxonomies, as the first examples of Dixon (1982) and Hundsnurscher and Splett (1982) show.

Moving the focus to the task of lexical entailment and the specific relation of hypernymy, we face more ambiguity. Traditional definitions of hypernymy, such as the proposal that one term A is a hypernym of another term B if A's meaning covers the meaning of B or much broader (Tjong Kim Sang, 2007), are more straightforward if applied to nouns and verbs, whose hierarchical relations and structures are clearer. However, adjectives introduce challenges related to gradability, scalar structure, and context dependency (Liu et al., 2023). Unlike nouns, which can be organized in chains where each hyponym naturally inherits properties from its hypernyms (Heyvaert, 2010), adjectives resist simple hierarchical arrangement and prefer more compact and less vertical scales.
Notably, hypernymy relations between adjectives are absent from foundational lexical resources such as Princeton WordNet (Miller, 1995; Fellbaum, 1998). Instead, adjectives are organized via antonymy (e.g., wet–dry), semantic similarity (e.g., dry–arid), and pertainymy (e.g., crime–criminal), with an absence of explicit modelling of hierarchical relations that could lead to a conflation between synonymy and hypernymy in adjectival semantics (Scheible and Schulte im Walde, 2014).

The question is now around the existence of hypernymy for adjectives. Recent works suggest that lexical relations should be conceptualized not in binary terms, but rather as graded semantic phe-

[3] https://huggingface.co/collections/loraug/hyper-discovery-68123c85fef0889c6559e674

nomena, with terms represented along a continuous scale according to principles of category membership (Vulić et al., 2017). However, given the fact that most of the existing lexical resources and datasets created for lexical semantics tasks derive from the structure of the Princeton WordNet - which does not incorporate adjectival hypernymy - the representation and formalization of this relation remains underdeveloped.

Some lexical resources for other languages have taken steps towards modelling adjectival hypernymy. For example, the GermaNet (Hamp and Feldweg, 1997) abandons the cluster-based approach of Princeton WordNet, adopting instead a hierarchical structuring for adjectives similar to that of nouns and verbs, where, for example, "gut" ("good") is the hypernym of "toll" ("great"). Similar hierarchical mappings are observed in the Polish WordNet, where the original "dumbbell model" of Princeton WordNet has been transformed into a vertical hyponymy structure (Rudnicka et al., 2016). In the Open Dutch WordNet, although antonymy remains the dominant relation for adjectives, instances of hierarchical hypernymy are also introduced, such as "knotsgek, stapelgek, krankjorum, knettergek" ("very mad") as hyponyms of "gek, dwaas" ("mad") (Maks et al., 2008).

In the context of the NLP community, the hypernymy relation has been explored in many works, from its automatic extraction from text corpora (Hearst, 1992) to investigations around language models' knowledge and classification capabilities (Ushio et al., 2021; Wang et al., 2021). Various evaluation studies have been conducted on language models' performances, but they overwhelmingly focus on nouns (Camacho-Collados et al., 2018), verbs (Greco et al., 2024), or general natural language inference tasks (Madaan et al., 2024). Specific techniques such as prompting-based evaluation of hypernymy knowledge (Hanna and Mareček, 2021), dataset augmentation and benchmarking strategies (Kober et al., 2021) have been proposed, but still none of them are dedicated specifically to adjective hypernymy. Thus, while significant advances have been made in understanding lexical entailment for other parts of speech, an investigation specifically tailored to adjectives remains an unexplored field.

3 Open English Wordnet

The *Open English Wordnet* (McCrae et al., 2019b) is a comprehensive, open-source lexical resource for the English language, derived from the original Princeton WordNet (Fellbaum, 1998), and developed using the W3C OntoLex-Lemon model (McCrae et al., 2017). As an OntoLex resource, OEWN publishes lexical data on the Web in accordance with linked data principles, thereby facilitating interoperability across linguistic and semantic resources.

OEWN plays a central role in the Linguistic Linked Open Data (LLOD) cloud (McCrae et al., 2016), acting as a foundational resource for representing lexical semantics, enabling cross-linguistic comparison, and linking to other language resources. It uses the OntoLex-Lemon model as a standardized vocabulary for representing lexical information as RDF. The Global WordNet Association has developed formats for the representation of wordnets (McCrae et al., 2021), which extend the core principles of OntoLex. These formats support multiple serializations in XML, JSON(-LD), and RDF/Turtle—providing interoperability with other OntoLex resources. These formats have been adopted by OEWN and other wordnets in the Open Multilingual Wordnet (OMW) to ensure compatibility and extensibility.

In RDF, the GWA format directly encodes `LexicalEntry`, `LexicalSense`, and `LexicalConcept` using OntoLex classes, while WordNet-specific metadata and relations (e.g., sense keys, synset ordering) are modelled via a dedicated ontology at `https://globalwordnet.github.io/schemas/wn`.

4 Theoretical proposal

Even though it is true that hypernymy for adjectives is difficult to define and has unclear boundaries, we think that it is a necessary relation that could help in the semantic organization of a language by introducing a hierarchy of meaning and usage that is not limited to a simple categorization in broad classes or to fuzzy and often confused horizontal and opposition relations. The distinction with synonymy is in fact very narrow, but we think that in the scope of the same semantic field, there are adjectives with a broader meaning than others, and those need to be distinguished (e.g., cognitive - mental - immaterial; displeasing - unpleasant - negative).

The Princeton WordNet has 117,659 synsets and

84,428 hypernym-hyponym relations, but those are only for nouns and verbs, and not for adjectives and adverbs. Furthermore, there is currently no reliable dataset for the English language that represents this relation for adjectives. Facing the limited availability of high-quality training data, one of the main objectives of this study was thus the creation of a reliable initial dataset containing adjective pairs and their definitions, drawn from the OEWN, in order to address the problem of word-sense disambiguation.

Throughout the annotation and reliability evaluation of hyponym-hypernym pairs extracted from existing lexical resources, a principle of substitution was followed (see McCarthy and Navigli (2007) for other applications of lexical substitution): the hypernym should be substitutable in place of the hyponym in context, such that the resulting sentence preserves the original meaning at an acceptable general level, without contradiction or need for forced interpretation. While a semantic broadening is both expected and required, the sense of the hyponym needs to be included within the possible interpretations of the hypernym without ambiguity. Thus, hypernymy is characterized by inclusion of meaning, but not equivalence, distinguishing itself from synonymy. In the annotation process, if the connection between the two adjectives of the candidate pair was not perceived as direct and intuitive by the annotators, the pair was discarded, as we do not want dubious inference or multiple plausible readings. Given that hypernymy is inherently difficult to define - being vague and often subject to flexible and debatable boundaries - only those pairs deemed most reliable were retained (even though, undoubtedly, some may still be debated, highlighting once again the subjective nature of such a relation).

One of the main challenges in identifying a hypernymy relation between two adjectives lies in the polysemous nature of many of them. A first example is "cold", which can refer either to the semantic field of temperature or to aspects of human behaviour and personality. Another case is "hard", which in one sense refers to a physically solid material, and in another to the more abstract concept of difficulty. Adjectives' meaning and their relations are therefore defined not only in isolation but, more importantly, depending on the associated noun and the semantic domain they evoke: "cold" can describe both an environment and a person; "hard" can qualify both a material and a problem. In this regard, context plays a crucial role. While this is already a challenge for human annotators, it becomes even more complex for language models that have to deal with semantic disambiguation (see Section 5). Moreover, many English words can assume more than one part of speech depending on context and usage, without undergoing inflectional changes, as English has a very limited morphological richness. A first example is the word "clean", which can function as both a verb and an adjective. A broader and more frequent case involves present and past participles, which are at times interpreted as verbs and at other times as adjectives. This ambiguity can be extended to nouns as well (the above mentioned word "cold" can have multiple adjectival meanings, but it can also be a noun: "a mild viral infection involving the nose and respiratory passages"), and given the pervasiveness of this phenomenon in English, it was deemed necessary to incorporate OEWN definitions for both hyponyms and hypernyms into the construction of the gold standard dataset (see also the experiments of Moskvoretskii et al. (2024b) for why definitions are crucial).

In contrast to the more straightforward case of nominal hypernymy, we think that adjectival hypernymy needs to be grounded in sense disambiguation, operationalized through substitution-based inclusion tests, and evaluated in a context. Our gold-standard dataset leverages those principles and aims to provide an initial but reliable resource for further studies of adjectives' semantics and its modelling in computational systems.

5 Dataset creation

The construction of the gold-standard dataset was based on pre-existing lexical resources, leveraging wordnets from languages that explicitly encode adjectival hypernymy relations: the Open Dutch WordNet (ODWN) and the Polish WordNet (plWN). These two resources were chosen because, unlike the Princeton WordNet, they organize adjectives hierarchically, making them suitable starting points for our purposes.

Hyponym-hypernym pairs of adjectives were automatically extracted from those two resources by using the wn Python library (Goodman and Bond, 2021), and an initial pool of 450 pairs was randomly picked from the total (166 from ODWN and 284 from plWN). Only adjectives were considered, and both child hyponym and father hypernym nodes were included. Each extracted pair was then manually translated into English using bilingual online dictionaries (Wiktionary and the Cambridge Dictionary), consistently selecting the first suggested translation to minimize subjective bias and arbitrary choice.

After this, each pair was reviewed individually by two annotators. If both adjectives of the pair appeared in the same synset within the OEWN - indicating synonymy rather than a hierarchical relation - the pair was discarded and the original relation was kept: for example, “difficult” and “hard”, which are hypernym-hyponym in plWN, were not included because they are synonyms in OEWN (“not easy; requiring great physical or mental effort to accomplish or comprehend or endure”). When multiple hypernyms were available for a hyponym, if they belonged to the same OEWN synset they were kept; otherwise, only the most reliable one was chosen and agreed upon by the annotators, while the others were discarded.

So, starting from the initial 450 pairs, 148 were discarded, and the final gold standard for English comprises 302 adjective pairs: 92 sourced from ODWN, 170 from plWN, and 40 derived from either of the two (18 ODWN, 22 plWN) but for which a more reliable alternative hypernym was proposed and approved by the annotators (e.g., the plWN had “effective” as hypernym for “deft”, but “skillful” was suggested and approved as an alternative and kept in the dataset). Acceptance rates were similar across resources, with a retention ratio of 0.55 for ODWN (92/166) and 0.60 for plWN (170/284). Furthermore, as already mentioned in Section 4, for each adjective included in the gold standard, the corresponding English definition was included after retrieving it from the OEWN. This was essential to ensure that the annotated relations reflected the intended senses of the adjectives rather than possible polysemous variants.

The annotation process began with an initial small dataset of 50 adjective pairs, annotated by two annotators using a three-label classification system (yes, no, maybe). All cases of negative agreement (no-no), strong disagreement (yes–no), weaker disagreement (no–maybe) and shared doubt (maybe-maybe) were discarded, while complete agreement (yes–yes) cases were retained and partial agreement (yes–maybe) ones were further discussed (examples are shown in Table 1). From this first sample, 62% (31 adjective pairs) were selected (Cohen’s kappa $\kappa = 0.65$). From a second sample of 100 pairs, 69% were retained ($\kappa = 0.64$), and from a third sample of 300 pairs, 67% (201 pairs) were selected ($\kappa = 0.61$). A third annotator was then included for the annotation of 100 pairs randomly selected from the third larger sample (300 pairs), providing additional discussion and validation (Fleiss’ kappa $\kappa = 0.48$).

Annotation	hypo	hyper
yes-no	intelligent	rational
yes-maybe	limitless	vast
yes-yes	lucid	aware
no-maybe	multiple	plural
maybe-maybe	unneeded	useless
no-no	productive	rich

Table 1: Examples of agreements and disagreements in the annotation of the pairs of adjectives performed by the two annotators.

Given the difficulty of semantic relations tasks which introduce a high level of subjectivity, inter-annotator agreement levels were moderate, but still acceptable, consistent and comparable across the different samples, with most disagreements involving “maybe” labels rather than direct contradictions (15 occurrences of the “maybe” label in the first sample, 24 in the second and 93 in the third).

In addition to the main gold standard where each hyponym is associated to one single exact hypernym, a second version of the dataset was later developed for model fine-tuning and evaluation (see Section 6.1). Here, synonyms of each hypernym were added based on synset membership in OEWN, in order to account for cases where multiple semantically correct hypernyms exist (see Table 2). This was motivated both by the natural multiplicity of hypernyms for a given adjective and for reducing the possible influence of definition-based prompts during the evaluation of the models.

6 Methodology

From the definition of hypernymy for adjectives to the issues in creating a benchmark dataset, it is clear that the task of lexical entailment already presents many issues for human understanding. As for the capabilities of language models, there have been attempts to assess their reliability following either a binary classification approach (choosing the correct hypernym) or a generation one (given a hyponym, predicting its most probable hypernym).

In this study, we use two models: TaxoLLaMa and SmolLM-360M-Instruct. We first test their

Dataset	hyponym-lemma	hypo_definition	hypernym-lemma	hyper_definition
single	relaxed	without strain or anxiety	calm	not agitated; without losing self-possession
multiple	relaxed	without strain or anxiety	calm, serene, tranquil, unagitated	not agitated; without losing self-possession

Table 2: One example of a pair taken from the two versions of the gold standard, first showing one single exact hypernym ("calm") for the input hyponym ("relaxed"), and then showing its synonyms found in OEWN too.

capabilities in the hypernymy discovery task and then fine-tune them on both our benchmark datasets in order to explore their capabilities of capturing semantic knowledge from them.

TaxoLLaMa (Moskvoretskii et al., 2024a) is the finetuned version of the LLaMA-2-7b model (Touvron et al., 2023), which was trained on the WordNet dataset for 16 taxonomy-related tasks and reached SoTA results on hypernymy discovery in different domains and languages. As it was neither trained nor tested on adjective examples, however, its performance in predicting them was expected to be lower.

SmolLM-360M-Instruct is part of SmolLM, a family of state-of-the-art small models[4]. It is optimized for instruction-following tasks through supervised fine-tuning on multiple instruction datasets. Even though it is much smaller than TaxoLLaMa, its size makes it ideal for rapid fine-tuning and evaluation on specialized tasks such as adjective hypernymy discovery, enabling us to explore if smaller models can acquire semantic relations when provided with limited data.

While recent models offer significantly improved performances across many tasks, our focus was on small to mid-sized models that could be fine-tuned efficiently, guided by high-quality and focused supervision. This emphasizes practical accessibility and cost-efficiency, but future work could investigate whether larger models perform significantly better in zero-shot or few-shot settings, or whether fine-tuning continues to provide a clear advantage.

6.1 Training Details

We performed the fine-tuning using the Unsloth method[5], which allowed us to quantize the models to 4 bits and train them using LoRA (Hu et al., 2022), reducing memory and computational requirements without affecting performance. This approach was particularly suited for using our small datasets and enabled the fine-tuning of both smaller and larger models like TaxoLLaMa with limited hardware resources and easy accessibility[6]. We fine-tuned the models using the SFTTrainer class, conducting the training for 60 optimization steps with a learning rate of 2e-4. The per-device batch size was set to 2, and gradient accumulation was used with 4 steps, simulating a batch size of 8. Leveraging our gold standard dataset, we fine-tuned the models on a training set composed of 211 items (70% of the dataset). The one below is a training sample, with an input question by the human user and an expected output answer by the GPT assistant, following a chat template:

> (from: human) "What are the hypernyms of the hyponym: "complicated" (definition: "difficult to analyze or understand")?",
>
> (from: gpt) "The hypernyms are: "difficult, hard" (definition: "not easy; requiring great physical or mental effort to accomplish or comprehend or endure")."

The training was performed in two separate settings: first the models were fine-tuned using the original gold standard dataset (hereafter referred to as "single") with a one-to-one correspondence between each hyponym and its relative hypernym; then a different gold standard was developed ("multiple"), implementing the single dataset with the synonyms of the hypernyms in order to have a one-to-many correspondence between each hyponym and its relative hypernyms (see Table 2). In order to have a consistent criterion and not fall into ambiguities, we relied on the OEWN and added all the adjectives pertaining to the same synsets of the original single hypernyms. This was done for two main reasons. First, given that a hyponym can have more than one hypernym, including only one correct exact hypernym in the gold standard

[4] https://huggingface.co/blog/smollm
[5] https://unsloth.ai/
[6] https://colab.research.google.com/

could leave out other possible candidates (e.g., "opportune" has "suitable" as its exact hypernym in the single dataset, but "appropriate" and "suited" were added in the multiple dataset, pertaining to the same synset: "meant or adapted for an occasion or use"). Secondly, the models trained on the single dataset would output only one hypernym for each hyponym when tested, while if trained with multiple possible hypernyms they would include more predictions and improve their semantic knowledge.

6.2 Prompting

Both during the zero-shot evaluation and after the fine-tuning, we used two different prompt settings: first, we just provided the models with an input hyponym adjective, and then we also gave them the hyponym definition. For the zero-shot prompting, we followed the below format originally used for TaxoLLaMa:

> <s>[INST] «SYS» You are a helpful assistant. List all the possible words divided with a comma. Your answer should not include anything except the words divided by a comma«/SYS»
>
> hyponym: humorous (full of or characterized by humor) | hypernyms: [/INST]

While after the fine-tuning we followed the below format:

> messages = ["from": "human", "value": "What are the hypernyms of the hyponym: "invigorating" (definition: "imparting strength and vitality")?",]

7 Results

The original base TaxoLLaMa model reached SoTA results and scored an MRR (Mean Reciprocal Rank) of 54.39 for English, but this was trained only on verbs and nouns sampled from the WordNet-3.0 graph. So, its performance on hypernymy discovery for adjectives was expected to be lower. In order to assess this, we test it in a zero-shot setting on our test set (91 pairs, 30% of the total gold standard), and we record an MRR of 9.4 when the model is not prompted with the definition of the input hyponym, and an increase to 25.8 when the definition is given. After the fine-tuning on the multiple dataset with the synonyms, those scores improve, respectively, to 23.6 and 33.3.

The difficulty in treating adjectives and distinguishing them from other parts of speech was a major one with the base TaxoLLaMa: when evaluated on the test set in a zero-shot setting without the definition, 58% of times it gives as output only nouns, 21% only adjectives, and 21% both (by providing it with the definition, the numbers change respectively to 14, 44 and 42%). After the fine-tuning on the single dataset, the amount of predicted adjectives significantly increases: 95% without the definition and 100% with the definition (both 100% when fine-tuned on the multiple dataset). This improvement is achieved by both TaxoLLaMa and SmolLM-360M-Instruct, as Table 3 shows. Given the difficulty with POS recognition and disambiguation, we consider this result to be almost as relevant and significant as the correct hypernym prediction.

Model	Setting	No def	With def
TaxoLLaMa	Zero-shot	0.39	0.78
TaxoLLaMa	ft-single	0.96	1.00
TaxoLLaMa	ft-multi	1.00	1.00
SmolLM-360M-Instr	Zero-shot	0.69	0.77
SmolLM-360M-Instr	ft-single	0.79	0.96
SmolLM-360M-Instr	ft-multi	1.00	1.00

Table 3: F1-score performances on predicting the correct POS (ADJ), before the fine-tuning (Zero-shot), when trained on the single dataset (ft-single) and when trained on the multiple dataset (ft-multi), without and with the definition.

After the fine-tuning, the models were tested in two settings against the two different datasets, in order to first evaluate their capability of inferring the exact correct hypernym of a given hyponym (against the single dataset), and then introducing also the possibility of giving synonyms in output (against the multiple dataset).

Table 4 shows the results in the first setting, where TaxoLLaMa-ft-multi reaches the best scores both without and with the definition in the prompt. Interestingly, providing the definition does not improve the performance of TaxoLLaMa-ft-single. At first glance, this may seem surprising. However, we need to consider that this model produces only one adjective as output, and this choice is often strongly influenced by lexical overlap with words in the input definition: e.g., "extant" (defined as "still in existence; not extinct or destroyed or lost") has "real" as its correct hypernym, but TaxoLLaMa-ft-single predicts "existent", being biased by the vocabulary of the definition.

This variability and influence due to definitions was also one of the motivations to introduce syn-

Model	Setting	No def	With def
TaxoLLaMa	Zero-shot	0.15	0.39
TaxoLLaMa	ft-single	0.32	0.31
TaxoLLaMa	ft-multi	0.35	0.44
SmolLM-360M-Instr	Zero-shot	0.13	0.16
SmolLM-360M-Instr	ft-single	0.14	0.15
SmolLM-360M-Instr	ft-multi	0.21	0.25

Table 4: Performances evaluated against the single dataset on predicting the exact correct hypernym before the fine-tuning, when trained on the single dataset and when trained on the multiple dataset, without and with the definition. Given that apart from TaxoLLaMa-Zero-shot all the other models output a single hypernym, precision and recall are equal and only one value is reported.

onyms into the multiple gold standard ("real" and "existent" are in the same synset in OEWN). By allowing synonyms, we can consider acceptable also cases where the model predicts a semantically correct hypernym which is not exactly the one we expected. Additionally, it is important to note that the base TaxoLLaMa is trained to output a list of hypernyms, whereas TaxoLLaMa-ft-single produces only a single prediction. This structural difference increases the chances of including a correct hypernym in the output for the base model (resulting in a higher F1-score of 0.39 against 0.31). However, when evaluating only the first-ranked hypernym from the base TaxoLLaMa's output list, its performance drops, with only 7 out of 14 correct hypernyms appearing in the first position.

The results shown in Table 5, obtained in the second evaluation setting against the multiple dataset, reveal that TaxoLLaMa-ft-multi performs better than the other models, and we observe a consistent improvement in the performance of the models when definitions are included in the prompt, unveiling the importance of incorporating them for disambiguation. Additionally, when compared to their counterparts trained on the single dataset (TaxoLLaMa-ft-single and SmolLM-ft-single), both TaxoLLaMa-ft-multi and SmolLM-ft-multi demonstrate clear gains, highlighting the benefit of allowing multiple valid hypernyms during training for capturing the nuanced nature of the hypernymy relation.

8 Conclusion

In this paper, we explore the understudied relation of adjective hypernymy, both from a theoretical and a computational perspective. We propose a definition for it, grounded in semantic inclusion and contextual substitutability, distinguishing it from synonymy and other relations. Making use of this framework, we construct a two-version gold-standard English dataset for adjective hypernymy by adapting the lexical information stored in the Polish, Dutch and Open English wordnets. Our dataset was validated through human annotation and synset-based disambiguation, offering a reliable, small and initial benchmark for future research.

We then evaluate the capabilities of language models - the large-scale TaxoLLaMa and the smaller SmolLM-360M-Instruct - on the task of hypernymy discovery, both in zero-shot and fine-tuned settings. Our results show that the models initially struggle with adjective hypernymy, particularly with issues of POS ambiguity and semantic polysemy. However, after fine-tuning on our dataset, especially the synonym-augmented variant, their performances improve, highlighting the value of task-specific training data. We also found that providing the models with explicit definitions of the input adjectives improves their ability to identify correct hypernyms. This underscores the central role of word sense disambiguation in first identifying and then modelling adjectival meaning.

As future work, we would like to a) expand the gold-standard dataset to reach a higher coverage and more generalizability, and allow for a better theoretical description and more reliability in training and evaluating language models, b) model adjective hypernymy in OEWN, identifying and representing hypernymy relations for all adjectives in the resource, c) support the interoperability between the OntoLex and NLP communities, promoting the use of adjective hypernymy in downstream applications, d) extend language models evaluation to other semantic relations between adjectives to explore how they differ theoretically and how well they are captured computationally.

Limitations

1. Dataset size: The gold standard is limited in size, which may restrict both its theoretical completeness (in modelling the full spectrum of adjectival hypernymy) and its practical utility (for model training and evaluation, especially on unseen or ambiguous adjective pairs), constraining generalizability and lexical coverage.

2. Models evaluated: Only two language mod-

Model	Setting	No def			With def		
		P	R	F-M	P	R	F-M
TaxoLLaMa	Zero-shot	0.04	0.13	0.06	0.07	0.23	0.11
TaxoLLaMa	ft-single	0.14	0.14	0.14	0.16	0.16	0.16
TaxoLLaMa	ft-multi	0.15	0.20	0.17	0.28	0.25	0.26
SmolLM-360M-Instr	Zero-shot	0.07	0.07	0.07	0.09	0.09	0.09
SmolLM-360M-Instr	ft-single	0.09	0.09	0.09	0.10	0.10	0.10
SmolLM-360M-Instr	ft-multi	0.16	0.10	0.12	0.20	0.14	0.16

Table 5: Performances evaluated against the multiple dataset on predicting a list of possible hypernyms before the fine-tuning, when trained on the single dataset and when trained on the multiple dataset, without and with the definition. For the models that output only one hypernym (TaxoLLama-ft-single, SmolLM-Zero-shot and SmolLM-ft-single, precision, recall and F-measure values are the same.

els were used, so this limits the conclusions regarding models capabilities by necessarily leaving out other existing architectures, sizes, and training data diversities.

3. Subjectivity of annotation: The annotation process, although based on carefully defined criteria and multi-annotator agreement, introduces a degree of subjectivity, which is even more accentuated by the graded, nuanced and polysemous nature of adjective hypernymy itself.

4. Language: The dataset, the theories and the evaluation are all limited to English. Consequently, some of the observed phenomena may not be generalized cross-linguistically, as the semantic behaviour of English adjectives may be different from those of other languages.

Acknowledgements

This research is supported by Taighde Éireann - Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.

R. M. W. Dixon. 1982. *Where have All the Adjectives Gone?* De Gruyter Mouton, Berlin, New York.

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

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Candida M. Greco, Lucio La Cava, and Andrea Tagarelli. 2024. Talking the talk does not entail walking the walk: On the limits of large language models in lexical entailment recognition.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Michael Hanna and David Mareček. 2021. Analyzing BERT's knowledge of hypernymy via prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, page 539–545, USA. Association for Computational Linguistics.

Frans Heyvaert. 2010. An outline for a semantic categorization of adjectives. In *Proceedings of the 14th EURALEX International Congress*, pages 1309–1318, Leeuwarden/Ljouwert, The Netherlands. Fryske Akademy.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Franz Hundsnurscher and Jochen Splett. 1982. *Grundlegung Einer Semantischen Beschreibung der Adjektive des Deutschen*, pages 16–47. VS Verlag für Sozialwissenschaften, Wiesbaden.

Cristopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics & Philosophy*, 30:1–45.

Thomas Kober, Julie Weeds, Lorenzo Bertolini, and David Weir. 2021. Data augmentation for hypernymy detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1034–1048, Online. Association for Computational Linguistics.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective scale probe: Can language models encode formal semantics information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290.

John Lyons. 1977. *Semantics*. Cambridge University Press.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2024. Lost in inference: Rediscovering the role of natural language inference for large language models.

Isa Maks, Piek Vossen, Roxane Segers, and Hennie van der Vliet. 2008. Adjectives in the Dutch semantic lexical database CORNETTO. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan. The COLING 2016 Organizing Committee.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

John P. McCrae. 2025. Renovating the verb hierarchy of english wordnet. In *Global WordNet Conference 2025*.

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

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. 2016. The open linguistics working group: Developing the linguistic linked open data cloud. In *10th Language Resource and Evaluation Conference (LREC)*, pages 2435–2441.

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

John P. McCrae, Francesca Quattri, Christina Unger, and Philipp Cimiano. 2014. Modelling the semantics of adjectives in the ontology-lexicon interface. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 198–209, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019a. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

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

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

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024a. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2331–2350. Association for Computational Linguistics.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024b. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

M. Lynne Murphy. 2003. Semantic relations and the lexicon: antonymy, synonymy, and other paradigms. *Acta Linguistica Hafniensia*, 36(1):185–189.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310, Bucharest, Romania. Global Wordnet Association.

I. Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155. Publisher Copyright: © 2023 The Author(s).

Victor Raskin and Sergei Nirenburg. 1995. Lexical semantics of adjectives a microtheory of adjectival meaning.

Ewa Rudnicka, Wojciech Witkowski, and Katarzyna Podlaska. 2016. Challenges of adjective mapping between plWordNet and Princeton WordNet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2413–2418, Portorož, Slovenia. European Language Resources Association (ELRA).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China. Association for Computational Linguistics.

Silke Scheible and Sabine Schulte im Walde. 2014. A database of paradigmatic semantic relation pairs for German nouns, verbs, and adjectives. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 111–119, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Erik Tjong Kim Sang. 2007. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 165–168, Prague, Czech Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2021. KEML: A knowledge-enriched meta-learning framework for lexical relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13924–13932.

Bringing IATE into the Semantic Web Family

Paula Diez-Ibarbia
Universidad Politécnica
de Madrid
paula.diez@upm.es

Patricia Martín-Chozas
Universidad Politécnica
de Madrid
patricia.martin@upm.es

Elena Montiel-Ponsoda
Universidad Politécnica
de Madrid
elena.montiel@upm.es

Abstract

This paper is an extension of previous work by the authors and other researchers that studies the application of the OntoLex-lemon model for representing the InterActive Terminology for Europe (IATE) database in the Semantic Web. While traditional XML-based approaches have been effective for multilingual terminological work, the Semantic Web enables richer, more interoperable representations. The study evaluates the suitability of OntoLex-lemon for modeling IATE's complex structure and identifies limitations in existing vocabularies. To address these, this paper tries to identify orher existing vocabularies and ontologies that could satisfy those limitations, which include term reliability, regional usage, lifecycle statuses, lookup forms, and concept cross-references. Still, some representation requirements are not covered by existing vocabularies and may need to be further discussed within the community.

1 Introduction

Traditional computational formats to structure terminological resources, such as those based on XML, have proven effective in supporting multilingual terminological work within commercial and industrial settings. These standards enable terminology teams to enforce consistent terminology and communication clarity while reducing time and cost.

Despite these strengths, representing terminologies using Semantic Web principles offers significant additional benefits. Unlike rigid XML hierarchies, Semantic Web standards such as RDF support more flexible, graph-based structures that facilitate incremental growth and easier integration. Furthermore, they enable interoperability across diverse datasets regardless of their origin, providing robust mechanisms for linking terminological data across languages and resources, enhancing reuse and connectivity in ways that traditional formats cannot easily achieve.

Among the Semantic Web models for representing terminologies and language resources, SKOS[1] and OntoLex-lemon[2] (hereforth, Ontolex) (McCrae et al., 2017) are the most widely adopted. The SKOS model has been successfully applied to large-scale thesauri such as EuroVoc (Díez et al., 2010), UNESCO Thesaurus (Pastor-Sánchez, 2016), AGROVOC (Caracciolo et al., 2013), TheSoz (Zapilko et al., 2013), and STW Thesaurus for Economics (Neubert, 2009).

In contrast to SKOS, Ontolex offers a wider linguistic framework as it provides a standardized model for representing lexical information. In fact, it has been widely applied to model lexical resources, such as the Apertium dictionaries (Gracia et al., 2018) and K Dictionaries (Bosque-Gil et al., 2016a). However, the potential of this model has also been thoroughly studied to represent terminological resources (Martín-Chozas et al., 2024).

In this line, the most representative and widely exploited terminological database in the European Union is the InterActive Terminology for Europe (IATE)[3], which is the object of this study.

2 IATE in the Semantic Web

IATE, now IATE2 (Zorrilla-Agut and Fontenelle, 2019), is the official terminology database of the European Union. Its primary purpose is to promote clarity and accuracy in the drafting and translation of EU documents, and it is freely accessible to the public. Developed collaboratively by EU bodies and maintained by the Translation Centre for the Bodies of the EU, IATE contains millions of entries covering a wide range of domains, from law and finance to agriculture and science. The entries registered in this database can offer a wide range of information, not only terminological, but also

[1] http://www.w3.org/2004/02/skos/core#

[2] https://www.w3.org/2016/05/ontolex/#core

[3] https://iate.europa.eu/home

lexicographical and conceptual.

The representation of this resource has already been addressed in the literature. First, Cimiano et al. (2015) proposed the representation of an IATE dump following the lemon vocabulary (foundation for the later Ontolex model) (McCrae et al., 2012) and complementary RDF properties from the TBX (Term Base eXchange) format, which was afterwards integrated in the TerminotecaRDF project (Bosque-Gil et al., 2016b), a platform to integrate semantically published terminological resources in Spanish. This first conversion of IATE was also the object of another work that aimed to enrich this resource with automatic translations (Arcan et al., 2018). This first attempt by Cimiano et al. (2015) to represent IATE in RDF was further extended by the creation of Terme-à-LLOD, a platform to convert and host terminological data based on TBX2RDF (di Buono et al., 2020).

Given this context, in this paper we explore the potential of Ontolex to meet the representation needs of IATE, identifying limitations of this standard and proposing complementary vocabularies to fill such void.

3 Representation requirements of IATE

As mentioned in the Introduction, the potential of Ontolex to represent terminological resources has previously been analysed by the authors in Martín-Chozas et al.(2024). This work revised a set of authoritative terminological resources, including IATE, and proposed an extension for the Ontolex model, Termlex[4]. Some of the requirements reported, that may require further discussion, are as follows:

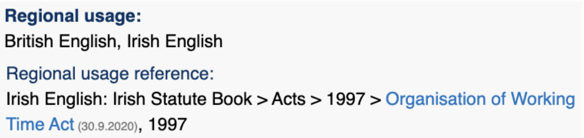
- The definitions and notes of a term, which often include additional data, such as the author or the source. To accommodate these requirements, we proposed the classes `termlex:Definition` and `termlex:Source`.

- The reliability of a term, which indicates the accuracy or the level of confidence of a given term. This factor varies amongst resources. For instance, in IATE is represented with stars, from one (lower reliability) to four. In other resources, such as Termium, this is represented as an *acceptability rating*, with values such as *correct*, *avoid* or *unofficial*. This indicator could be represented with `tbx:reliabilityCode`, but we believe that it should be stantardised, as this scale varies amongst resources, and for this reason we proposed the `termlex:ReliabilityCode` class pointing at a fixed set of values.

Continuing with this work, in the following sections we propose newly identified elements that need further discussion.

3.1 Regional Usage

A common element across IATE entries is the *regional use* indicator, as in Figure 1. This would not represent an issue itself, since Lexinfo[5] already provides support for this element with the properties `lexinfo:geographic` and `lexinfo:geographicVariant`. However, as shown in the figure, there are also references attached to this marker. Therefore, it may be required to reify this property to a class, so that the source information could be represented.



Regional usage:
British English, Irish English

Regional usage reference:
Irish English: Irish Statute Book > Acts > 1997 > Organisation of Working Time Act (30.9.2020), 1997

Figure 1: Regional use for the entry *Pact for Skills*

3.2 Lifecycle

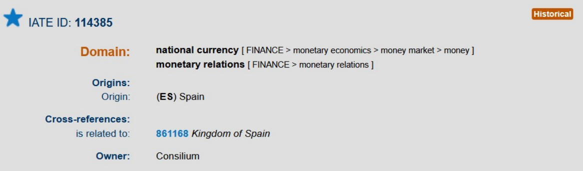
Another feature present in some IATE entries is the *lifecycle* (Figure 2). This indicator can adopt four values: *historical* (no longer in use or in existence), *proposed* (but not yet adopted), or *abandoned* (proposed but ultimately not adopted), and it is used to mark the status of a terminological entry. A close property to model this is `lexinfo:normativeAuthorization`. However, its range acquires a fixed set of values (*admitted*, *deprecated*, *preferred*, etc.) that do not match the values of IATE. Moreover, this property is indicated to represent the status of a specific term, and the lifecycle indicator refers to the entire terminological entry. Therefore, this issue requires further discussion.

3.3 Lookup Form

The *lookup form* is an interesting element that refers to any term or spelling variation that is

[4] https://termlex.oeg.fi.upm.es/myDocumentation3/index-en.html

[5] https://lexinfo.net/



IATE ID: **114385**

Historical

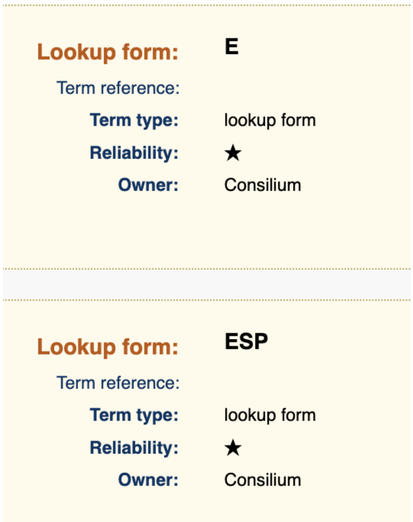
Domain: **national currency** [FINANCE > monetary economics > money market > money]
monetary relations [FINANCE > monetary relations]

Origins:
Origin: (**ES**) Spain

Cross-references:
is related to: **861168** *Kingdom of Spain*

Owner: Consilium

Figure 2: Lifecycle for the entry *Peseta*



Lookup form: **E**
Term reference:
Term type: lookup form
Reliability: ★
Owner: Consilium

Lookup form: **ESP**
Term reference:
Term type: lookup form
Reliability: ★
Owner: Consilium

Figure 3: Lookup form for the entry *Reino de España*

Cross-reference type	Concept ID
is narrower than	750475
is broader than	901212
is related to	114385
is not to be confused with	1620578
is antonym of	750475
is part of	3588819
has as part	901212
is capital city of	1891420
has as capital city	861168
is currency of	901212
has as currency	861168
is demonym of	1891744
has as demonym	883501
is caused by	1255366
is cause of	3640243
is predecessor of	2246619
is successor of	3591743
is seat of	3630354
has as seat	126540

Table 1: IATE examples of cross-references

searchable, but not displayed as a term, such as common spelling mistakes, alternative spellings, plural or inflected forms, etc. Figure 3 shows an example of two lookup forms for *Reino de España*. In this specific case, `lexinfo:shortForm` could work, as *E* and *ESP* are short forms for *España*. However, this is not always the case, such as *European Assembly* which is the lookup form for *European Parliament*[6], which could be modelled as a variant; or *Kronkolonie Anguilla* which is the lookup form for *Anguilla*[7], which could be modelled as a narrower concept.

Still, as observed in Figure 3, additional data may be added to the lookup form, so it might be necessary to propose a class to accommodate this information.

3.4 Concept Cross-References

Certain IATE concepts include cross-references; in other words, they provide information regarding the relationships between concepts. In particular, 19 types of relations have been identified, as displayed in Table 1.

Some of the relations have a certain linguistic nature, such as taxonomic ones. These types of relationships can be effectively modelled using SKOS (Simple Knowledge Organisation System)[8], a structured vocabulary specifically developed for the representation of thesauri. SKOS facilitates the expression of several conceptual relations, including *is narrower than* and *is broader than*. It is important to note that in SKOS terminology, the IATE relation *is narrower than* corresponds to `skos:broader`, whereas *is broader than* aligns with `skos:narrower`. Similarly, the IATE relation *is related to* is represented using the SKOS property `skos:related`. All these SKOS properties require instances of `skos:Concept` in both their domain and range. However, as the class `ontolex:LexicalConcept` is defined as a subclass of `skos:Concept`, no issues of semantic incompatibility arise in this context.

Nonetheless, SKOS alone is insufficient to capture all types of cross-references. For example, it cannot be used to represent antonyms, which are denoted in IATE by the value *is antonym of*. While this relationship cannot be expressed using SKOS, it can be modelled through LexInfo[9], an ontology that complements the Ontolex framework and provides a set of linguistic data categories. Specifically, the property `lexinfo:antonym` may be employed;

[6] https://iate.europa.eu/entry/result/126540

[7] https://iate.europa.eu/entry/result/883501

[8] http://www.w3.org/2004/02/skos/core#

[9] http://www.lexinfo.net/ontology/3.0/lexinfo#

however, it should be noted that this property is designed to operate between Lexical Senses, rather than at the level of Lexical Concepts. As such, the IATE data structure would need to be adapted this modelling approach. For instance, in IATE, the concepts 750475 and 3627400 are regarded as antonyms. In order to represent this accurately using `lexinfo:antonym`, the Lexical Sense of *random error* (a term associated with concept 750475) would need to be linked to the Lexical Senses of *systematic error*, *systematic error of measurement*, and *systematic measurement error* (all of which are associated with concept 3627400). Nevertheless, as we would prefer to be the most loyal possible to the original structure, other options will have to be explored.

In addition to linguistic relations, other types of associations may be identified, such as *has as capital city*, *has as currency*, or *has as demonym*, among others. As these relations are extra-linguistic in nature, it is necessary to employ alternative ontologies, such as the DBpedia Ontology (DBO)[10]. For instance, the property `dbo:capital` has been proposed to model the relation *has as capital city*. However, this property imposes constraints on both its domain and range. Consequently, the subject of the triple must be declared as both a `ontolex:LexicalConcept` and a `dbo:PopulatedPlace`. Similarly, the object must simultaneously belong to the classes `ontolex:LexicalConcept` and `dbo:City`.

Similarly, the cross-reference *has as currency* has been proposed to be modelled using the property `dbo:currency`. In order to satisfy the property's constraints, the object of the triple must be classified as a `dbo:Currency`, in addition to being declared a `ontolex:LexicalConcept`.

However, some cross-references, although present in DBO, cannot be utilised due to the nature of the property type. For example, to represent the cross-reference *has as demonym*, the property `dbo:demonym` was identified. Nevertheless, this is a data property, as it takes a string as its object. Nevertheless, the representation of IATE cross-references requires the use of an object property. As a result, the use of Wikidata[11] has been proposed, specifically the property `wdt:P1549`.

To summarise, in order to comply with the constraints imposed by certain properties, it is often necessary for instances to be assigned an additional class alongside `ontolex:LexicalConcept`. Furthermore, some properties may involve a restructuring of the data model to ensure conformity with their restrictions (e.g., with the property `lexinfo:antonym`). Finally, Table 2 presents a number of proposed solutions for modelling cross-references, although some are still ongoing work and have no modelling suggestions yet.

4 Conclusion and Future Work

In this paper, we have explored the application of the Ontolex model to the representation of IATE. While this task has been addressed in previous studies—primarily focusing on parameters such as definitions, notes, sources, and reliability codes—our work has concentrated on modelling other distinctive features of IATE. These include regional usage of terms, the life cycle of entries, lookup forms, and concept cross-references.

On the one hand, we have encountered limitations with certain existing properties, such as in the representation of antonymic relations between concepts. On the other hand, some features of IATE currently appear to be beyond the scope of representation. This is the case, for instance, with the source attribution for regional usage, or the treatment of lookup forms.

As for future work, the representation of several cross-reference types remains an open issue. Likewise, the modelling of the lifecycle of terminological entries is ongoing and may benefit from further discussion. Beyond IATE, we also aim to examine other terminological resources and assess their modelling requirements. One such example is TERMDAT[12], a database which considers the validation status[13] of a term record.

Acknowledgments

This work has received funding from the INESData project (Infrastructure to Investigate Data Spaces in Distributed Environments at UPM - TSI-063100-2022-0001, under the UNICO I+D CLOUD call) and TeresIA (Terminology and AI), both funded by the Ministry for Digital Transformation and the Civil Service, within the framework of the recovery plan PRTR financed by the European Union (NextGenerationEU).

[10] https://dbpedia.org/ontology

[11] https://www.wikidata.org/wiki/Wikidata:Main_Page

[12] https://www.termdat.bk.admin.ch/

[13] https://github.com/ontolex/ontolex/blob/master/notes/terminology-requirements.md

Cross-reference type	Property proposal
is narrower than	`skos:broader`
is broader than	`skos:narrower`
is related to	`skos:related`
is not to be confused with	`owl:differentFrom`
is antonym of	`lexinfo:antonym`
is part of	`dcterms:isPartOf`, `rico:isOrWasPartOf`, `dul:isPartOf`
has as part	`dct:hasPart`, `rico:hasOrHadPart`, `dul:hasPart`
is capital city of	
has as capital city	`dbo:capital`
is currency of	
has as currency	`dbo:currency`
is demonym of	
has as demonym	`wdt:P1549`
is caused by	`dbo:causedBy`
is cause of	
is predecessor of	`rico:precedesInTime`
is successor of	`rico:rico:followsInTime`
is seat of	`dul:hasLocation`
has as seat	`dul:isLocationOf`

Table 2: Property proposals for IATE cross-reference representation

References

Mihael Arcan, Elena Montiel-Ponsoda, John Philip McCrae, and Paul Buitelaar. 2018. Automatic enrichment of terminological resources: the IATE RDF example. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Https://doi.org/10.5281/zenodo.2599942.

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Guadalupe Aguado-de Cea. 2016a. Modelling multilingual lexicographic resources for the web of data: The k dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, page 65.

Julia Bosque-Gil, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de Cea. 2016b. Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain. *Term Bases and Linguistic Linked Open Data*.

Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. 2013. The agrovoc linked dataset. *Semantic Web*, 4(3):341–348.

Philipp Cimiano, John P McCrae, Víctor Rodríguez-Doncel, Tatiana Gornostay, Asunción Gómez-Pérez, Benjamin Siemoneit, and Andis Lagzdins. 2015. Linked terminologies: applying linked data principles to terminological resources. In *Proceedings of the eLex 2015 Conference*, pages 504–517. ISBN 978-961-93594-3-3.

Maria Pia di Buono, Philipp Cimiano, Mohammad Fazleh Elahi, and Frank Grimm. 2020. Terme-à-LLOD: Simplifying the conversion and hosting of terminological resources as linked data. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 28–35, Marseille, France. European Language Resources Association.

Luisa Alvite Díez, Beatriz Pérez-León, Mercedes Martínez-González, and Dámaso-Javier Vicente Blanco. 2010. Propuesta de representación del tesauro Eurovoc en SKOS para su integración en sistemas de información jurídica. *Scire: representación y organización del conocimiento*.

Jorge Gracia, Marta Villegas, Asuncion Gomez-Perez, and Nuria Bel. 2018. The Apertium bilingual dictionaries on the web of data. *Semantic Web*, 9(2):231–240.

Patricia Martín-Chozas, Thierry Declerck, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. 2024. Representing terminological data in the semantic web. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*.

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.

John P. McCrae, Dennis Spohr, and Philipp Cimiano. 2012. The lemon Lexicon Model for Ontologies. In *Proceedings of the 3rd International Conference* *on Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*.

Joachim Neubert. 2009. Bringing the "Thesaurus for Economics" on to the Web of Linked Data. *LDOW*.

Juan-Antonio Pastor-Sánchez. 2016. Proposal to represent the unesco thesaurus for the semantic web applying iso-25964. *Brazilian Journal of Information Studies: Research Trends*, 10(1):1–8.

Benjamin Zapilko, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. TheSoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web*, 4(3):257–263.

Paula Zorrilla-Agut and Thierry Fontenelle. 2019. IATE 2: Modernising the EU's IATE terminological database to respond to the challenges of today's translation world and beyond. *Terminology*, 25(2):146–174.

Ontologies for historical languages: using the LiLa and OntoLex-Lemon framework to build a Lemma Bank for Old Irish

Theodorus Fransen
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo A. Gemelli, 1
20123 Milan, Italy
theodorus.fransen@unicatt.it

Abstract

This paper presents a Linked Data approach to digitising and structuring Old Irish linguistic resources using the LiLa (Linking Latin) ontology, which is itself largely based on the OntoLex-Lemon framework (Cimiano et al., 2016). Old Irish, as a historical Celtic language with fragmented textual traditions, presents unique challenges for the creation and interoperability of digital resources. This work is part of the MOLOR project, whose aim is to create a knowledge base for Old Irish by interlinking texts, lexicons, and inflectional data. The first step in this ambitious endeavour is described here: the creation of an RDF linguistic Linked Data hub known as a Lemma Bank, similar to the one created as part of the LiLa project, addressing specific linguistic challenges and opportunities while adhering to the LiLa ontology.

1 Introduction

The digitisation of ancient and medieval languages presents significant challenges for computational linguistics and Digital Humanities scholars. Old Irish (600–900 CE) constitutes the earliest attested period of the Irish language—and of any Celtic language—for which the surviving documentary evidence is sufficiently comprehensive to allow a complete synchronic linguistic analysis; it is important for both the study of Indo-European linguistics and later stages of the Irish language (Stifter, 2009, 59). However, the combination of morphological complexity and orthographic variation (see Section 2.2), along with the use of different editorial standards, annotation schemas, and data formats in linguistic resources, creates substantial barriers to the successful use of standard natural language processing (NLP) methods (Doyle et al., 2019; Doyle and McCrae, 2024; Dereza et al., 2023). These factors also hinder resource compatibility and interoperability (Doyle and McCrae, 2025). The most important challenge in the current work is the variation seen in lemmatisation practice across lexical resources, particularly with regard to inflectional categorisation (see Section 4.1).

Recent advances in Linked Data technologies and semantic web frameworks offer promising solutions to address these challenges. The LiLa (Linking Latin) ontology, originally developed for Latin linguistic resources, provides a robust framework for representing ancient and historical language data in machine-readable formats that support interoperability and scholarly research (Passarotti et al., 2020).

This paper describes the implementation of the LiLa ontology—in turn adhering to OntoLex-Lemon—in the context of the Old Irish MOLOR project[1], presenting both methodological approaches and practical solutions to the unique challenges posed by this medieval Celtic language. The work contributes to the growing field of digital resources for ancient languages while demonstrating the adaptability of existing ontological frameworks to new linguistic contexts.

2 Background

2.1 Linked Data for ancient and historical languages

Chiarcos et al. (2018) address the underexplored application of Linked Open Data to ancient and historical languages and report on a case study applying LLOD principles to Assyriology by creating a Linked Data edition of the Electronic Text Corpus of Sumerian Royal Inscriptions. This linguistically annotated Sumerian corpus is connected to lexical resources, annotation terminology repositories, and museum collections housing the original cuneiform artifacts. The work serves as a foundation for expanding Linked Data approaches to other cuneiform corpora, including Ur III administrative

[1] https://tinyurl.com/molor-project

and legal texts, as part of the broader Machine Translation and Automated Analysis of Cuneiform Languages project (MTAAC, 2017–2020) and in close collaboration with the Cuneiform Digital Library Initiative or CDLI (CDLI contributors, 2025).[2]

Tittel and Chiarcos (2018) discuss the conversion of a medieval French medical treatise from a traditional scholarly edition into a semantically enriched digital format using RDFa (Adida et al., 2015) to link vocabulary entries to the Dictionnaire étymologique de l'ancien français (DEAF), offering a technological bridge between TEI/XML standards and Linked Open Data resources in digital philology.

The LiLa project[3] (Passarotti et al., 2020) represents a particularly relevant model, creating a comprehensive Linked Data ecosystem for Latin linguistic resources. LiLa's ontology provides sophisticated mechanisms for representing morphological, syntactic, and semantic information while maintaining interoperability across diverse resource types and scholarly traditions.

These projects demonstrate the feasibility and scholarly value of Linked Data approaches for historical language materials, enhancing discoverability, interoperability, and analytical capabilities.

2.2 Old Irish and linguistic challenges

Old Irish presents unique challenges for digital resource development that distinguish it from better-resourced ancient languages such as Latin or ancient Greek. Although Old Irish represents the first Celtic language with sufficient written evidence to enable comprehensive grammatical analysis, its associated contemporary text corpus—predominantly glosses on Latin manuscripts—is relatively small.

The language differs significantly from other Indo-European languages in several key ways. Its syntax follows a verb-first word order pattern (Stifter, 2009, 60), characteristic of Insular Celtic languages. Stifter (2009, 60) says the following about the linguistic complexity of Old Irish:

> Old Irish is almost prototypical for a language whose grammatical behaviour cannot be described adequately by synchronic rules. The bewildering complexities of some of its grammatical subsystems, especially that of verbal morphology, become transparent only when viewed from a diachronic position, and in order to understand allomorphic variation correctly it is essential to work with underlying forms and their often quite dissimilar surface representations

The same author continues with an illustrative example: "both *do·sluindi* /doˈslunʲdʲi/ '(s)he denies' and negated *ní·díltai* /ˈdʲiːlti/ '(s)he does not (*ní*) deny' regularly reflect the same diachronically underlying structure *dī-slondīθ" (Stifter, 2009, 60).

Phonologically, Old Irish has an extensive consonant inventory and displays what some scholars have termed a vertical vowel system (Anderson, 2016).[4] Although not unique to Old Irish, the language also exhibits initial mutations—changes to consonants and sometimes vowels at the start of the word based on grammatical context—whose orthographic encoding is neither systematic nor consistent in early texts. For example, a lenited *f* (which is silent) may be represented as *f*, *ḟ*, or may disappear altogether in the orthography.

These distinctive linguistic characteristics make the development of quality computational tools for Old Irish a pressing scholarly need.[5]

Old Irish orthography shows variation across manuscripts and time periods, reflecting both scribal practices and genuine linguistic change during the Old Irish period. Often, this orthographic variation is intertwined with morphological variation and change, creating additional challenges for automated processing and cross-referencing of textual materials and—most relevant for the purposes of the current paper—the selection and harmonisation of an exhaustive set of representative citation forms, i.e. lemmas, as illustrated in Section 4.1.

What may be viewed as orthographic or phonological variation may point to morphological variation, which may in turn be obscured by particular spellings. Taking the example of *cladaid* and *claidid* (see Appendix, Table 2), the difference here is the consonance (non-palatal vs palatal) of root-final *d* (i.e. non-palatal *clad-* vs palatal *claid-*), which

[2] The data structure of CDLI is illustrated at `https://cdli.ox.ac.uk/wiki/doku.php?id=data_structure`.

[3] `https://lila-erc.eu/`

[4] A vertical vowel system in phonology refers to a vowel inventory organised primarily along the height dimension (high, mid, low) with minimal or no distinctions based on frontness/backness or rounding.

[5] For a more extensive discussion on the morphological complexity of the Old Irish verb in the context of computational modelling see Fransen (2020).

signifies a difference in inflection class. An orthographic representation such as *cladid*, however, could represent either morphological variant.

The fragmentary nature of the Old Irish corpus also creates challenges for comprehensive coverage, as many forms and morphonological contexts are only attested rarely or in ambiguous contexts. This requires careful balance between exhaustive representation and practical utility in ontological design decisions. The next subsection discusses the lexicographical landscape and lists the resources instrumental for compiling an Old Irish Lemma Bank.

2.3 The Old Irish resource landscape

Griffith et al. (2018) give an overview of lexicographical resources available for early medieval Irish. Dereza (2018) is a first valuable and instructive attempt at building a lemmatiser for Old Irish using rule-based and machine learning techniques (based on DIL, see below).

Stifter et al. (2022) call for greater interoperability between linguistic resources for early medieval Irish. Indeed, there has recently been a push towards resource interoperability and standardisation. Doyle and McCrae (2025) report on a new lexical resource and the publication of two treebanks following the Universal Dependencies (de Marneffe et al., 2021) standard of annotation, noting that these resources "have been created with the express purpose of ensuring lexical compatibility between them" (p. 393). The equally novel resource Goidelex (Anderson et al., 2024) incorporates normalised orthographical forms and is compatible with other frameworks (see below).

Despite these promising developments, resources currently do not speak the same language, i.e. there is a lack of a unified ontology and controlled vocabularies following Semantic Web standards. The current work is a first step in overcoming this limitation, by creating a collection of canonical forms, i.e. lemmas, used to interlink resources using Linguistic Linked Data methods following the LiLa framework (Passarotti et al., 2020). The current work uses the following three resources for the collection of lemmas.

Dictionary of the Irish Language (DIL) (eDIL, 2019)—The standard dictionary for medieval Irish covering the period 700–1700CE, which transitioned from print to digital format in 2007. While DIL offers extensive lexical coverage, it suffers from non-exhaustive and inconsistent annotation of examples and limited data extraction functionalities for large-scale research. Furthermore, headwords are not always representative of Old Irish.

Corpus PalaeoHibernicum (CorPH) (Stifter et al., 2021)—CorPH constitutes the most morphosyntactically detailed and comprehensive lexical resource for Old Irish. It contains over 10,500 word entries from 77 analysed texts, available as downloadable CSV files. While not immediately relevant for the task of building a Lemma Bank, its complex word structure breakdown makes it difficult to link back to source texts.

Würzburg lexicon (Kavanagh and Wodtko, 2001)—A print dictionary accompanied by PDF files for the highly important 8th-century Würzburg glosses (not covered in CorPH).

Goidelex (Anderson et al., 2024)—This novel resource currently contains 671 entries from the Würzburg glosses, extracted from Kavanagh and Wodtko (2001). It provides detailed inflectional and phonological data, uses normalised spelling, links to other resources, and follows modern data standards, including Paralex, a novel standard for inflectional lexicons (Beniamine et al., 2023). Paralex includes tools for converting data into the RDF OntoLex-Lemon format. It is also compatible with the Cross Linguistic Data Format (Forkel et al., 2018).

3 Modelling: the LiLa lemma ontology

The LiLa ontology provides a comprehensive framework for representing linguistic data through Linked Data principles. The LiLa knowledge base centres on a comprehensive collection of Latin lemmas that serve as connection points between different language resources. Since the system is vocabulary-focused, these lemmas link together dictionary entries, corpus texts and NLP output that reference the 'common denominator', enabling seamless integration across resources (Passarotti et al., 2020, 186–187). The ontology incorporates multiple levels of linguistic analysis, from graphemic representation through morphological, lexical, and syntactic annotation to semantic and pragmatic information. The ontology employs standardised vocabularies and URI schemes that enable cross-referencing between different resources and projects, supporting both human-readable scholarly annotation and machine-processable data that can be queried and analysed computationally. The LiLa Lemma ontology is described and exemplified in

Listing 1: LiLa Lemma Class Definition

```
lila:Lemma a rdfs:Class,
            owl:Class ;
    rdfs:comment "A Lemma must have a POS, but it cannot have more than 1",
                 "In LiLa, a Lemma is a form in the word inflection that is used (or may
     potentially be used) to lemmatize tokens in a corpus." ;
    rdfs:label "Lemma" ;
    rdfs:subClassOf ontolex:Form ;
    rdfs:subClassOf [ a owl:Restriction ;
                      owl:onClass lila:POS ;
                      owl:onProperty lila:hasPOS ;
                      owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger
                    ] .
```

Listing 2: The entry *sequor* ‘to follow’ as modelled according to the LiLa Lemma Class

```
<data/id/lemma/124461>
        a                       lila:Lemma ;
        rdfs:label              "sequor" ;
        lila:hasBase            <data/id/base/417> ;
        lila:hasInflectionType  lila:v3d ;
        lila:hasPOS             lila:verb ;
        lila:lemmaVariant       <data/id/lemma/124462> ;
        dcterms:isPartOf        <data/id/lemma/LemmaBank> ;
        ontolex:writtenRep      "sequor"@la , "secor"@la .
```

Listings 1 and 2, respectively.

Although a discussion on the application of LiLa and OntoLex-Lemon classes and properties to Old Irish lexemes has already been provided in Fransen et al. (2024), it might be prudent to briefly explain the `lila:lemmaVariant` property here again. This property was created in LiLa to cater for the use of alternative canonical forms used for the same lexeme as represented in lexical entries while at the same time maintaining resource interoperability. Consider Figure 1. Here we see four inflectional variants—first vs second conjugation, active vs deponent—representing four citation forms for the same Latin lexeme ‘to limp’. Using the commutative property `lemmaVariant`, “LiLa harmonises different lemmatisation strategies and annotation styles, thus granting interoperability” (Passarotti et al., 2020, 193). This is exactly because each resource or token linked to one of those lemmas is linked to any other token or resource lemmatised using one of the variant lemmas. Note that this elegantly circumvents the restriction that an `ontolex:LexicalEntry` can have at most one canonical form (Cimiano et al., 2016, §3.1).

For purely orthographic variation (or certain phonological variants in the case of Old Irish (Fransen et al., 2024)), the different spellings are modelled according to the more general property `representation` (`ontolex:WrittenRep`) and, crucially, *as part of the same lemma and hence URI*—compare the written variants *sequor* and *secor* in Listing 2 and *claudo* and *cludo* in Figure 1, respectively.

For the Old Irish implementation, LiLa’s core concepts have been adhered to—not without challenges—as detailed in Section 4.1.

4 Implementation

4.1 Harmonisation challenges

Adapting the LiLa ontology for Old Irish presents several significant challenges that require careful methodological consideration. As discussed in Section 2.2, Old Irish is characterised by a high degree of synchronically unpredictable (or at least opaque) allomorphy, and in this respect arguably exceeds the morphological complexity of Latin, particularly in verbal inflection.[6]

Admittedly, a high degree of allomorphy, combined with spelling variation, is not necessarily problematic for the task of collecting and aligning lemmas from already existing lexical resources, as

[6] Although few Celtic and classical scholars would disagree, the author is not aware of any empirical study that compares Old Irish and other historical Indo-European languages such as Latin using features of morphological complexity; however, the reader may want to consult Fransen (2019, 30–34) for some quantitative observations on the Old Irish verbal system.

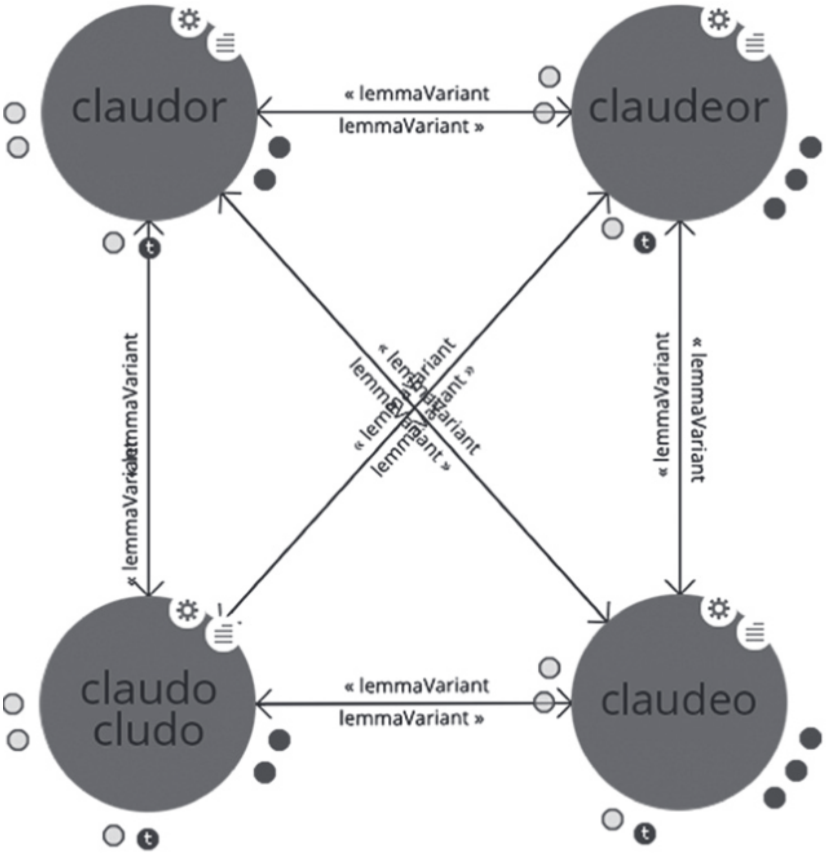


Figure 1: Four Latin lemmas with different inflection patterns representing different citation forms, connected through the commutative property `lila:lemmaVariant`; taken from Passarotti et al. (2020, 193).

described in this paper (as opposed to automatic morphological analysis and lemmatisation of raw text). The challenge at hand, more so than morphological complexity itself, has proven to be the lack of uniformity between resources in the categorisation of morphological (or more specifically, inflectional) variation. Linguistic complexity and variation and lack of descriptive uniformity are obviously related, with other factors at play, such as uncertainty due to gaps in attestation.

Inflectional variation is the pillar of the lemma variant property. Notwithstanding this property's usefulness, mapping the variation seen in Old Irish data onto clear-cut inflectional variants proved rather challenging, especially with the nominal system. Tables 1 and 2 in the Appendix provide an overview of the inflectional variation and microclasses seen with Old Irish lemmas. Notably in the case of the nominal system, one can observe 1) unbalanced categorisation of lemmas (one-to-many relationships) across resources; 2) different, yet sometimes overlapping inflection classes (`fl_cat`); and 3) different resolution (i.e. macro- vs. microclasses). The noun *fius* 'knowledge', for example, is described as a *u*-stem or *o*-stem, and can be masculine or neuter. Creating four lemmas for all four permutations seems excessive and is not reflective of Old Irish—the linguistic reality is a mixed (and not fully attested) inflectional paradigm due to confusion between stems and a general shift of *u*-stems towards *o*-stem inflection (Thurneysen, 1946, §309).

Furthermore, as can be partially gleaned from the footnotes accompanying the tables in the Appendix, the annotation of inflectional classes is at times arbitrary, inconsistent, or incorrectly suggestive of differences; a good example of the latter is the inclusion of both the lemmas *bádaid* and *báidid* 'to drown' in CorPH—see Table 2, footnote *c*. All this is compounded by the occasional slip in CorPH.[7]

Orthographic variation presents another significant challenge, as Old Irish manuscripts show considerable spelling variation between different scribes and across chronological periods. The on-

[7] As part of the data extraction process and resource alignment, the author has already identified and corrected hundreds of mistakes in CorPH, ranging from typos to the assignment of the wrong language to an entry.

tology must support multiple orthographic representations while maintaining scholarly precision in distinguishing between genuine linguistic (mostly inflectional) variation and scribal variation and inconsistency. The written representation data property of `ontoLex:Form`, like in LiLa, was considered sufficient to encode orthographic variation (as opposed to encoding morphological variation, for which the `lila:lemmaVariant` property, as detailed in Section 3, was used).

That being said, the primary purpose of a Lemma Bank is to accommodate and unify variation found in lemmatisation practice in order to make lexical resources interoperable; it is not the place for prioritising certain spellings or providing a highly principled and systematic morphological categorisation of forms. Of course, a Lemma Bank can be a first point of call and as such might benefit from linguistic means that facilitate search queries, linguistic description, or research purposes (e.g. using typographical means consistently to make compounding more explicit, see Section 4.2), as long as the matching of lemmas with lexical entries (or perhaps even tokens in a text) remains computationally trivial.

Since it contains lemmas from existing resources, a Lemma Bank naturally inherits some of the lemmatisation inconsistencies found in those resources. However, by means of 1) exhaustive coverage of (potential) lemmas and 2) the principles of Linked Data, and especially the SPARQL query language (Prud'hommeaux and Seaborne, 2008), interoperability and effective retrieval of linguistic information in linked specialist resources, possibly built with standardisation or normalisation in mind, is warranted.

Goidelex (Anderson et al., 2024), which employs a normalised orthography for Old Irish (Fransen et al., 2023), may serve as an example. Since it is built according to the Paralex standard for inflected lexicons, which, as mentioned earlier, includes an ontology for conversion into OntoLex-Lemon lexicons, there exists the theoretical possibility of navigating from a lemma in the Lemma Bank to the associated lexical entry in Goidelex and retrieving inflectional paradigms in normalised orthography.

4.2 SQL tables

RDF conversion was preceded by semi-automatic creation and integration of SQL tables based on the extraction of data from the resources mentioned in Section 2.3, currently limited to nouns (including verbal nouns and proper nouns), numerals, and verbs.

For nouns, the author integrated CorPH's lemma table, a selection of compositional forms entries from CorPH's morphology table, and Goidelex and Kavanagh and Wodtko (2001)[8] (for the Würzburg lemmas, which are not in CorPH). For verbs, CorPH was again used, manually aligned with the verbal subset in CSV files extracted from the Würzburg lexicon (Kavanagh and Wodtko, 2001)[9] and verb entries from DIL. The mid-high dot has been invariably employed with compound verbs (rather than the hyphen, as used in, e.g, DIL), separating the pretonic preverb from the stressed part of the verb, e.g. *do·beir* 'to give, bring'. Compound nouns were hyphenated broadly following Kavanagh and Wodtko (2001), even where they were not hyphenated in other source data, e.g. *dag-athair* 'good father', primarily with a view to creating typographical consistency among lemmas.[10] Table 2 in the Appendix closely mirrors (a snippet of) the initial spreadsheet (apart from the URIs, of course) used to manually align verb lemmas—subsequently converted into a TSV file and imported as an SQL table.

4.3 RDF conversion

The Old Irish lemma data in the relational databases—*lemma*, *lemma_wr* and *variant_group*—was subsequently converted into RDF using the D2RQ mapping language (Cyganiak et al., 2012), emulating the URI schemes for the LiLa Lemma Bank. However, at least in the first instance, fewer properties have been used, the absence

[8] More precisely, the Goidelex lexemes table which represents (normalised) entries with more than one attestation, plus hapax lemmas manually added from the Würzburg lexicon.

[9] The lexicon was automatically parsed and converted into CSV files by Dr Aaron Griffith on the basis of accompanying PDF files, with assistance from the Utrecht Digital Humanities Lab. The parsing script is found at https://github.com/CentreForDigitalHumanities/wurzburg-glosses-extraction while the CSV files were generously shared privately with the author. Admittedly, the parsed files only cover a selection of POS categories and some entries are missing (some verbs had to be manually added). Moreover, the extraction is noisy in places.

[10] These typographical separation devices reflect morphological boundary markers which are linguistically insightful, even though their inclusion might arguably go beyond the remit of a Lemma Bank. Having said this, ignoring these markers in queries is trivial, while inserting them post hoc is not. Furthermore, they can easily be deleted in a string manipulation step to facilitate matching with linguistic resources that do not employ these markers, such as diplomatically edited text resources.

Listing 3: The form *breth* ‘bearing’ as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/1490>
      a       molor:Lemma ;
      rdfs:label "breth" ;
      molor:hasPOS molor:noun ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/4924> ;
      ontolex:writtenRep "breth" .
```

Listing 4: The form *brith* ‘bearing’ as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/4924>
      a       molor:Lemma ;
      rdfs:label "brith" ;
      molor:hasPOS molor:noun ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/1490> ;
      ontolex:writtenRep "brith" .
```

of `lila:hasInflectionType` probably being the most significant difference (see Section 5). The Lemma Bank currently totals 6,000+ lemmas, a fifth of which are verbs.

The URI schemes otherwise follow LiLa conventions, ensuring future compatibility with ancient and historical language Linked Data resources. Listings 3–6 exemplify the RDF version of the entries for the nouns *breth* and *brith* ‘bearing’ as well as for the verbs *molaithir* (deponent) and *molaid* (active) ‘to praise’, illustrating the author’s choices in employing the written representation datatype vs the lemma variant property with these forms (the reader may want to refer to Tables 1 and 2 in the Appendix, respectively, for more details).

5 Discussion

5.1 Recapitulation: scope and function of a Lemma Bank

Inconsistent or divergent annotation of inflection types has presented the most complex aspect of the collecting and modelling lemmas from legacy resources, as discussed in Section 4.1. It was decided to not try and facilitate divergent inflectional annotation practices as part of the MOLOR RDF Lemma Bank, as this would have entailed having to focus on the linguistic exercise of (further) correcting and harmonising annotation in existing resources, which would most likely have meant choosing one categorisation system over another. Echoing what was discussed in Section 4.1, the goal of a Lemma Bank is to capture variant lemmatisation practices rather than aiming for standardisation and normalisation. Moreover, taking a principled and fine-grained approach to morphological variation would redundantly emulate work as part of Goidelex (Anderson et al., 2024), which is focused on providing high-resolution inflectional information employing a normalised orthography. Furthermore, considering the fact that a `molor:Lemma` (and `lila:Lemma`) is a subclass of `ontolex:Form`, the absence of morphological information is actually in line with the OntoLex-Lemon core model, where it is the lexical entry that is assigned morphological properties and not the form.

More generally, leaving specialised information to individual resources conforms to the philosophy of the Linked Data paradigm and its premise of knowledge being distributed (even if potentially divergent or conflicting in nature).

5.2 Applications to computational tasks

Despite preserving variation rather than enforcing standardisation, the harmonised lemma representations in the RDF Lemma Bank could significantly assist computational lemmatisation efforts for Old Irish. Work such as Dereza (2018) demonstrates the challenges of automatic lemmatisation for historical languages, where morphological complexity and orthographic variation create substantial obstacles. By providing a comprehensive, structured repository of lemma-form relationships across multiple lexical resources, the Lemma Bank offers a rich training resource that could improve lemmatisation accuracy. The ontological structure allows for sophisticated querying of lemma variants and their attestations, potentially enabling more robust handling of the orthographic and morphological variation that characterises Old Irish texts.

Listing 5: The form *molaithir* ‘to praise’ as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/5744>
      a       molor:Lemma ;
      rdfs:label "molaithir" ;
      molor:hasPOS molor:verb ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/5745> ;
      ontolex:writtenRep "molaidir" , "molaithir" .
```

Listing 6: The form *molaid* ‘to praise’ as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/5745>
      a       molor:Lemma ;
      rdfs:label "molaid" ;
      molor:hasPOS molor:verb ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/5744> ;
      ontolex:writtenRep "molaid" .
```

5.3 Advantages of RDF over traditional data formats

The choice of RDF over traditional relational databases or flat file formats (TSV, CSV) reflects the distributed and interconnected nature of lexical knowledge. While SQL databases excel at structured queries within closed systems, RDF graphs enable seamless integration across heterogeneous resources and institutions. This is particularly valuable for historical linguistics, where lexical data often originates from multiple scholarly traditions and projects. The graph-based model naturally represents the complex relationships between lemmas, forms, and attestations, while SPARQL queries can traverse these relationships in ways that would require complex joins in relational systems. Moreover, the use of standardised vocabularies like OntoLex-Lemon ensures interoperability with other linked lexical resources, facilitating broader comparative and cross-linguistic research that would be challenging to achieve with isolated database systems.

6 Conclusion and future work

The current work has focused on building an RDF Lemma Bank for Old Irish to interconnect linguistic resources according to semantic web principles and the LiLa ontology in particular. The application of the LiLa ontology to Old Irish demonstrates both the potential and challenges of existing Linked Data frameworks for under-resourced ancient and historical languages. A clear-cut mapping to LiLa lemma properties is not always trivial due to morphological and orthographic variation, inconsistency, or different resolution in inflectional annotation (to which we can add uncertainty originating in gaps in attestation). The decision was made not to enforce a single, harmonised morphological annotation system within the Lemma Bank, but instead to leverage the `lila:lemmaVariant` property (currently without using `lila:hasInflectionType`) to interlink alternative lemmas and `ontolex:writtenRep` for orthographic variation, thus respecting the distributed nature of Linked Data.

The Lemma Bank is expected to grow in size (more lemmas and more POS categories), with the linking of resources to the Lemma Bank constituting the beginnings of a knowledge base for Old Irish, with SPARQL endpoints that support complex morphological and syntactic searches throughout the Old Irish corpus.[11] Such a knowledge base will hopefully lead to enhanced search and analysis capabilities, while also highlighting areas where traditional philological approaches remain necessary supplements to computational methods.

User feedback from the scholarly community will be the best measure of project success; the author hopes to report on use cases during a future edition of the OntoLex workshop series.

Acknowledgments

MOLOR—Morphologically Linked Old Irish Resource—has received funding as part of the European Union’s Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Ac-

[11] The author is particularly interested in making two lexical resources interoperable with the Lemma Bank: Goidelex (Anderson et al., 2024) and the Irish section of PaVeDa (Roma and Zanchi, 2025), the latter of which currently consists of some hundred Old Irish verb entries marked for valency patterns.

tions (MSCA), grant agreement No 101106220. The author thanks the (former) LiLa project team for their generous sharing of expertise and resources. The initial indexing of verb lemmas was performed by Federico Simone Samperi (Università di Pavia). The author also acknowledges the contributions of Celtic Studies scholars, especially Prof. Elisa Roma, who has been kindly providing linguistic assistance over the course of the MOLOR project.

References

Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. 2015. RDFa core 1.1 - third edition. W3c recommendation, World Wide Web Consortium.

Cormac Anderson. 2016. *Consonant colour and vocalism in the history of Irish*. Ph.D. thesis, Adam Mickiewicz University, Poznań.

Cormac Anderson, Sacha Beniamine, and Theodorus Fransen. 2024. Goidelex: A lexical resource for Old Irish. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 1–10, Torino, Italia. ELRA and ICCL.

Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. Paralex: a DeAR standard for rich lexicons of inflected forms. In *International Symposium of Morphology*. https://www.paralex-standard.org.

CDLI contributors. 2025. About CDLI. https://cdli.earth/about. [Online; accessed 2025-07-13].

Christian Chiarcos, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 2018. Towards a linked open data edition of sumerian corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), May 7–12, 2018, Miyazaki, Japan*, pages 2437 – 2444.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community report. W3C community group final report, World Wide Web Consortium. https://www.w3.org/2016/05/ontolex/.

Richard Cyganiak, Chris Bizer, Jörg Garbers, Oliver Maresch, and Christian Becker. 2012. The D2RQ mapping language. v0.8 – 2012-03-12.

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

Oksana Dereza. 2018. Lemmatization for ancient languages: Rules or neural networks? In *Artificial Intelligence and Natural Language*, pages 35–47, Cham. Springer International Publishing.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023. Do not trust the experts - how the lack of standard complicates NLP for historical Irish. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Adrian Doyle and John McCrae. 2025. Development of Old Irish lexical resources, and two Universal Dependencies treebanks for diplomatically edited Old Irish text. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 393–402, Albuquerque, USA. Association for Computational Linguistics.

Adrian Doyle and John P. McCrae. 2024. Developing a part-of-speech tagger for diplomatically edited Old Irish text. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 11–21, Torino, Italia. ELRA and ICCL.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 2019. A character-level LSTM network model for tokenizing the Old Irish text of the Würzburg glosses on the Pauline epistles. In *Proceedings of the Celtic Language Technology Workshop*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.

eDIL. 2019. An electronic dictionary of the Irish language. Based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913–1976).

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5(180205).

Theodorus Fransen. 2019. *Past, present and future: Computational approaches to mapping historical Irish cognate verb forms*. Ph.D. thesis, Trinity College Dublin, Dublin.

Theodorus Fransen. 2020. 3 Automatic morphological analysis and interlinking of historical Irish cognate verb forms. In Elliott Lash, Fangzhe Qiu, and David Stifter, editors, *Morphosyntactic Variation in Medieval Celtic Languages. Corpus-Based Approaches*, chapter 3, pages 49–84. De Gruyter Mouton, Berlin, Boston.

Theodorus Fransen, Cormac Anderson, and Sacha Beniamine. 2023. Towards a normalised orthography for Old Irish. Paper at *36th Irish Congress of Medievalists*, Dublin, 22–23 June 2023.

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

Aaron Griffith, David Stifter, and Gregory Toner. 2018. Early Irish lexicography – a research survey. *Kratylos*, 63(1):1–28.

Séamus Kavanagh and Dagmar S. Wodtko. 2001. *A lexicon of the Old Irish glosses in the Würzburg manuscript of the epistles of St. Paul*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Kim McCone. 1987. *The Early Irish verb*. An Sagart, Maynooth.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL query language for RDF.

Helmut Rix and Martin Kümmel. 2001. *LIV, Lexikon der indogermanischen Verben: die Wurzeln und ihre Primärstammbildungen*, 2., erw. und verb. aufl. edition. Reichert, Wiesbaden.

Elisa Roma and Chiara Zanchi. 2025. Old Irish in the PaVeDa: Issues, perspectives, and two case studies. In Dylan R. Cooper, Rachel Martin, Graham O'Toole, and Samuel Ezra Puopolo, editors, *Proceedings of the Harvard Celtic Colloquium 42: 2023*, volume 42, pages 212–237. Harvard University Press, Boston.

David Stifter. 2009. Early Irish. In Martin Ball and Nicole Müller, editors, *The Celtic Languages*. Routledge.

David Stifter, Bernhard Bauer, Elliott Lash, Fangzhe Qiu, Nora White, Siobhán Barrett, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021. Corpus PalaeoHibernicum (CorPH) v1.0. `https://chronhib.maynoothuniversity.ie`.

David Stifter, Nina Cnockaert-Guillou, Beatrix Färber, Deborah Hayden, Máire Ní Mhaonaigh, Joanna Tucker, and Christopher Guy Yocum. 2022. Developing a digital framework for the medieval Gaelic world. Project report, Queen's University Belfast.

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*. Dublin Institute of Advanced Studies, Dublin. Translated by D. A. Binchy and Osborn Bergin.

Sabine Tittel and Christian Chiarcos. 2018. Historical lexicography of Old French and linked open data: transforming the resources of the dictionnaire étymologique de l'ancien français with OntoLex-Lemon. In *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, co-located with LREC2018, 12 May 2018, Miyazaki, Japan*.

A Comparison of Old Irish lexical entries

URI	CorPH					Goidelex					DIL[a]					Meaning
	ID	Entry	POS	fl_cat	Gender	ID	Entry	POS	fl_cat	Gender	ID	Entry	POS	fl_cat	Gender	
http://molor.eu/data/id/lemma/4670	4627	*fius*	noun	u	n	fius-282	*fius*	noun	u	n	dil.ie/22221	1 *fis*, *fius*	noun	u, o	n, m	know-ledge[b]
	—	—	—	—	—	fius-282-1	*fius*	noun	u	m	—	—		—	—	
	—	—	—	—	—	fius-282-2	*fius*	noun	o	n	—	—	—	—	—	
—	—	—	—	—	—	—	—	—	—	—	dil.ie/21774	*fes(s)*[c]	x[d]	—	—	
http://molor.eu/data/id/lemma/1490	430	*breth*	verbal_noun	ā	f	—	—	—	—	—	dil.ie/6785	*breth*	verbal_noun	ā, i	f	bearing[e]
http://molor.eu/data/id/lemma/4924	9582	*brith*	verbal_noun	i[f]	f	brith-97	*brith*	verbal_noun	i	f	dil.ie/6860	*brith*	x	—	—	
—	—	—	—	—	—	breith-97-1	*breith*	verbal_noun	ā	f	dil.ie/6698	*breith*	x	—	—	
http://molor.eu/data/id/lemma/4410	5230	*adaig*	noun	ī	f	adaig-524	*adaig*	noun	n1[g]	f	dil.ie/256	1 *adaig*	noun	iā	f	night
—	—	—	—	—	—	—	—	—	—	—	dil.ie/856	*aidche*	x	—	—	
—	—	—	—	—	—	—	—	—	—	—	dil.ie/33617	*oidche*	x	—	—	
http://molor.eu/data/id/lemma/4935	3261	*pendaind*	noun	ī	f	—	—	—	—	—	dil.ie/34272	*pennaind*	noun	—	f	penance
http://molor.eu/data/id/lemma/4807	3262	*pendait*	noun	ī	f	pennait-362	*pennait*	noun	ī2	f	dil.ie/34273	*pennait*	noun	ā	f	
http://molor.eu/data/id/lemma/4638	5506	*eipistil*	noun	ā/i/ī	f	eipistil-551	*eipistil*	noun	ī2	f	dil.ie/20187	*epistil*	noun	—	f	epistle, letter
http://molor.eu/data/id/lemma/4873	8894	*talam*	noun	n	m	talam-889	*talam*	noun	n1	m	dil.ie/39932	*talam*	noun	n	m	earth, land
http://molor.eu/data/id/lemma/4708	7502	*gein*	noun	n	n	gein-750	*gein*	verbal_noun	n2	n	dil.ie/25530	1 *gein*	verbal_noun, noun	—	n, f	the act of pro-cre-ation, birth
http://molor.eu/data/id/lemma/4808	6739	*persan*	noun	ā/n	f	persan-673	*persan*	noun	ā	f	dil.ie/34285	*persa*	noun	n	f	person

Table 1: Comparison of Old Irish nominal lexical entries across CorPH, Goidelex, and DIL resources

[a] Nouns have not yet been systematically aligned with DIL

[b] Also used as the verbal noun of the verb *ro·finnadar* ‘to find out’.

[c] Neuter plural form.

[d] The *x* denotes a cross-reference to the main DIL entry.

[e] Also as common noun with meaning ‘judgment’, as part of the same entry in DIL.

[f] This verbal noun has a gen.sg in *-e* and as such does not fully adhere to *i*-stem inflection, even though this stem designation may be etymologically correct; see Thurneysen (1946, §§256, 294) on the blurred lines between *ā*- and *i*-stem inflection with certain verbal nouns.

[g] For a description of inflectional microclasses see https://github.com/cormacanderson/Goidelex/blob/main/inherent_properties.csv.

URI	CorPH				Kavanagh		DIL				Meaning
	ID	Entry	POS	fl_cat	Entry	POS	ID	Entry	POS	fl_cat	
http://molor.eu/ data/id/lemma/6097	3165	*ad·roilli*	verb	H2[a]	*ad·roilli*	verb	dil.ie/558	*ad·roilli*	verb	—	to deserve
http://molor.eu/ data/id/lemma/6098	6153	*as·roilli*	verb	H2	—	—	dil.ie/4482	*as·roilli*	x[b]	—	
http://molor.eu/ data/id/lemma/6087	318	*báidid*	verb	W2b[c]	—	—	dil.ie/5172	*báidid*	verb	—	to submerge
	3517	*bádaid*	verb	W2a	—	—	—	—	—	—	
http://molor.eu/ data/id/lemma/6170	3711	*cladaid*	verb	S1a	—	—	dil.ie/9297	*cladaid*	x	—	to dig
http://molor.eu/ data/id/lemma/6171	8275	*claidid*	verb	S2	*claidid*	verb	dil.ie/9329	*claidid*	verb	—	
http://molor.eu/ data/id/lemma/5744	5395	*molathair*[d]	verb	W1	*molaidir*	verb	dil.ie/50393	*molaithir*	x	—	to praise
http://molor.eu/ data/id/lemma/5745	—	—	—	—	*molaid*	verb	dil.ie/32491	*molaid*	verb	ā	
http://molor.eu/ data/id/lemma/5906	5893	*taraisnigidir*	verb	W2a	—	—	dil.ie/39730	*tairisnigidir*	verb	g depon	to trust in
	6012	*toraisnigidir*	verb	W2a	—	—	—	—	—	—	

Table 2: Comparison of Old Irish verbal entries across CorPH, Kavanagh, and DIL resources

[a] The classification system used here is from McCone (1987); another widely used classification is Thurneysen (1946).

[b] The *x* denotes a cross-reference to the main DIL entry.

[c] This verb derives from the PIE causative formation *$g^{u̯}oh_2d^h$-*éi̯e-* (Rix and Kümmel, 2001, 206), which explains the class type W2b, reserved for causatives with (mostly) an *o* or *u* as root vowel (McCone, 1987, 28). Synchronically, however, the distinction between W2a and W2b is of little relevance with this verb—albeit confusingly suggestive of a difference—as it has no bearing on either the conjugation pattern or the neutral vs palatal consonance; as such, these lemmas have been merged, each having been assigned a `ontolex:writtenRep` as part of the same URI.

[d] This should be *molaithir* (Prof. David Stifter (Maynooth University), pers. comm.) and has been corrected accordingly; see also Thurneysen (1946, §575).

A Lightweight String Based Method of Encoding Etymologies in Linked Data Lexical Resources

Anas Fahad Khan
CNR-ILC
Italy
fahad.khan@ilc.cnr.it

Maxim Ionov
University of Zaragoza
Spain
mionov@unizar.es

Paola Marongiu
CNR-ILC
Italy
paola.marongiu@ilc.cnr.it

Ana Salgado
NOVA CLUNL
Portugal
anasalgado@fcsh.unl.pt

Abstract

In this submission, we propose an approach to encoding etymological information as strings with formal syntax ("etymology strings"). We begin by discussing the advantages of such an approach compared to modelling etymologies and etymons explicitly as RDF individuals. Next we give a formal description of the regular language underlying our approach as an Extended Backus-Naur Form grammar (EBNF). We use the Chamuça Hindi lexicon as a test case for our approach and show a practical application of the approach using SPARQL queries that extract necessary information from etymological strings.

1 Introduction

Linked Data best practices encourage the use of HTTP URI's to name things and representing every bit of information as RDF triples (statements) based on a formally defined data model. In other words, the preference is for modelling data explicitly in the form of knowledge graphs in which all or most of the entities are represented as RDF classes or individuals each with its own individual URI (individuals can be represented as blank nodes, although this limits their usability). However, certain kinds of data do not lend themselves easily to being modelled this way. This is the case for descriptions of dynamic or temporal phenomena: these latter are most naturally modelled via the addition of a temporal parameter to standard RDF subject-predicate-object triples something which, in general, can only be done via workarounds, e.g., via the reification of properties. Linguistic Linked Data offers many examples of such dynamic phenomena, notably in the form of etymologies, i.e, hypothetical word histories (Khan, 2020). Indeed, when it comes to etymologies, aside from the requirement to represent temporality we have a number of other modelling considerations that make the comprehensive description of such resources potentially very complicated from an RDF point of view. For instance, the following:

- The requirement to represent the hypothetical status of etymologies. In many cases etymologies carry with them a high level of uncertainty and many works will often provide more than one etymology for the same word.
- Related to the previous requirement, in a large number of cases it is important to be able to represent references to the scholarly literature/dictionaries, corpus citations, etc.

In addition, in order to carry out a 'deep' modelling of etymologies, we should, according to linked data best practices, create URIs/individuals for etymons and cognates as well as for etymologies themselves and for cognate sets, that is, to reify all of these elements, as well as adding other elements in order to represent the temporal duration of relationships and properties between individuals. A vocabulary that meets all or most of these requirements will end up being complicated and difficult to use, going well beyond the basic constructions and elements of languages such as RDFS and OWL or more specialsed vocabularies such as OntoLex (see for instance the proposal given in (Khan, 2018)). In addition, such a vocabulary would be hard to create in a theory-agnostic way hence it might end up being unusable for some resources. At the same time, these are fairly standard requirements to cover phenomena that are found in etymological descriptions provided by a lot of lexicographic resources. However, in a large number of use-cases not all the complexity is warranted: What we are looking for are shallow descriptions of etymologies which are of limited complexity and that lend themselves to fast querying and/or processing.[1]

[1] Note that full modelling can still be integrated with this approach using SPARQL UPDATE queries given that there is a vocabulary that can accommodate that.

Consequently, in the current work, instead of proposing a standard, 'deep' RDF based modelling, we propose the use of strings to model etymologies in a 'shallow' way. In order to help ensure interoperability and to facilitate querying of such strings in SPARQL, we define a regular language for representing such 'etymological strings', one that is based on textual conventions for the representations of etymologies in lexicographic works as well as in other kinds of literature.

The rest of the paper is structured as follows. In Section 2 we motivate the use of our approach and provide further details of the kinds of use cases to which we recommend applying this approach. Next, in Section 3 we give a full description of the regular language which we propose as a solution. Afterwards, in Section 4 we give the description of a use case with which we illustrate the way we recommend to query data modelled using this kind of language and present a web interface that uses this approach.

2 Motivation and Use Cases

This work aims to provide a lightweight method of encoding etymological information as string literals. This is particularly useful when (i) the original information is fairly simple; (ii) only a shallow representation of the etymology is required; (iii) a more involved kind of RDF modeling would introduce unnecessary complexity or overhead (and we may not want to e.g., explicitly represent etymologies as hypotheses or model time as a parameter). For instance, in many cases source etymologies will be given in a form similar to the following one for the word *friar* (example from (Durkin, 2009)):

> Latin *frāter* "brother" > Old French *frère* "brother, member of a religious order" > Middle English *frere*, *friar* > Modern English *friar*

where '>' is a standard symbol that marks the direction of the etymological development of the target word (in this case Modern English *friar*).

This example gives a description of the history of a word, but etymologies can also describe other linguistic elements, such as senses as in the following case where the semantic development of the English word is given *sad* (example taken, again, from (Durkin, 2009)):

> Satisfied, having had one's fill (of something) [metamorphized and narrowed] > weary or tired (of something) [borrowed] > sorrowful, mournful

Our argument is that in these, and a large number of similar cases, following a more involved RDF modelling (such as that proposed in (Khan, 2018)) would create too much overhead in terms of RDF triples, given that most often users are interested only in basic kinds of etymological information or an entry as a whole. Indeed, in such simple cases, we could encode the essential information in the form of a string literal allowing users to extract necessary pieces of information via SPARQL queries that make use of regular expressions or via the use of simple string matching functions on the results of a SPARQL query, thus providing complexity-on-demand while preserving all the information. An alternative to this approach (that would also avoid the prolixity which we have just mentioned) would be to define an RDF vocabulary, or series of vocabularies, that captured only some of the requirements which we listed above but that e.g., didn't take the explicit representation of time into consideration or that associated only limited kinds of information with different individual stages of an etymology. However, we feel ours is a cleaner alternative, and one that better fits the current trend towards minimal computing (and of course in more complicated kinds of use cases we would suggest using a more extensive ontology based approach).

In order to preserve interoperability as far as possible, our idea is to define a simple regular language (i.e., a set of strings that can be described with a regular expression) which our etymological strings must belong to. This regular language is based on the simple string representation used in etymological dictionaries, and example of which we saw above. We describe this language in the following section in EBNF for the purposes of explanation.

3 Description of the Etymological String Regular Language

In the rest of the article we lay out and describe our first proposal for an etymological string language. This language has been designed with the following features:

- It allows one or more complete etymologies in a single string, each of which is separated by the '|' symbol; these are alternative etymologies for the lexical element in question (nesting is not possible in a regular language, potentially there could be a context-free superset of this language that allows a shorthand for this using brackets). Each of these etymologies can be associated with a bibliographic source between parentheses.

- Each step is an etymology separated by a '>' symbol and individual steps adhere to the format of: *language code* followed by one or more alternative forms for a *lemma* followed by one or more *senses* separated by an ampersand '&'.

- We allow for an additional specification of each step with a transition note between square brackets.

According to this language, our previous *friar* example can be rewritten as follows:

> lat frater > fro frere 'brother' & 'also member of a religious order of 'brothers'' > enm friar, frere > eng friar (Source: Durkin) .

Similarly we can rewrite the *sad* example above as follows:

> 'Satisfied, having had one's fill (of something)' [metamorphized and narrowed] > 'weary or tired (of something)' [borrowed] > 'sorrowful, mournful'

This language is clearly limited in the kinds of etymological information which can be captured. At the same time it fits a large number of simpler cases as found in many lexicographic works. The clear benefit of the formalism is that it provides a middle ground between human- and machine-readability: while still looking familiar to lexicographers and etymologists, it can be validated and processed with efficient and simple computational methods.

EBNF Grammar

In this section we present an EBNF grammar of our regular language[2]. This permits us to give a precise formal definition of our language in a fairly (for humans) understandable and transparent way, something that would not be the case if we presented our language as one long regular expression.

```
etymologies = etymology , { "| " , etymology } , [ " " ],"."
      ;

etymology = step , { [ transition_note ] , ">" , [ " " ] ,
      step } ;

transition_note = "[" , printable_no_quote_seq , "]" ;

step = [ lang , " " ] , ["? ", " "],
      [ lemmas , " " ] ,
      [ senses , " " ] ,
      [ source ] ;

lang = letter , { letter | "-" } ;

lemmas = lemma , { ", " , lemma } ;

lemma = letter , { letter } ;

senses = "‘" , sense , "’" , { " & " , "‘" , sense , "’" } ;

sense = printable_no_quote_seq ;

source = "(Source:" , printable_no_quote_seq , ")" ;

letter = ? UnicodeLetter ? ;

printable_no_quote_seq = printable_no_quote , {
      printable_no_quote } ;

printable_no_quote = ? isPrintable & notQuote ? ;
```

As will be seen, our language already allows us to capture a lot of different kinds of etymologies (as hopefully the next section will demonstrate). However there may be small elements which will be added in subsequent versions to make the formalism even more useful.

4 The CHAMUÇA Test Set

As a test set for our regular language we decided to use the CHAMUÇA Hindi language lexicon (CHAMUÇA-Hi), one of the outputs of an ongoing project, CHAMUÇA, which aims to trace the impact of Portuguese on the languages of Asia (Khan et al., 2024). CHAMUÇA-Hi consists of just over a hundred entries in Hindi, each of which at least plausibly derive from an original Portuguese etymon.[3] For many of these entries there are alternative etymologies positing a non-Portuguese origin of the entry in question. Overall, however, the etymological information included in the dataset is fairly shallow; it is therefore ideal for showcasing our lightweight approach.

We had already converted our dataset into RDF

[2] Note that although we have used EBNF to present our langauge it is regular without introducing features that would place it higher in the Chomsky hierarchy.

[3] The lexicon can be downloaded here: https://github.com/anasfkhan81/Etymologies_as_Strings/blob/main/chamuca_hi_lex.ttl.

(Khan et al., 2024), but decided to generate another version with the addition of etymological strings which follow the regular language proposed in this article, associating the entries with their etymologies using the lexinfo[4] `etymology` property. For instance, see the following entry for the word अनन्नास (anannaas) meaning 'pineapple':

```
अनन्नास:_entry a ontolex:LexicalEntry,
   ontolex:Word ;
lexinfo:domain <http://lari-datasets.ilc.cnr.it/
   chadoms#botany> ;
lexinfo:etymologicalRoot <http://lari-datasets.ilc.cnr.
   it/chamuca_pt_lex#ananás> ;
lexinfo:etymology "tpn naná ''pineapple (Source:
   Wiktionary) > pt ananás ''pineapple (Source:
   Dalgado) ." ;
lexinfo:gender lexinfo:masculine ;
lexinfo:partOfSpeech lexinfo:commonNoun ;
frac:frequency [ a frac:Frequency ;
      rdf:value 0 ;
      frac:observedIn :hiTenTen21 ] ;
ontolex:canonicalForm अनन्नास:_lemma ;
ontolex:lexicalForm अनन्नास:अनन्नास_dp_form_,
   अनन्नास:अनन्नास_os_form_,
   अनन्नास:अनन्नास_vs_form_,
   अनन्नासो:अनन्नास_vp_form_,
   अनन्नासों:अनन्नास_op_form_ ;
ontolex:sense अनन्नास:_sense .
```

Other examples of etymological strings from the dataset include:

`"pt ? carabina (Source: Dalgado) | French carabine (Source: Dalgado) | fa qarābīn (Source: McGregor) ."`

and

`"pt ? baptismo (Source: Dalgado) | en baptism (Source: Dalgado) ."`

The question mark at the beginning of each lemma here signals the fact that the etymology is regarded as being doubtful in the source itself.

In each of these cases, the creation of an RDF graph explicitly encoding the same etymological information would have meant creating individuals for each of these etymons as well as for the etymology itself with a high cost in the number of resulting triples. Encoding the information as strings as we have done in this case, we are still able to extract quite a lot of relevant etymological information either via SPARQL queries or using basic string processing functions from different programming languages. For instance we can write a simple query which gives the number of alternative entries for each etymology:

```
PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/
      lexinfo#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex
      #>

SELECT ?entry
      ((STRLEN(STR(?etymology)) - STRLEN(REPLACE(
      STR(?etymology), "\\|", ""))) + 1 AS ?
      numAlternatives)
WHERE {
  ?entry a ontolex:LexicalEntry ;
        lexinfo:etymology ?etymology .
}
```

Another example would be a query which searches of all words which potentially have an ancient Greek etymon:

```
PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/
      lexinfo#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex
      #>

SELECT ?entry ?etymology
WHERE {
  ?entry a ontolex:LexicalEntry ;
        lexinfo:etymology ?etymology .

  # grc in the beginning or after >
  # or after |
  FILTER (
    REGEX(STR(?etymology),
        "(^|>|\\|)\\s*grc\\b")
  )
}
```

Using a combination of queries like this users can extract specific parts of etymologies in which they are interested. An additional advantage of this approach is that the users are not limited to using SPARQL queries to process the data: while etymological strings has to be extracted with SPARQL, further processing can be done in a variety of ways, thanks to wide support of regular expressions by software and programming languages.

To demonstrate a possible way to use etymological strings in a user-friendly way, we created a web interface that provides basic etymological data for a chosen entry from this dataset.[5]

5 Conclusions and Future Work

In this article we have presented a first draft of our work on etymological strings, a formalised way to represent etymological information in a string,

[4] https://lexinfo.net/

[5] https://anasfkhan81.github.io/Etymologies_as_Strings/.

without expanding it to complex graph representations. While unconventional, it is a convenient and efficient way to hide unnecessary data complexity without losing it, all while providing the data in both machine- and human-readable way.

In the future we plan to enrich these strings with further kinds of information such as e.g., part of speech and gender. However we feel that the proposed formal language is already suitable for a large number of use cases. However we do not want to make the language or strings too complicated since this would defeat the purpose of using our approach in the first place. Also in the future we plan to look into whether our approach is also useful for other kinds of information, i.e., morphological information.

Acknowledgments

The work is in part supported by the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

References

Philip Durkin. 2009. *The Oxford Guide to Etymology*. OUP Oxford.

Anas Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12):304.

Anas Fahad Khan. 2020. Representing Temporal Information in Lexical Linked Data Resources. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 15–22, Marseille, France. European Language Resources Association (ELRA).

Anas Fahad Khan, Ana Salgado, Isuri Anuradha, Rute Costa, Chamila Liyanage, John Philip McCrae, Atul Kumar Ojha, Priya Rani, and Francesca Frontini. 2024. Chamuça: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages. In *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 44–48.

Ontolex-Lemon in Wikidata and other Wikibase instances

David Lindemann
UPV/EHU University of the Basque Country
Unibertsitateko Ibilbidea, 5
01006 Vitoria-Gasteiz
david.lindemann@ehu.eus

Abstract

This paper provides insight into how the core elements of the Ontolex-Lemon model are integrated in the *Wikibase Ontology*, the data model fundamental to any instance of the Wikibase software. This includes *Wikidata lexemes*, which today is probably the largest Ontolex-Lemon use case, a dataset collaboratively built by the community of Wikidata users. We describe how lexical entries are modeled on a Wikibase, including the linguistic description of lexemes, the linking of lexical entries, lexical senses and lexical forms across resources, and links across the domain of lexemes and the ontological part of a Wikibase knowledge graph. Our aim is to present Wikibase as a solution for storing and collaboratively editing lexical data following Semantic Web standards, and to identify relevant research questions to be addressed in future work.

1 Introduction

Wikibase,[1] a set of extensions to MediaWiki[2], is a software solution for storing, collaboratively editing and exhibiting structured data on the web, in the shape of a knowledge graph. The software is used, first and foremost, by Wikidata (Vrandečić and Krötzsch, 2014; Erxleben et al., 2014).[3] Many other instances of the software have emerged since the software packages (and hosting solutions) are freely available.[4] The types of entities described in a Wikibase include *items*, which represent all kinds of real-world objects and ontological concepts, and *lexemes*, describing words. As it will be explained in section 2.1, in a Wikibase, lexemes are described following the core of the Ontolex-Lemon model (McCrae et al., 2017), and so they are on Wikidata, which is today probably the largest open and collaboratively editable Ontolex-Lemon use case. By May 2025, Wikidata described 1.42 Million lexical entries in 1,379 different languages, which compared to earlier figures (Nielsen, 2020) means an exponential growth. German, Russian, Danish, Estonian, English and Malayalam are, in this order, the languages with the most described lexemes. The potential uses of the linguistic descriptions contained in this growing resource, which we present in more detail in section 2.2, are manifold.

Section 2.3 is devoted to Wikibase as a linking hub for descriptions of lexical *entries*, *senses* and *forms* across different resources; cross-resource links are encoded as *external ID* properties.[5] External identifiers do not only constitute hyperlinks for a user to jump between the resources presented in different web portals, but also enable SPARQL federation,[6] that is, a Wikibase's content may be integrated also with the lexical and ontological content of any other Wikibase, including Wikidata, or an RDF database of other kind.[7] In this regard, it is important to point out how RDF is used on a Wikibase, which will be explained in section 2.4. Related to this, an important and distinguishing feature of Wikibase is its multi-layer integration of lexical and ontological entities inside the same database, which will be discussed in detail in section 2.5.

While the lexemes collection on Wikidata is open for continued enrichment, some use cases may require a separate Wikibase instance, in a first phase of a contribution project, or even for a whole project lifetime: A language may be already described on Wikidata, so that any addition would involve lexeme, sense and form disambiguation

[1] See https://wikiba.se.

[2] See https://mediawiki.org.

[3] See https://www.wikidata.org.

[4] See https://wikiba.se/showcase/ and https://wikibase.world, a catalogue of Wikibase instances.

[5] See https://www.wikidata.org/wiki/Wikidata:External_identifiers.

[6] See https://www.w3.org/TR/sparql11-federated-query/.

[7] See https://www.mediawiki.org/wiki/Wikibase/Federation.

and deduplication tasks, which when working on an own instance could be left to the final phase of a project. Also, a dataset to be curated may be regarded too noisy (e. g. in a legacy dictionary digitisation project), or too esoteric (e. g. when dealing with dialectal or historical language data) to be included in Wikidata without previously ensuring relevance and quality. In addition, to work on an own Wikibase instance means freedom in data modeling and community management. Licensing might also be an issue, since Wikidata content is obligatorily licensed according to CC0. Exploring these potentials, we briefly discuss the Wikibase ecosystem for lexeme descriptions in section 2.6.

In the closing section 3, we provide an outlook for open research questions to be worked on, focusing on the relation between the fine-grained modeling proposals made by the Ontolex Community, on the one hand, and the conventions emerging in the community working on Wikidata lexemes, on the other.

2 Ontolex-Lemon on Wikibase

2.1 Lemon core classes

The core of Ontolex-Lemon,[8] that is, the classes `ontolex:LexicalEntry`, `ontolex:LexicalSense`, and `ontolex:Form`, is reused in the Wikibase Ontology,[9] the backbone model of any Wikibase instance.

Wikibase treats the lexical entry, with its own numeral identifier preceded by letter L, as primary entity describing a *lexeme*, with forms and senses as sub-entities, and presents the instances of those three Ontolex-Lemon core classes together on one editable lexeme page.[10] This structure is pre-set in the data model fundamental to any Wikibase instance and cannot be modified by the user, and the same is true for a small number of properties to be attached to the three core classes, listed in table 1; the three obligatory properties describing a lexical *entry* must have a value, a *sense* must have a *gloss* (a short sense-disambiguating text, represented in RDF using `skos:definition`), and a *form* must have a representation (a value for `ontolex:representation`).[11]

`ontolex:LexicalEntry`
`wikibase:lemma`
`wikibase:lexicalCategory`
`dct:language`
`ontolex:sense`
`ontolex:lexicalForm`
`ontolex:LexicalSense`
`skos:definition`
`ontolex:Form`
`wikibase:grammaticalFeature`
`ontolex:representation`

Table 1: Obligatory basic classes and properties in their domain for describing lexemes in Wikibase

In detail, this obligatory structure entails the following restrictions:

- A lexical *entry* must point to exactly one item in the same Wikibase as value for *lexical category*.

- A lexical *entry* must point to exactly one item in the same Wikibase as value for its *language*.

- A lexical *entry* must have at least one *lemma*. Several lemmata can co-exist for the same *entry* for covering different spelling variants (e. g. British and American English).[12] In the RDF representation, lexeme lemmata appear attached to the entry using a property named `wikibase:lemma`. Lemmata are indexed for the MediaWiki text search index,[13] so that a user can search for lexemes, manually in the interface, or via API.

- A lexical *sense* must have at least one *sense gloss*, in any language, i. e. not necessarily or not only in the language associated to the *entry*. Purpose of the gloss is to provide the information necessary to discriminate word senses.

- A *form* must have at least one *written representation*. Alike different values for *lemma*

[8] For the original Lexical Model for Ontologies, see `https://lemon-model.net/`.

[9] See `https://wikiba.se/ontology/`.

[10] See `https://www.wikidata.org/wiki/Lexeme:L1` for the lexeme page describing lexeme `wd:L1`.

[11] Although `ontolex:writtenRep` would be the best match here, the Wikibase Ontology uses `ontolex:representation`; in Ontolex-Lemon, the former is defined as subproperty of the latter.

[12] See an example at `https://www.wikidata.org/wiki/Lexeme:L1347`, where the English *color/colour* is described, an example for an entry with lemmata in distinct scripts at `https://kurdi.wikibase.cloud/wiki/Lexeme:L3447`.

[13] See `https://www.mediawiki.org/wiki/Elasticsearch`.

on entry level, form representations in several spelling variants can be attached to the same *form* entity. A written representation in most cases will be a string as found in text of that language, but written representations also include code transcriptions, such as those describing sign language forms.[14] To describe a form without providing any type of written representation is not foreseen.

- A *form* can have zero or more items in the same Wikibase as values for *grammatical feature*.

Lemmata, *sense* glosses (skos:definition), and also *form* representations (ontolex:representation) are associated to a language code. The available codes are the same that are available throughout the mediaWiki instance, e. g. for labels, descriptions, and *monolingualtext* strings.[15]

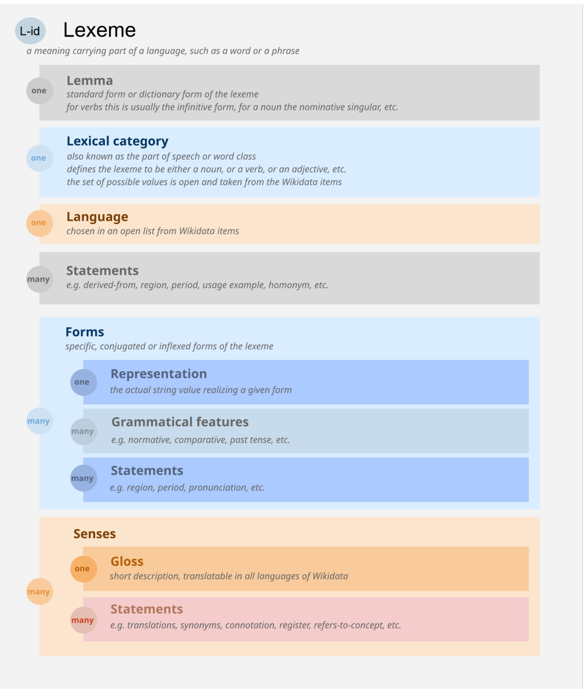

L-id
Lexeme
a meaning carrying part of a language, such as a word or a phrase
one
Lemma
standard form or dictionary form of the lexeme
for verbs this is usually the infinitive form, for a noun the nominative singular, etc.
one
Lexical category
also known as the part of speech or word class
defines the 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: The Wikibase lexeme, as illustrated in the Wikibase documentation

These restrictions guarantee interoperability on a basic level. Beyond that, any additional relation involving *entry*, *sense* or *form* objects is not predefined and can be modeled according to the use case using self-defined properties in Wikibase *statements*,[16] as illustrated in Fig. 1.[17] The RDF classes and properties mentioned above are part of the Wikibase Ontology, that is, they are used in the RDF representation of an entity (an *item*, a *lexeme*, a *property*), and accordingly, they appear in RDF entity data dumps.[18] In opposition to that, all other relations added to *lexeme*, *sense* or *form* as Wikibase *statements* (see section 2.4) always involve a property defined in the namespace of the Wikibase instance itself, identified by a number preceded by the letter P, and that has a range restricted to one Wikibase datatype.[19]

2.2 Wikidata lexemes

Wikidata lexemes is an open and editable collection, where everybody is invited to collaborate. A documentation of the data model based on the three Ontolex-Lemon core classes is given on the Wikidata documentation pages.[20] Concerning advanced modeling questions, contributors get support from each other.[21] A core group of more experienced and active users provides advice to newcomers and occasional contributors, also through dedicated outreach events.[22] A manually curated list of lexical entries provides examples in several languages of good and complete modeling practice.[23]

Instead of following prescribed models concerning morphology, etymology, multilingual equivalents, etc., on Wikidata, a bottom-up grown set of properties is used for describing entries, senses, and forms; see table 2 for the ten most frequent properties for each of the classes, pointing to other

[14] For an example, see wd:L991786-F1.

[15] See https://www.wikidata.org/wiki/Help:Monolingual_text_languages, and for a list of all codes implemented in a Wikibase instance, e. g. https://www.wikidata.org/w/api.php?action=query&meta=wbcontentlanguages.

[16] See https://www.wikidata.org/wiki/Help:Statements.

[17] The picture is used at https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model.

[18] In Turtle serialization, see, for example, https://www.wikidata.org/wiki/Special:EntityData/L1347.ttl.

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

[20] See https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/Data_Model.

[21] Discussions take place on-wiki (see https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data) and in a dedicated channel on the *Telegram* platform.

[22] The pages dedicated to the 2021 and the 2024 *Lexicodays* provide video recordings, presentation slides and links to other pages containing lexicographical guidelines and descriptions of tools related to Wikidata lexemes, see https://www.wikidata.org/wiki/Wikidata:Events/Lexicodays_2024 and https://www.wikidata.org/wiki/Wikidata:Events/30_lexic-o-days_2021

[23] See https://www.wikidata.org/wiki/Wikidata:Showcase_lexemes.

entities on Wikidata, or to a data value (in table 2, excluding *external id* properties). Without going into much detail, we point out some of them; in some cases, the modeling is unambiguous (and straightforwardly alignable to Ontolex), and sometimes alternative modeling approaches coexist:

- As it will be explained in section 2.5, **translation equivalence** is expressed using `wd:P5972`, a property set (in both directions) directly between senses; the property used in Ontolex-Lemon for this purpose is `vartrans:translation`. In May 2025, Wikidata contained 119,847 translation links between senses of different languages.

- For representing **etymology**, a lexeme is linked directly to the lexical entry describing the etymon using `wd:P5191`.[24] In May 2025, Wikidata contained 40,540 etymology links from lexeme to lexeme.

- For representing **decomposition**, `wd:P5238`, the *combines lexeme* property links a compound or multiword entry to its constituents, in the same way `decomp:subterm` is used according to Ontolex.[25] This property is used in May 2025 207,799 times.

- **Pronunciation** is described in multiple ways, but always associated to *forms*; the value of a general *pronunciation* property `wd:P7243` is in general identical to the form representation, but may include stress indicators and other diacritics, e. g. to indicate vowel length. It is recommended that pronunciation audio files and/or IPA transcriptions are attached as *qualifiers* to the pronunciation claim, but audio files can also be directly attached to *form*.[26] Today, 74% of 252,652 pronunciation audio files are linked directly to a *form*. Only audio files hosted on Wikimedia Commons can be used.[27]

- As seen in table 2, properties devoted to certain **transcription** and **transliteration** systems exist; each of them is, like all Wikidata properties, described through its entity data, and on an own discussion page.[28] In parallel, a general property *transliteration or transcription* (`wd:P2440`) can be used, and its value qualified using the property *determination method or standard* (`wd:P459`). The former, defined as subproperties of `wd:P2440`, are heavily used on Wikidata in general,[29] but in only about 1.2% of its use cases (i. e., not more than 3,808), that property is attached to *forms*,[30] on which the general property is used around ten times more often: The 70,165 uses of `wd:P2440`, with a `wd:P459` qualifier pointing to the transliteration system, are almost half-divided between *items* and *forms*.[31]

- **Usage examples** (`wd:P5831`) are recommended to be attached to *entry*, and not to *sense*. However, about 7% of the 31,271 usage examples in Wikidata lexemes today remain attached to *sense*; the distribution across languages shows a diverse picture, but a clear preference for *entry* as subject of the property.[32] If attached to *entry*, one or more senses can be declared *subject sense* (`wd:P6072`); this is done using that property as qualifier in the example statement. The advantage of this modeling, apart from being able to declare a usage example to be pertinent to more than one sense, lies in the ability to have examples attached to the entry also if in the moment of upload and without or before whatever sense disambiguation procedure it is not clear which sense should be marked as *subject sense*, e. g. when dealing with examples stemming from corpora, or if (still) no senses are described for the lexeme: As soon as the correct sense can be determined, the `wd:P5831` *claim* is enriched with with a `wd:P6072` *qualifier*, without having to delete and re-write the whole *statement* (with its references), and attach it to *sense*. Another strong reason for attaching

[24] See, for example, `https://www.wikidata.org/wiki/Lexeme:L630740#P5191`.

[25] See `wd:L625224` for the German phrase "frohes neues Jahr".

[26] For an example, see `wd:L3338-F2`. *Qualifiers* as part of a *statement* are explained in section2.4

[27] The property is of datatype *Commons Media File*, see `https://www.wikidata.org/wiki/Help:Data_type#commonsMedia`.

[28] For example, about the property `wd:P5825` *ISO 15919 transliteration*, a documentation is available at `https://www.wikidata.org/wiki/Property_talk:P5825`, and about `wd:P4187` *Tibetan to Latin transliteration*, `https://www.wikidata.org/wiki/Property_talk:P4187`.

[29] See `https://w.wiki/EBWF`.

[30] See `https://w.wiki/EBYi`.

[31] See `https://w.wiki/EBZ2` for usage counts according to the different transliteration systems.

[32] See `https://w.wiki/ECBm` for the use of this property across languages.

usage examples to entry is to enable their annotation with a subject *form*, i. e. the word form appearing in the example.[33] In addition, having examples at both levels complicates their retrieval in queries.

The Wikidata lexemes collection is quantitatively described on dedicated pages,[34] and it can be explored using the *Ordia* tool (Nielsen, 2019),[35] which generates dictionary-like exhibitions of lexical data; it also features a tool to look up Wikidata *forms* matching to tokens in text. *Ordia* also provides statistics on the Wikidata lexeme collection, such as, for example, lists of the most frequently used properties in the domains of *entry* (a. k. a. *lexeme*), *sense* and *form*.[36] The *Synia* tool also shows statistics on lexemes, e. g. counts of values for `wd:P6191` *language style* in different languages.[37]

Table 3 lists overall counts for the ten languages with best absolute coverage at the three levels.[38] Asking for relative coverages, without data about the total amount of corpus lemmata, corpus types, and dictionary senses in a language, we might ask for a relation between the coverage on Wikidata and the number of speakers of a language. A query like that (taking into account languages with more than 100,000 speakers) results in a different ranking, with Estonian, Breton, Basque and Danish leading lexemes,[39] and Basque, Breton, Dagbani and Norwegian Bokmål as top four languages for senses.[40]

Apart from those already mentioned, a range of tools[41] has emerged around Wikidata lexemes, designed to help creating entries and append senses,[42] for linking senses to ontological references,[43] for creating forms collections through templates,[44] searching for usage examples and adding them to

LexicalEntry	
`wd:P5185`	*grammatical gender*
`wd:P5911`	*paradigm class*
`wd:P5238`	*combines lexeme*
`wd:P31`	*instance of*
`wd:P5187`	*word stem*
`wd:P1552`	*has characteristic*
`wd:P5402`	*homograph lexeme*
`wd:P2348`	*time period*
`wd:P5191`	*derived from lexeme*
`wd:P5186`	*conjugation class*
`wd:P5831`	*usage example*
`wd:P7486`	*grammatical aspect*
LexicalSense	
`wd:P5137`	*item for this sense*
`wd:P5972`	*translation*
`wd:P5973`	*synonym*
`wd:P1343`	*described by source*
`wd:P18`	*image*
`wd:P9488`	*field of usage*
`wd:P6191`	*language style*
`wd:P8394`	*gloss quote*
`wd:P9970`	*predicate for*
`wd:P6271`	*demonym of*
`wd:P6084`	*location of sense usage*
`wd:P10339`	*semantic gender*
Form	
`wd:P7243`	*pronunciation*
`wd:P898`	*IPA transcription*
`wd:P443`	*pronunciation audio*
`wd:P5279`	*hyphenation*
`wd:P5825`	*ISO 15919 transliteration*
`wd:P8881`	*ITRANS (Indic scripts)*
`wd:P8530`	*alternative form*
`wd:P5276`	*Slavistic Phonet. Alphab. transcr.*
`wd:P10822`	*homophone form*
`wd:P1721`	*Hanyu Pinyin transliteration*
`wd:P11950`	*appears before phonolog. feat.*
`wd:P11951`	*appears after phonolog. feat.*

Table 2: The 12 most frequently used properties describing Wikidata lexemes, senses and forms (May 2025)

[33] See, for example, the usage examples for `wd:L87`.

[34] See `https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Statistics`

[35] Accessible at `https://ordia.toolforge.org/`.

[36] Accessible at `https://ordia.toolforge.org/property/`. Wikidata SPARQL queries as used in *Ordia* can be modified for custom searches, e. g. for listing the property statistics for a single language, as in the following queries derived from those in *Ordia*: *entry*, `https://w.wiki/E4pg`, *sense*, `https://w.wiki/E4pw`, *Form*, `https://w.wiki/E4p$`; note the filter for *external id* properties.

[37] See `https://synia.toolforge.org/#languagestyle`.

[38] See up-to-date statistics at `https://w.wiki/ECLq` for *lexeme*, at `https://w.wiki/EDkc` for *sense*, and at `https://w.wiki/EDjh` for *form*.

[39] See `https://w.wiki/EDna`.

[40] See `https://w.wiki/EDnQ`.

[41] See also `https://www.wikidata.org/wiki/Wikidata:Tools/Lexicographical_data`.

[42] For a lexeme creation UI, see `https://hangor.toolforge.org`; for python, in addition to the general *wikibaseintegrator* library (`https://github.com/LeMyst/WikibaseIntegrator`), specially for lexemes, *tfsl* (`https://gitlab.wikimedia.org/toolforge-repos/twofivesixlex`.)

[43] See `https://lexica-tool.toolforge.org/`.

[44] See `https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms`.

LexicalEntry	language	LexicalSense	language	Form	language
1,422,331	all Wikidata	**585,893**	all Wikidata	**14,396,207**	all Wikidata
239,465	German	48,308	Bokmål	2,802,106	Estonian
102,150	Russian	41,973	English	1,257,083	Basque
96,753	Danish	30,756	Basque	1,246,745	Russian
83,218	Estonian	29,790	Nynorsk	1,199,194	Latin
68,758	English	24,127	Czech	871,813	Czech
67,367	Malayalam	24,055	Swedish	753,118	Malayalam
64,445	Italian	22,000	Italian	692,671	Spanish
63,128	Spanish	21,036	Japanese	641,454	Danish
56,296	Latin	20,484	Persian	571,091	German
48,214	Swedish	18,851	Egyptian	522,107	Italian

Table 3: Absolute coverage of some languages in Wikidata lexemes (May 2025)

entries,[45] for massively recording pronunciation audios,[46] or for graph visualisations of lexical relations.[47]

One goal of the Wikidata lexemes collection is to enable natural language generation for drafting Wikipedia article text from abstract knowledge representations (Vrandečić, 2021; Morshed, 2024),[48] and another possible application is corpus annotation (Lindemann and Alonso, 2024); all these depend on the degree the lexemes, senses and forms in the collection cover the languages to process.

2.3 Wikibase lexemes as linking hub: The case of Wikidata

In addition to these and other properties for the linguistic description of the lexeme, Wikibase lexical entries contain pointers to external resources encoded as *external id* properties. These lead a human user to an entry or sense description in a dictionary web portal. In some cases, federated database queries can access content in several graph databases at a time using such external ID. For example, a query can involve Wikidata and the LiLa Latin Knowledge Base (Passarotti and Mambrini, 2022),[49] exploiting the LiLa URI attached to Wikidata Latin lexemes, and calling the LiLa SPARQL endpoint from within the Wikidata Query Service, or vice versa.

Wikidata lexemes, on *entry* level, by May 2025 count 2.3 Million *external id* statements. Table 4 shows usage counts for the most frequent 20 properties.[50] As for *sense*, all lexical resources aligned to Wikidata on sense level today still lack significant coverage,[51] although we point out the fact that e. g. English Wordnet synset identifiers are aligned to Wikidata *items* (McCrae and Cillessen, 2021), which, as explained, are referenced by a significant number or senses. An alignment of *forms* to external resources such as corpus-based form repositories is at large still not present, although it would be interesting, for instance in rich-morphology languages, where the morphologically possible forms outnumber the forms that are actually attested in corpora, so that forms attestation is very valuable information, with a similar value lemma attestation has in languages with a comparably reduced number of different inflected forms, like English.

2.4 Reification in Wikibase

Wikibase *statements* include by default a mechanism for further describing the main *claim* of a statement using *qualifiers*, *ranks*, and *references*. A graphical model of a Wikibase statement is given in figure 2, where "entity" represents an *item*, *lexeme*, *sense*, *form*, or *property* node, each of which has its own URI in the main entity namespace of the Wikibase instance,[52] and where the blue-colored "value" nodes, depending on the property datatype, represent entities of the same Wikibase, or data val-

[45] The *Luthor* tool uses Wikisource content as corpus, see https://luthor.toolforge.org/.

[46] See https://lingualibre.org/.

[47] For an example, an etymological network, see https://lucaswerkmeister.github.io/wikidata-lexeme-graph-builder/?subjects=L184995&predicates=P5191.

[48] See https://meta.wikimedia.org/wiki/Abstract_Wikipedia.

[49] See https://lila-erc.eu; the LiLa ID property is wd:P11033.

[50] For up-to-date counts, see https://w.wiki/E9j3.

[51] An example for one of the most used external identifiers is *DWDS sense ID*, see https://www.wikidata.org/wiki/Property_talk:P12550.

[52] On Wikidata, e. g. wd:Q1 for an item, wd:L1 for a lexeme, wd:L1-S1 for a sense, wd:L1-F1 for a form, and wd:P31 for a property.

Count	Property	Property label (en)	Lang. (ISO-639-1)
231,859	wd:P9940	*DWDS-Lemma-Identifikator*	de
153,322	wd:P8376	*Duden-Lexem-Identifikator*	de
139,067	wd:P11519	*elexiko ID*	de
122,350	wd:P11138	*Sõnaveeb entry ID*	(45 lang.)
84,951	wd:P9947	*WDG-Lemma-Identifikator*	de
66,512	wd:P13258	*Presisov večjezični slovar ID*	fr de en sq sh
65,599	wd:P9529	*Den Danske Ordbog article ID*	da
61,333	wd:P5912	*Oqaasileriffik online dictionary ID*	da en kl nb
54,755	wd:P12630	*Aragonario ID (6th version)*	es an
51,566	wd:P11033	*LiLa Linking Latin URI*	la
49,987	wd:P9385	*DWB Lemma ID*	de
45,751	wd:P5275	*OED Online ID*	en
39,052	wd:P9962	*Ordbog over det danske sprog ID*	da
37,925	wd:P12420	*Il Nuovo De Mauro ID*	it
37,137	wd:P10042	*Bokmålsordboka-ID*	nb nn
36,535	wd:P11838	*Svenska Akademiens ordlista ID*	sv
35,124	wd:P9387	*GWB Lemma ID*	de
31,014	wd:P12690	*New Oxford American Dictionary ID*	en
29,803	wd:P11319	*Little Academic Dictionary ID*	ru
29,316	wd:P12828	*DAKA Danish-Greenlandic Dictionary ID*	da

Table 4: Most frequently used *external id* properties on Wikidata lexemes, and lexeme languages

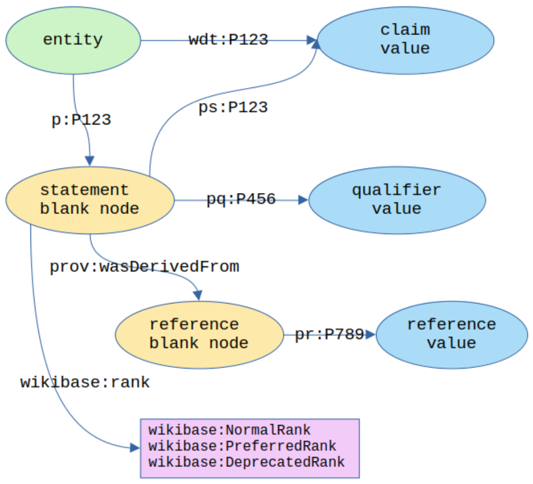


Figure 2: Graphical model of a Wikibase Statement

ues, including strings, external identifiers, date objects, globe coordinates, et cetera. Each statement node may be linked to several *qualifier* values, and to several *reference* nodes (in the RDF representation attached to the statement blank node using prov:wasDerivedFrom), which make up blocks of references. *Ranks* are used for annotating multiple statements for the same property with one of the three values *normal*, *preferred*, or *deprecated*.[53]

In order to keep all data pertaining to the same lexical entry "together", so that it would all get stored in the same entity data JSON blob,[54] and displayed on the same lexeme entity page, it is convenient to keep the modeling of the linguistic description as shallow as a list of Wikibase *statements*. Unlike in the Ontolex modules, where the linguistic description often involves several reification layers, the model followed in Wikidata, and, for the same reason, in any other Wikibase, will try, wherever possible, to stay with certain reification approaches, which can be combined, but will be limited to the following:

- Using statement *qualifiers*, i. e. semantic triples describing the main *claim* of the statement.

- Using *references* for provenance annotations.

- Using subproperties such as wd:P1721

[53] See https://www.wikidata.org/wiki/Help:Ranking.

[54] At https://www.wikidata.org/wiki/Special:EntityData/L1.json, the entity data JSON for wd:L1 as stored in the database, and from which the entity page display is generated, can be obtained.

Hanyu Pinyin transliteration, subproperty of wd:P2440 *transliteration or transcription*.

If we compare, for example, how a translation relation between two senses can be further described, according to Ontolex this can be modeled introducing a blank node of class vartrans:Translation into the dataset (option A), linking that to both *sense* nodes using typed properties (*source* and *target*), and attaching to that node supplemental information about the translation relation (Bosque-Gil et al., 2015). As a shallower alternative (option B), which does not involve any additional blank node in the structure, and consequently allows no additional description of the translation relation, Ontolex uses the vartrans:translation property.[55] The second option is more suitable for a Wikibase, because the additional node in option A, since it does not fit into the Wikibase statement structure, would have to be created as named individual entity, i. e. a Wikibase *item* with its own Q-identifier, its labels to be indexed in the Wikibase ElasticSearch, its class declaration, and descriptions. However, in a Wikibase, option B caters for a description as rich as option A in Ontolex: In the discussed example, the translation relation type, its direction, or any other translation restriction feature can be expressed using a *qualifier* on the translation *claim*. This, in turn, is not possible when using option B in Ontolex, since the semantic triple describing the translation relation there cannot be further described or qualified.

2.5 Lexicon-Ontology interface and multilinguality

According to Ontolex-Lemon, a property named ontolex:reference links a lexical sense to an ontological concept (Bosque-Gil et al., 2015). On Wikidata, wd:P5137 *item for this sense* has the equivalent function. For example, Wikidata's lexeme wd:L3549, describing the English noun *foot*, has three senses, each of them pointing to a different conceptual *item* node in the graph, using that property. wd:L3549-S1, the first listed sense, is linked to an item describing a unit of length, while the second sense points to an item describing a furniture part, and the third sense, wd:L3549-S3, glossed as "anatomical structure found in vertebrates", links to the anatomical entity. Each of the three linked Wikidata *items* is itself further described, for example, with links to Wikipedia articles in multiple languages (the property used here is schema:about), which have a title in that language, and which provide encyclopaedical descriptions of the concept. That means in general terms that Wikibase provides a framework where lexical and ontological (conceptual) descriptions converge, and where text pages (for Wikidata, Wikipedia articles) *about* concepts also have their habitat. Since, in addition, Wikibase items are annotated with multilingual labels (rdfs:label and skos:altLabel) and descriptions (schema:description), the wd:P5137 reference of a lexeme sense into the ontological part of the Wikidata graph already provides three facets of multilinguality: The labels, textual concept descriptions, and entire text pages attached to an ontological item referenced by lexeme senses may cover multiple languages.

Since several languages' word senses can be linked to the same *item*, wd:P5137 provides translation (and, inside the same language, synonymy) information. In Ontolex, this way of modeling translation relations is referred to as *translation as shared reference*.[56] As of May 2025, Wikidata contains 227,908 *item for this sense* claims that link senses to items.[57] Calculating the number of translation links through shared references for every item, that sums 5.23 million translation (and intralingual synonymy) links between lexeme senses, counting the connecting links twice, i. e. as translation link in both directions.[58] This by far outnumbers wd:P5972 *sense translation* relations (see section 2.2), which constitute an alternative without leaving the domain of lexemes, as needed for senses without an ontological reference in Wikidata (most prominently, senses of lexemes with a lexical category other than *noun*). As described above, multilingual sense *glosses* (very short sense descriptions attached using the built-in skos:definition) provide another facet of multilingual sense description.

Fig. 3 shows an example for a lexeme's relations in the graph, including *monolingualtext* values in two languages: the lexical entry describing a German noun *Pferd* ("horse"), linked to entities of dif-

[55] See examples with figures for both at https://www.w3.org/2016/05/ontolex/#translation-as-a-relation-between-lexical-senses.

[56] See https://www.w3.org/2016/05/ontolex/#translation-as-shared-reference.

[57] See the distribution of *item for this sense* across languages at https://w.wiki/ECCA.

[58] See https://w.wiki/EcDh for a list of all items linked to from senses, and the corresponding number of translation links (includes intralingual translation, i. e. synonymy).

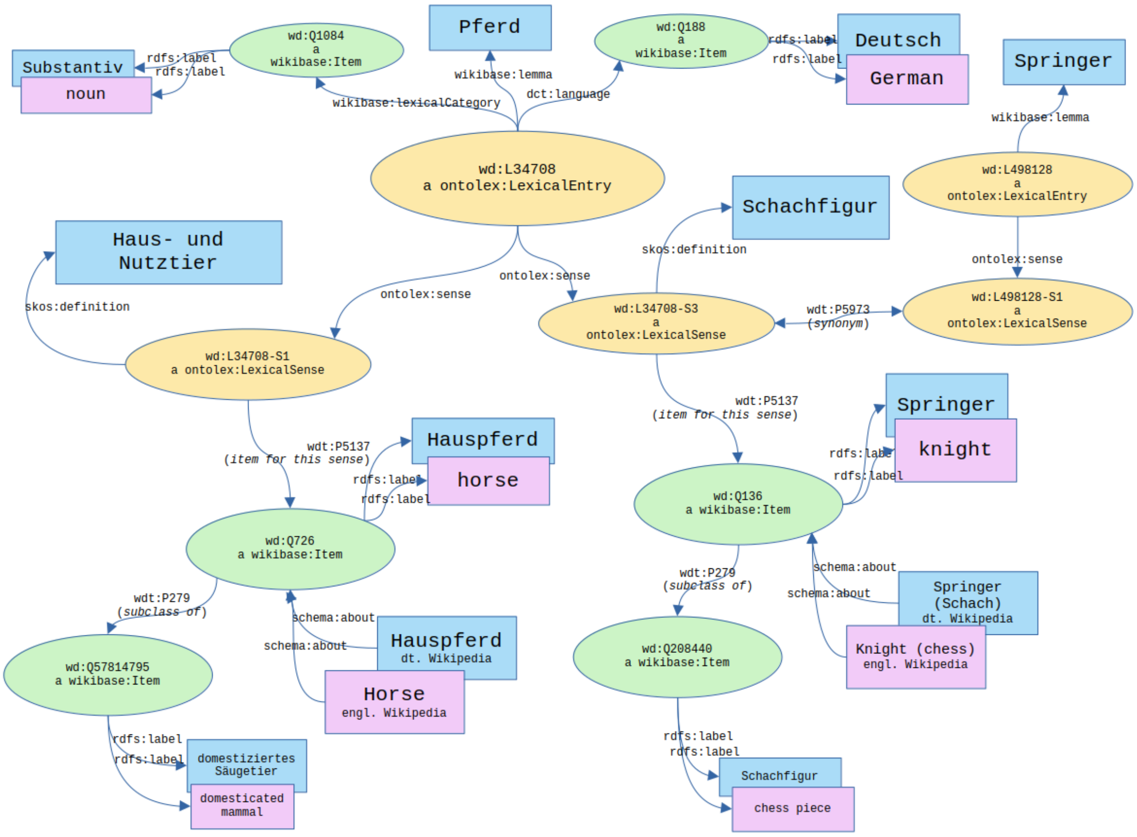


Figure 3: Wikidata entry wd:L34708 describing the German lemma *Pferd* and some relations

ferent types: ontological concepts (*items*), other lexical entries (*lexemes*), and lexical senses. Entities of type *item* are used to represent the *language*, which allows querying for lexemes according to language features such as the language family or countries where languages are spoken or native to, to represent the *lexical category*,[59] and, as explained above, for ontological sense references. *Label*, *definition* and *lemma* values are always strings associated to a language code (figure 3 only shows English and German, while the cited entities on Wikidata cover more languages here).

2.6 Wikibase as an infrastructure for lexical datasets

In terms of the *FAIR Guiding Principles for scientific data management and stewardship*[60], a lexical dataset on a Wikibase can safely be called to be state of the art, since permanent URI on the level of the three Ontolex-Lemon core classes assure *findability*, answering human user calls with the display of an editable entity page, and programmatical calls with machine-readable data in various formats.[61] *Accessibility* is furthermore given through the graphical query service,[62] and through a SPARQL endpoint. *Interoperability* is sustained by re-using W3C-recommended RDF vocabularies for a set of basic classes and properties defined in the Wikibase Ontology, such as Ontolex-Lemon for lexical data. Finally, *reusability* is ensured by open licenses.[63]

Wikimedia Deutschland, the WMF chapter responsible for Wikidata, is providing the Wikibase software for self-hosting,[64] and also provides a

[59] On Wikidata, a range of 320 different *items* is used here, some defined as subclasses of others, e. g. wd:Q1166153 *intransitive verb subclass of* wd:Q24905 *verb*; see https://w.wiki/ECKP for use counts.

[60] See https://www.go-fair.org/fair-principles/

[61] Entity data dumps are available in JSON and TTL format.

[62] For Wikidata, accessible at https://query.wikidata.org/.

[63] Wikidata declares its terms of use at the bottom of every displayed page, and in detail at https://foundation.wikimedia.org/wiki/Policy:Terms_of_Use, and any other Wikibase may contain declarations in a similar form.

[64] See https://www.mediawiki.org/wiki/Wikibase/Docker.

hosting cloud.[65] In community discussions, the vision connected to that embraces an ecosystem of independent but federated Wikibases, with Wikidata as central linking hub.[66] Regarding several domains of knowledge, this is already becoming reality; also related to lexical data, some projects have been able to showcase the potential this infrastructure provides for interlinked, FAIR datasets, in experiments with lexical datasets derived from different kinds of sources, such as dictionaries in CSV format (Huaman et al., 2023), Ontolex-Lemon TTL (Lindemann et al., 2023), and, recently, DMLEX (Krek et al., in press).

3 Outlook to further research

In this paper, we have been revisiting the model for lexical entries on Wikibase, and in Wikidata lexemes, the largest collection of lexical data on a Wikibase today, and which is also, probably, the largest Ontolex use case. It can be stated that, by reusing Ontolex-Lemon, the Wikibase software enables the community to perform steps towards the vision of a *Linked Lexical Data Cloud* (Declerck, 2018).

Since the first publications of the Lemon model (McCrae et al., 2012), which had been available by the time of defining the Wikibase Ontology, the Ontolex community has been publishing modeling proposals as modules, regarding, among other aspects, the description of morphology, etymology, and corpus attestations.[67] In the same timespan and in parallel, there has been emerging a tradition of modeling lexemes on Wikidata. One research question seems obvious in this context: Where and how do both models differ, where do they come together? Where can they benefit from each other? What advantages and disadvantages have the two strategies, compared to each other: A top-down model with strict recommendations, aiming at interoperable lexical datasets also at a more fine-grained level (Ontolex), and a model that limits obligatory interoperability to the core, leaving decisions regarding fine-grained descriptions up to the user (Wikibase), aiming at higher levels of interoperability through on-the-fly grassroot community discussions (Wikidata)? This can shed lights on questions about whether the Lemon core has proven its functionality, if the obligatory core should be extended, or if even some of the minimum requirements might turn out problematic for certain use cases. Comparing both universes may lead to useful insight: Can the Wikidata lexemes collection provide data-driven evidence for modeling decisions that can be regarded universal, beyond the Ontolex-lemon core? And, in the other direction: Can collaborators or automatic tools informed in Ontolex-Lemon modeling proposals help grassroot communities to improve consistency, quality assessment and interoperability? A continued dialogue between the Ontolex and the Wikidata lexemes communities, and those around other instances of the Wikibase software, will help to address these questions in detail.

Acknowledgements

The research leading to the presented results has been supported by DLTB research group (Basque Government, IT1534-2).

References

Julia Bosque-Gil, Jorge Gracia, Guadalupe Aguado-de Cea, and Elena Montiel-Ponsoda. 2015. Applying the OntoLex Model to a Multilingual Terminological Resource. In *The Semantic Web: ESWC 2015 Satellite Events*, pages 283–294.

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.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the Linked Data Web. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014*, number 8796 in Lecture Notes in Computer Science, pages 50–65. Springer International Publishing, Cham.

Elwin Huaman, David Lindemann, Valeria Caruso, and Jorge Luis Huaman. 2023. QICHWABASE: A Quechua Language and Knowledge Base for Quechua Communities.

Simon Krek, Primož Ponikvar, Andraž Repar, Iztok Kosem, and David Lindemann. in press. DMLEX on Wikibase: Legacy dictionaries as collaboratively editable dataset. In *Proceedings of eLex 2025: Electronic Lexicography in the 21st Century - Intelligent Lexicography*, Bled.

[65] See https://wikibase.cloud.

[66] See https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Joint_Vision.

[67] See an overview and source data at https://github.com/ontolex/ontolex.

David Lindemann, Sina Ahmadi, Anas Fahad Khan, Francesco Mambrini, Federicia Iurescia, and Marco Carlo Passarotti. 2023. When OntoLex Meets Wikibase: Remodeling Use Cases. *CEUR Workshop proceedings*, 2773.

David Lindemann and Mikel Alonso. 2024. Linking Historical Corpus Data and Annotations using Wikibase. In Kristina Štrkalj Despot, Ana Ostroški Anić, and Ivana Brač, editors, *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*, pages 743–748. Institute for the Croatian Language, Zagreb.

John Philip McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. The lemon cookbook.

John Philip 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.

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

Mahir Morshed. 2024. Using Wikidata lexemes and items to generate text from abstract representations. *Semantic Web*, 16.

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. Springer, Cham.

Finn Årup 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.

Marco Carlo Passarotti and Francesco Mambrini. 2022. Linking Latin: Interoperable Lexical Resources in the LiLa Project. In *Building new resources for historical linguistics*, pages 103–124. Pavia University Press, Pavia.

Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Communications of the ACM*, 64(4):38–41.

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

Philosophising Lexical Meaning as an OntoLex-Lemon Extension

Veruska Zamborlini* and Jiaqi Zhu and Marieke van Erp
DHLab, KNAW Humanities Cluster, Amsterdam, the Netherlands
*Ontology & Conceptual Modeling Research Group, Federal Univ. of Espírito Santo, Brazil
{veruska.zamborlini,jiaqi.zhu,marieke.van.erp}@dh.huc.knaw.nl

Arianna Betti
Universiteit van Amsterdam, Amsterdam, the Netherlands
a.betti@uva.nl

Abstract

OntoLex-Lemon is a model for representing lexical information, focusing on the use of lexical entries in texts rather than their definitions. This work proposes an extension to the model that aims to capture the definition of senses attributed to lexical entries. We explicitly represent a conceptual setup authored by an agent that operates on lexical content. It either proposes new senses for existing lexical entries in a language or coins new terms to express proposed senses. It provides textual and/or formal definitions to senses/concepts, and can serve as an interpretation of other senses/concepts through rephrasing, translation, formalization, or comparison. Because a conceptual setup and its interpretations may not be unanimously accepted, it is important to support the selection of relevant meanings, as for example, those proposed by a certain author. We illustrate the application of our proposed extension with two case studies, one about the philosophical definition of the concept of *idea* and its interpretations, and one about historical attributions of meaning to the Dutch East India Company (VOC).

1 Introduction

The OntoLex-Lemon[1] W3C recommendation for representing lexical information focuses on the various usages of lexical entries in texts. While this approach has proven effective in many contexts, it was not designed to capture the definitions that underpin the lexical senses attributed to the entries. Several extensions have been proposed to enhance the expressiveness of the Ontolex-Lemon model in different aspects, such as capturing morphological decomposition (*decomp* module), representing translations and lexical variation (*vartrans* module), describing metadata about lexical resources (*lime* module), and also linking multilingual linguistic resources (through Linguistic Linked Open Data (LLOD) initiatives) (Khan et al., 2022; Gromann et al., 2024). However, to the best of our knowledge, none of them directly addresses the need to represent the definitional and interpretative foundations of lexical senses or concepts.

To address this gap, we propose an extension to the OntoLex-Lemon model at a conceptual level, that is, not yet implemented. The extension enables explicit representation of a conceptual setup providing meaning attributed to lexical entries as textual or formal definitions by original authors and/or by other authors interpreting the original ones. In this work we use the term *conceptual setup* as generic label for a (loose) view/conceptualization (e.g., a term coined in journalism), an expert-level conceptualization/theory (e.g., a domain-specific definitions in a scholarly text), or a fully developed theory (e.g., a formal philosophical framework). We also refer to definition of *lexical sense* and *lexical concept* somewhat interchangeably, as the latter is typically lexicalized through the former in a particular language. Finally we consider interpretation as rephrasing, translating, explaining or formalizing someone else's conceptual setup with the intention of preserving the intended meaning, as opposed to (i) intentionally changing the meaning (as in correcting or complementing it) or (ii) directly/originally describing a conceptual set up (as in "interpreting reality").

While not all lexical senses have a specific source/author for their definitions, and usage may diverge from original definitions, our proposed extension aims to systematically capture those definitions and their interpretations for which there is traceable and verifiable evidence. This approach thus aims to enrich representations of lexical meaning, ultimately supporting the analysis and understanding of how concepts evolve over time.

To illustrate and motivate our proposal, we present two case studies from the digital humanities domain in Section 3. The parallel of these two

[1] www.w3.org/2016/05/ontolex/

cases is that both combine computational methods and digital humanities expertise to deal with the challenge of how concepts evolve or change over time in their specific domain, namely philosophy and history. They address similar research questions, such as: How have certain concepts changed/evolved? What kind of changes have they undergone? And how to model these conceptual changes in a way that is computationally manageable and interpretable for answering humanities research questions? By introducing our extension to the OntoLex-Lemon model applied to case studies from the domains of philosophy and history, we demonstrate its potential for supporting and enriching digital humanities research in general. However, the extension is broadly conceived and therefore applicable to other domains where lexical definitions also evolve or diverge, such as medicine, law, and science (Oortwijn et al., 2021).

The first example is from the *eIdeas* project within the "Concepts in Motion" lab,[2] a group aiming to trace computationally how concepts evolve over time. In philosophy, (re)interpreting a theory, or a concept within a theory, often requires a profound and concrete understanding of the implications behind the words that are used to formulate the concept or theory in question. During the interpreting process, interpreters usually need to assign meanings to the word/lexicon that is core to the concept/theory according to their understanding. Based on different underlying assumptions or philosophical perspectives, interpreters can have various interpretations and applications of the original concept/theory. To have new insights or approaches to a concept/theory, philosophers usually need to engage with the arguments and counterarguments that have been proposed about this concept/theory over time. In this background, our extension to the OntoLex-Lemon model can help trace the evolution and (re)interpretations of a philosophical concept/theory by providing a dynamic and multifaceted perspective on its meaning and applications.

The second example is from the "Trifecta" project,[3] which combines computational linguistics and semantic web technologies to extract and model, from the maritime and food history domains, concepts in their contexts, such as the Dutch East India Company, slavery, coffee, and cinnamon. (van Erp, 2023) points out that Large Knowledge Graphs (KGs) such as Wikidata and DBpedia only express a limited representation of the concepts and entities they represent. For instance, at the time of writing (van Erp, 2023), DBpedia focuses in representing the concept *coffee* on the food dimension, while it could be explored through multiple aspects, such as a plant, the activity of drinking the drink, a colonial good, and more. The project aims to automatically capture different dimensions of concepts in various contexts and represent this multi-dimensionality in Knowledge Graphs. Towards this goal, Trifecta focuses on dealing with key challenges: a. identity (what the concept is and how it is perceived), b. change (how this concept evolved over time), and c. the long tail (what low-frequency contexts are connected to this concept). Linguistic information supported by the Ontolex-Lemon ontology can play a role in tackling these challenges. The schema in Figure 1 that illustrates scholarly and historical texts representing different meanings and interpretations attached to the concept of the VOC.

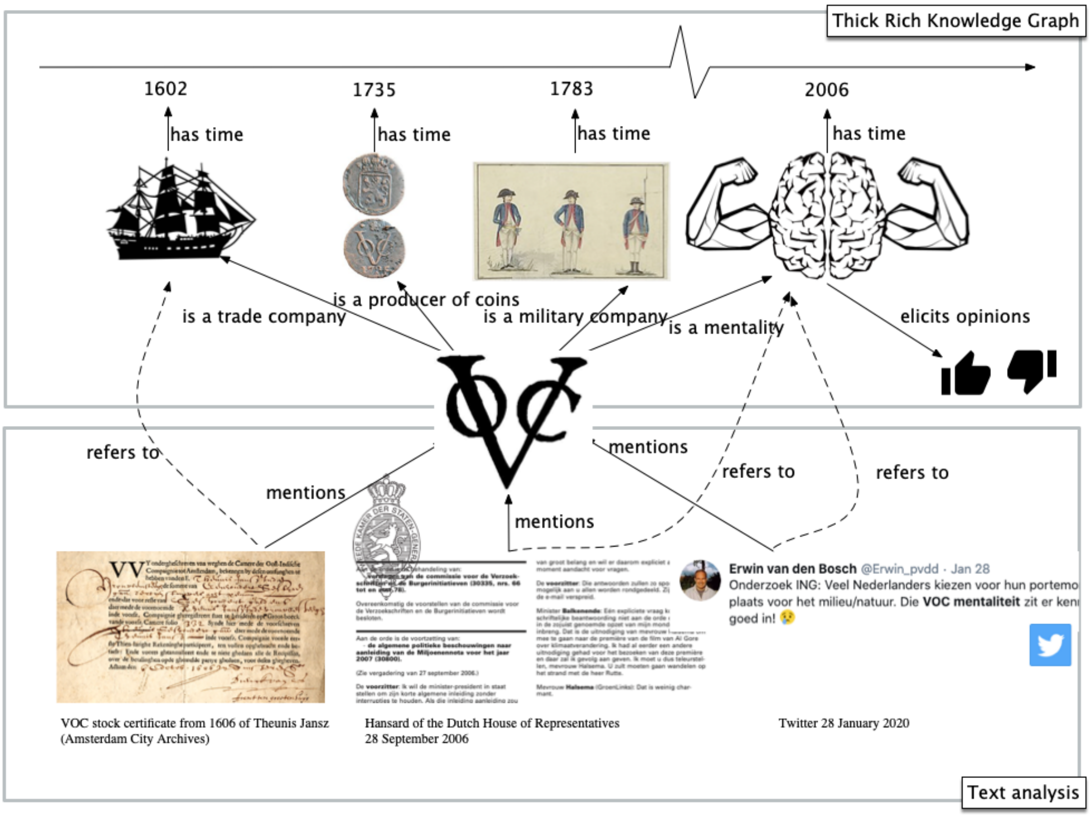
The remainder of this paper is organised as follows. In Section 2 we discuss related work, followed by two case studies stemming from digital humanities scenarios and competency questions for the proposed model to address in Section 3. In Section 4 we present our proposed extension to the Ontolex-Lemon model by providing UML representation, and we illustrate our extension with two schemas that instantiate the model for our case studies. We present our discussion revolving around the relationship to other modules and models in Section 5 and to what extent the competency questions are addressed in Section 6. Our conclusions and directions for future work respectively are presented in Section 7.

2 Related Work

Ontolex-Lemon (Mccrae et al., 2017) results from an effort of the Ontology Lexicon (Ontolex) community group becoming a W3C standard model for providing rich linguistic grounding to ontologies. It provides means to connect ontology entities to lexical entries with their morphological and syntactic properties. It is designed to be combined with the other four OntoLex modules: syntax and semantics (***synsem***), decomposition (***decomp***) variation and translation (***vartrans***) and linguistic metadata (***lime***). Several modules and extensions to Ontolex are reviewed in (Gromann et al., 2024) (and similar

[2] https://conceptsinmotion.org/

[3] https://trifecta.dhlab.nl/



Erwin van den Bosch @Erwin_pvdd · Jan 28
Onderzoek ING: Veel Nederlanders kiezen voor hun portemo
plaats voor het milieu/natuur. Die **VOC mentaliteit** zit er ken
goed in! 😢

Figure 1: Schematic illustration of meanings VOC can take on over time

examples throughout).

The ***synsem*** module is concerned with providing semantics to the lexical entries by connecting them to existing ontologies that may provide a formal specification to constrain the meaning of a concept. The goal of our proposal is (i) not to rely on the existence/adequacy of an OWL[4] ontology, but on documented conceptualizations/theories backing the senses attributed to the entries; (ii) to allow for formal definitions to be provided in formal languages beyond OWL and (iii) to allow for an OWL ontology to be derived from well-annotated theoretical sources ultimately providing also a detailed provenance for the concepts in the ontology.

The ***vartrans*** module models translation as a relation between senses, defining an exact (non-questionable) correspondence between them. In contrast, we propose an interpretation relation among lexical concepts that accommodates subjectivity and variation. It could ultimately serve as translations when the provided interpretation is accepted as valid.

The ***lime*** module (Linguistic MetaData) provides a standardized way to describe metadata about linguistic resources, such as the lexicon or the conceptualization set. It includes information about the language(s) covered, the number of lexical entries, the structure of the lexicon, and its linkage to other resources. This module supports interoperability and facilitates the discovery and integration of lexical datasets on the Web. However, it does not suffice to address the requirements for evidence supporting the attribution of meaning to lexical entries as envisaged in our proposal.

(Khan et al., 2022) provides an overview of research projects that use linguistic linked data vocabularies to create and publish lexical resources in various languages using the OntoLex-Lemon model (and its extensions). The paper introduces representative projects across various domains and use cases, including digital humanities, and discusses the influence of these projects on the use or definition of linguistic linked data models and vocabularies in detail. Two examples of related projects designed for the lexical modeling of historical domain-specific vocabulary are DitMAO–LexO–MAIA (Giovannetti et al., 2024) and ALMA (Tittel, 2023). Both initiatives share a similar goal: to capture the context in which certain senses were used (or proposed), particularly within historical or scholarly sources. However, their so-

[4] https://www.w3.org/OWL/

lutions do not explicitly model the provenance, interpretation, and formalization of definitions.

3 Case Studies and Competency Questions

This section presents two case studies that were devised with domain experts in the humanities domain including a set of competency questions for the proposed model to address.

3.1 Bolzano's Theory of Ideas & Interpretations

In his book *Wissenschaftslehre* (1837) (Bolzano, 1837), the Bohemian philosopher Bernard Bolzano proposes a theory in which he defines, among others, the term *Vorstellung* as "*Vorstellung [ist] dasjenige, was als Bestandtheil in einem Satze vorkommen kann, für sich allein aber noch keinen Satz ausmacht.*" (Bolzano, 1837) §. 48, which is translated as "that which can occur as a component in a sentence, but which on its own does constitute a sentence".

More than a century later, the Italian philosopher Betti, in their book chapter "Bolzano's Universe: Truth, Logic and Metaphysics (2012)" (Betti, 2012), renders Bolzano's *Vorstellung* as *Idea*, and rephrases the definition as "*an idea is that part of a proposition that is not itself a proposition*".

This interpretative chain continues: Betti's student Hungerbühler, in his thesis "A computational method for philosophical interpretation (2018)" (Hungerbühler, 2018), offers yet another layer by formalizing the concepts from Betti's interpretation using OWL Description Logics, such as the formal definition of *Idea* described in Listing 1. By reasoning over these formal definitions, Hungerbühler's thesis provides interesting insights on the definitions of concepts and their interpretations.

```
Class: Idea
    SubClassOf: partOf some Proposition
    DisjointWith: Proposition
```

Listing 1: Manchester OWL Syntax Example

This chain of provenance is essential, for example, when discrepancies arise: if an inconsistency is found using formalisations such as Hungerbühler's, the issue can be traced back to verify if it stems from his own reinterpretation, from Betti's interpretation, or from Bolzano's original theory. This case study illustrates possible benefits of a representation that allows for keeping the provenance of the original documents from which the definitions are taken along with the chain of interpretations. Furthermore the ability to use formal syntaxes (besides OWL) to describe concepts allows for later extraction of a formal model as input for reasoners and analysis of the results.

3.2 VOC as a "company-state"

In the 17th and 18th centuries, the Vereenigde Oostindische Compagnie (Eng., *Dutch East India Company*) (VOC) played an important role in early modern world history. The VOC was set up in the Dutch Republic as a trading company to trade with and in Asia, and soon created a trading network of colonies and settlements in Asia and Africa (Gaastra, 2003). Relying on the archives of the VOC as source material, historians discuss the VOC and its role in history from different perspectives, for example, in early modern global trade (Israel, 1989), in cross-cultural encounters (Blussé, 1986), and also in colonisation in Asia and Africa (Schrikker, 2007; Emmer, 2003). However, the VOC is a complex concept as it has conducted various kinds of activities and thus can be interpreted in various ways. In this paper, we focus on a certain historical perspective which understands the VOC as both commercial and political for its functions in both, and show how our extension to the model can help represent the relationship between different interpretations.

Inspired by historian Philip Stern's analysis of the English East India Company (EIC) as a "company-state" in the book (Stern, 2011), historian Arthur Weststeijn argued that the VOC should also be considered as a "company-state" as in (Weststeijn, 2014). Reinterpreting Stern and Weststeijn's "company-state" arguments, historian Erik Odegard further applied this perspective of understanding the VOC both as a ruler and a merchant in formulating his argument in (Odegard, 2020). Similar to the first case study on Bolzano's Theory of Ideas and Interpretations, our extension to the OntoLex model provides a structure that allows various interpretations or perspectives on the VOC to be presented and compared, which historians could benefit from. We propose this structure to enable researchers to work with interpretive complexity rather than flattening concepts under investigation. Our extension links interpretations to their authors and sources, which we expect will enable computational tracking of how arguments develop and circulate. The intended outcome is that researchers will be able to analyze not just what previous scholars claim about a concept, but how these claims relate to different theoretical frameworks and textual sources. We anticipate this will create new possibilities for understanding historical knowledge.

3.3 Competency Questions

We defined the following competency questions with the domain experts. These questions were chosen based on their relevance for the type of research the domain experts want to conduct and serve as guidance and evaluation for the types of information our extension needs to cover.

CQ1 What are all the definitions of a given lexical entry, along with their direct or indirect authors? *(e.g., Idea as defined by Betti or Hungerbühler and VOC as defined by Weststeijn and Odegard)*

CQ2 Which concepts have been (re)defined by a particular author? *(e.g., all concepts defined by Bolzano or Odegard)*

CQ3 What are the various interpretations that have been proposed for a specific conceptual setup? *(e.g., Betti's interpretation of Bolzano's theory; or Odegard interpretations of Stern and Weststeijn's theories.)*

CQ4 What are all the interpretations proposed by a specific author? *(e.g., all interpretations by Betti, for example, for Bolzano, as well as all interpretations by Odegard for example, for Stern and Weststeijn.)*

CQ5 What is the formal representation of all concepts included in a conceptual setup? *(e.g. Hungerbühler's formal interpretation of the Theory of Ideas in Manchester OWL syntax)*

CQ6 How has a concept evolved over time, both in general and through contributions by particular authors? *(e.g. how the several definitions provided for the concept VOC have evolved in time and through different narratives, or have an author such as Bolzano provided different definitions or refinements for the concept of Idea in different works)*

CQ7 Which definitions of terms are closer or farther away from each other? How close are they? *(e.g. how is the definition of Bolzano for Idea close the one by Aristotle, or is it closer to Aristotle's than Locke's definition? Or yet, how the definitions of the VOC relate to each other, as in agreement, complementarity, contradiction or others).*

4 OntoLex-Lemon Extension

The proposed extension is depicted in Figure 2 using a UML[5] diagram. The classes and relations from Ontolex-Lemon and its modules are prefixed accordingly (*olex* as short for *Ontolex*) and depicted in shades of green and yellow, while the proposed ones are not prefixed and are depicted in purple color. Our proposal is intended as a modular extension, specializing or complementing the entities of the OntoLex-Lemon framework, enabling the representation of the provenance of lexical entries, senses and concepts, and the modeling of interpretive or derivational relationships between them.

A *View/Conceptualization* is composed of *Defined Lexical Concepts* lexicalized as *Defined Lexical Senses*. It is *authored by* an *Agent* and *authored at* at a certain point in time (*temporal extension*) in an *Creation Event* possibly *described in* a *Document*. It is *expressed as* a *Lexicon* and may also coin a *CoinedLexicalEntry*. When the author is an *Specialised Agent* (for the subject in question), then the Conceptualization can be considered as a *Theory*. If it provides *FormalisedLexicalConcept* with a *formal definition*, it is then a *FormalisedTheory*. A formalization can be provided in any language/syntax, such as OWL Manchester Syntax[6] or SWI-Prolog[7]. As long as they are provided with the appropriate "language" annotation, e.g. @*manchester* or @*swiprolog* or ^^*:manchester* or ^^*:swiprolog*, a script can select the formal definition of selected concepts to compose an output description that can be input for a proper reasoning service.

Moreover, a *View/Conceptualization* can be an *Interpretation* of one or more *Views/Conceptualizations*, if it provides *LexicalConcepts* that are *interpretations of* concepts in other *Views/Conceptualizations*. If it provides interpretations for all the entries in another *View/Conceptualization*, the *Lexicon* expressing the interpreting one can be modeled as providing an *interpretation of* the *Lexicon* expressing the other under interpretation.

To illustrate the proposed extension, we present two schemas that instantiate the model for our case studies. The color code refers to the respective classes according to Figure 2, having the lexical entries and their forms grouped in a gray box. First, Figure 3 illustrates the case study **Bolzano's Theory of Ideas & Interpretations** (subsection 3.1).

[5] https://www.uml.org/

[6] www.w3.org/TR/owl2-manchester-syntax/

[7] www.swi-prolog.org/

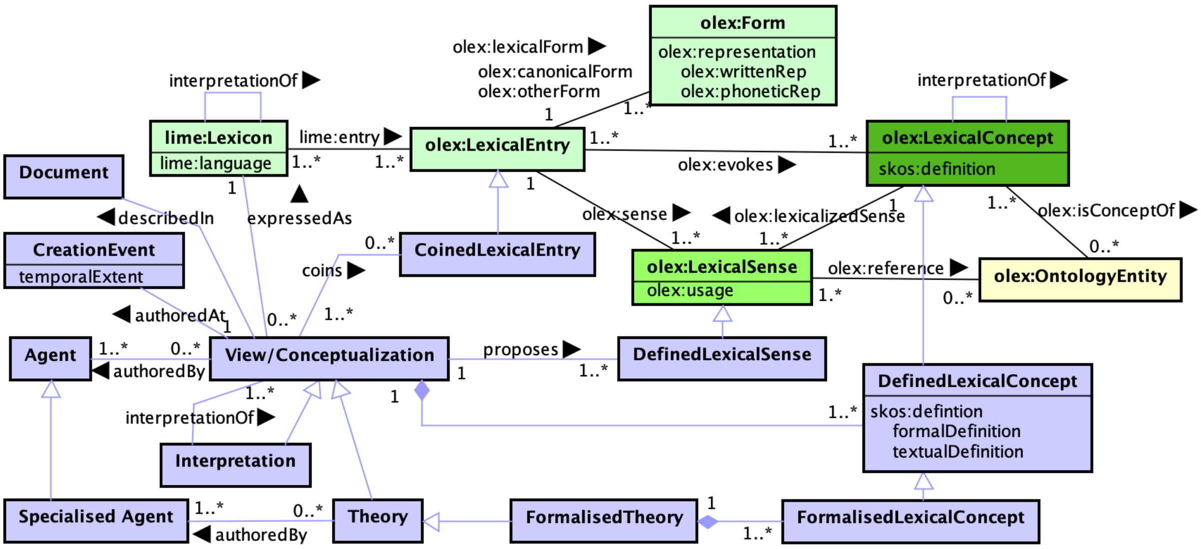


Figure 2: UML representation of (part of) OntoLex-Lemon model (indicated with prefixes and depicted in green shades and yellow) and the proposed extension (depicted in purple).

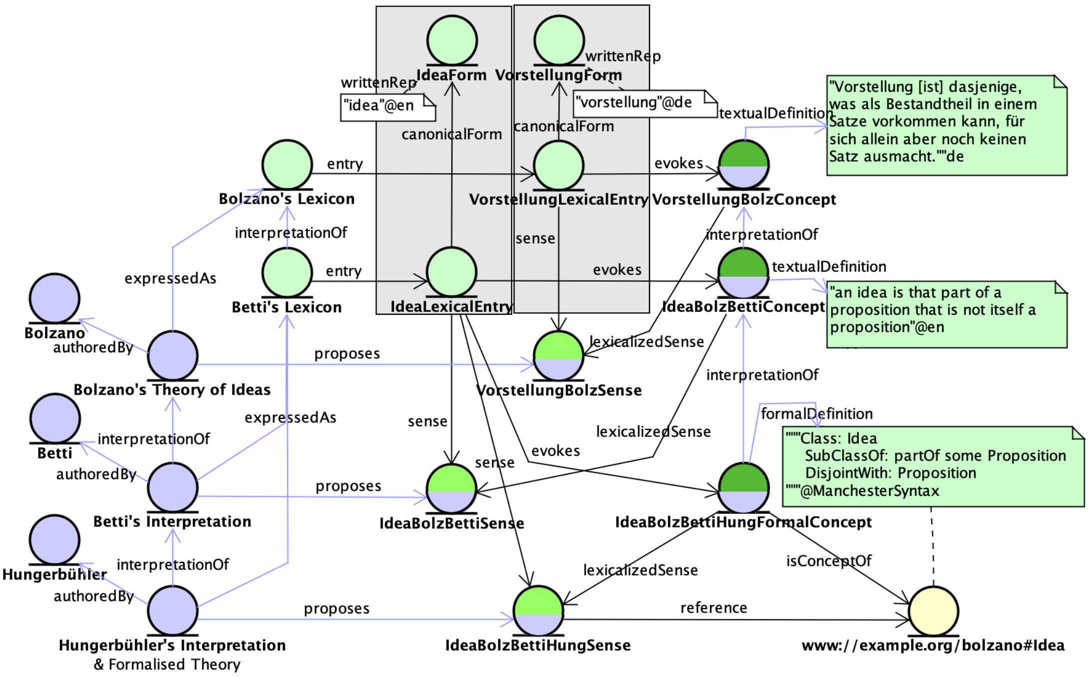


Figure 3: Schema representing an instantiation of the model for Bolzano's Theory of Ideas and its interpretation by Betti. Specifically for the entry *Vorstellung* in German, Betti proposes it as *Idea* in English rephrasing its definition, while Hungerbühler provides an interpretation of Betti's interpretation with a formalization in Description Logics using Manchester syntax (the color code refers to the respective classes in Figure 2).

The lexical entry for which the canonical form is *Vorstellung* in German, has its corresponding sense proposed by the mentioned theory, which is the lexicalized sense of a lexical concept having as definition the original text by Bolzano (detailed web annotation provided later). This provides provenance for the sense, namely, that it originates from Bolzano's theory.

Analogously, Betti's theory proposes a sense for the lexical entry with canonical form *Idea* that evokes the lexical concept whose textual definition in English is the rephrasing in the original text by Betti. It is an interpretation of the lexical concept lexicalized by the sense proposed by Bolzano. Rather than asserting that the senses for *Idea* and *Vorstellung* refer to exactly the same sense or are a translation of each other, which would imply a perfect equivalence, we instead represent that *Idea*, as proposed in Betti's chapter, as linked to a distinct sense that is an interpretation of Bolzano's original sense. In this way, we preserve both the nuance of interpretation and the provenance of each contribution.

Furthermore, because Betti's entire chapter is dedicated to interpreting Bolzano's work, we model their theory as an interpretation of Bolzano's original theory. The lexicon that expresses Betti's theory thus provides interpretations of the lexical entries that express Bolzano's theory, or translations for them if Betti's interpretations are taken as valid.

Finally, Hungerbühler's theory provides a formalization in OWL-DL that is an interpretation of Betti's theory, which in turn interprets Bolzano's original theory. The sense proposed by Hungerbühler is thus expressed through an OWL class (identified here by the illustrative URI `www://example.org/bolzano#Idea`) which is defined as equivalent to the formalization proposed by him. This formalization not only establishes a semantic anchor for the sense in question but also enables its use in automated reasoning tasks. By expressing the definition in a formal language such as OWL-DL (e.g., in Manchester Syntax), an OWL ontology can be generated and reasoned over using standard semantic web tools. It is important to note that although OWL-DL was selected for this particular case study, the proposed approach is not restricted to it; any other formal representation language could be employed to capture the definitions and support similar reasoning workflows.

Next, Figure 4 illustrates the case study **VOC as a "company-state"** (subsection 3.2). Here we have two lexical entries, for which the form is *EIC* in English (English abbreviation for the "English East India Company") and another one for which the form is *VOC* in Dutch (Dutch abbreviation for the "Dutch East India Company"). First, Stern proposes a sense for *EIC* entry that evokes the lexical concept whose textual definition in English describes it as a company-state. Next Weststeijn applies Stern definitions as an analogy to the *VOC* concept, actually proposing to it also a sense to the corresponding entry that evokes a similarly defined lexical concept. Finally, Odegard agree with them both, rephrases their definitions applied to both *EIC* and *VOC* entries, this providing a reinterpretation of the lexical concepts lexicalized by the senses proposed by Stern and Weststeijn. Important to notice, an dotted red arrow connecting Stern and Weststeijn's concepts is meant to express the analogy relation between the concepts, which however has not being yet included in our proposal and is therefore object of investigation for future work.

Finally, Figure 5 illustrates how the Web Annotation Vocabulary[8] can be used to document the provenance of both lexical entries and concepts. It describes two annotations having as source the same book of Bolzano. One has as body the lexical entry *Vorstellung* and the other has as body the lexical concept the entry evokes. They have selectors that describe the exact text referring to the bodies of the annotation (respectively the lexical entry and its definition) and indicating their location in the whole text by assigning a prefix and suffix.

5 Relation to other modules and models

In this section we discuss the possible relations of our proposal with two Ontolex modules, in particular ***lime*** and ***vartrans***, and with ***Prov-O***. We will further investigate the positive or negative consequences before incorporating them into the proposed extension.

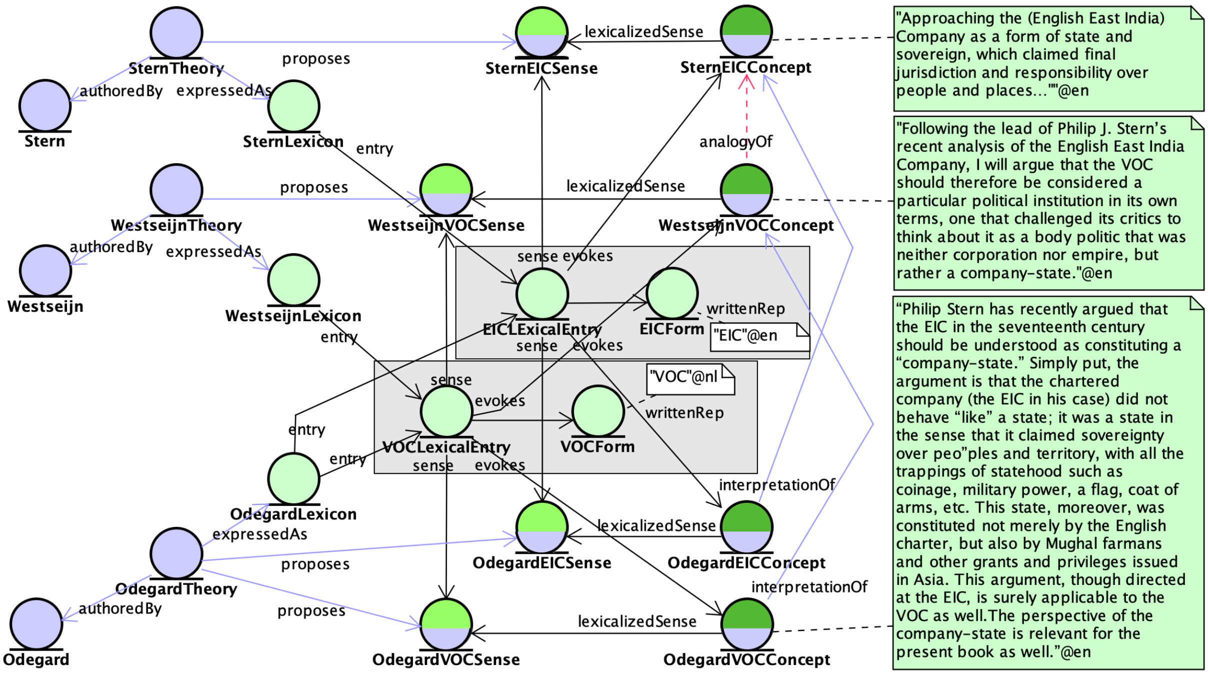
The ***lime*** (The LInguistic MEtadata)[9] module defines a `lime:ConceptualizationSet` as associating a `ontolex:ConceptSet` with a `lime:Lexicon`. One could consider a *View/Conceptualization* as a specialization of `ontolex:ConceptSet`, although the latter is clearly more than a just set of concepts. It may be a related to a `lime:ConceptualizationSet` since it is meant to bind the lexical concepts in the concept set and entries in the lexicon.

The relation to the ***vartrans***[10] module, which defines lexico-semantic relations such as `vartrans:Translation` between `ontolex:Senses` brings a more complex issue with respect to how to define the *Interpretation* and what is its relation between `vartrans:Translation` and *Interpretation*. We see a few possibilities: (i) a `vartrans:Translation` is a specialization of *Interpretation*, which would mean, among other things, that they would have to hold both between `ontolex:LexicalSenses` or between `ontolex:LexicalConcepts`; (ii) a variation of the

[8] https://www.w3.org/TR/annotation-vocab/

[9] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Metadata_(lime)

[10] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Variation_&_Translation_(vartrans)

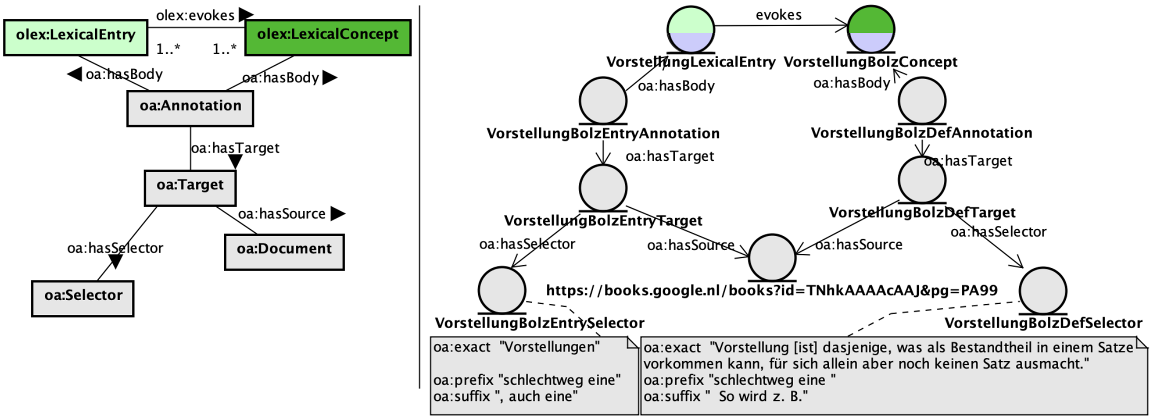


"Approaching the (English East India) Company as a form of state and sovereign, which claimed final jurisdiction and responsibility over people and places..."'@en

"Following the lead of Philip J. Stern's recent analysis of the English East India Company, I will argue that the VOC should therefore be considered a particular political institution in its own terms, one that challenged its critics to think about it as a body politic that was neither corporation nor empire, but rather a company-state."@en

"Philip Stern has recently argued that the EIC in the seventeenth century should be understood as constituting a "company-state." Simply put, the argument is that the chartered company (the EIC in his case) did not behave "like" a state; it was a state in the sense that it claimed sovereignty over peo"ples and territory, with all the trappings of statehood such as coinage, military power, a flag, coat of arms, etc. This state, moreover, was constituted not merely by the English charter, but also by Mughal farmans and other grants and privileges issued in Asia. This argument, though directed at the EIC, is surely applicable to the VOC as well.The perspective of the company-state is relevant for the present book as well."@en

Figure 4: Schema representing an instantiation of the model for EIC and VOC concepts. Each of them have two senses by different historians, namely Stern, Weststeijn and Odegard (the color code refers to the respective classes in Figure 2).



oa:exact "Vorstellungen"

oa:prefix "schlechtweg eine"
oa:suffix ", auch eine"

oa:exact "Vorstellung [ist] dasjenige, was als Bestandtheil in einem Satze vorkommen kann, für sich allein aber noch keinen Satz ausmacht."
oa:prefix "schlechtweg eine "
oa:suffix " So wird z. B."

Figure 5: Provenance annotation for the lexical entry *Vorstellung* alongside the concept it evokes (the color code refers to the respective classes in Figure 2 while Open Annotation ones are prefixed with *oa* and depicted in gray).

previous one is that they could be overlapping classes, meaning not all interpretations are translations and not all translations are interpretations (in the sense that they are not questionable); (iii) more aligned with our current proposal is that *Interpretation* hold between ontolex:LexicalConcepts, vartrans:Translation hold between ontolex:LexicalSenses, and the former could be derived from the latter; and (iv) it could also be that *Intepretation* hold between ontolex:LexicalSenses AND between ontolex:LexicalConcepts.

The ***Prov-O***[11] aims to support the representation of provenance information, either by being directly used or by serving as a reference model for creating domain specific provenance information. Its main entities are prov:Entity, prov:Activity and prov:Agent among which several provenance relations hold, for example, prov:wasDerivedFrom indicates that an entity is changed or created based on another, while and prov:wasAttributedTo ascribes an entity to an agent. Since our domain does require more specific provenance, such as the inter-

[11] www.w3.org/TR/prov-o/

Table 1: Competency Questions and Support by Models

Competency Question	OntoLex Base	OntoLex+ Others	Proposed Extension
CQ1.1 What are all the definitions of a given lexical entry?	✓	n.a.	n.a.
CQ1.2 What are all the definitions of a given lexical entry, along with their authors?	✗	~	✓
CQ1.3 What are all the definitions of a given lexical entry, along with their direct or indirect authors?	✗	~	✓
CQ2.1 Which concepts have been defined by a particular author?	✗	~	✓
CQ2.2 Which concepts defined by a particular author have also the term coined by him?	✗	~	✓
CQ2.3 Which concepts have been interpreted by a particular author?	✗	~	✓
CQ3 What are all the interpretations proposed by a specific author?	✗	~	✓
CQ4 What are the various interpretations that have been proposed for a specific view/conceptualization/theory?	✗	~	✓
CQ5.1 What is the (formal) definition of a concept given a (formal) syntax?	✓	n.a.	n.a.
CQ5.2 What is(are) the formal definition(s) of a concept?	✗	✗	✓
CQ5.3 What are the formal definitions of all concepts in a given theory?	✗	✗	✓

pretation of a concept as another one, or an analogy among them, we consider that Prov-O should not be used as is, but it can be a reference model from which our proposed extension can specialize.

6 Addressing the Competency questions

In this section, we discuss whether the competency questions and some variations can be addressed by the OntoLex-Lemon Base model, by combining it with other modules or vocabularies, or by the proposed extension. Table 1 indicates if the questions are fully, partially or not addressed using, respectively, the symbols ✓, ~, ✗. Moreover, we use n.a. when OntoLex-Base address the issue and therefore no extension is necessary.

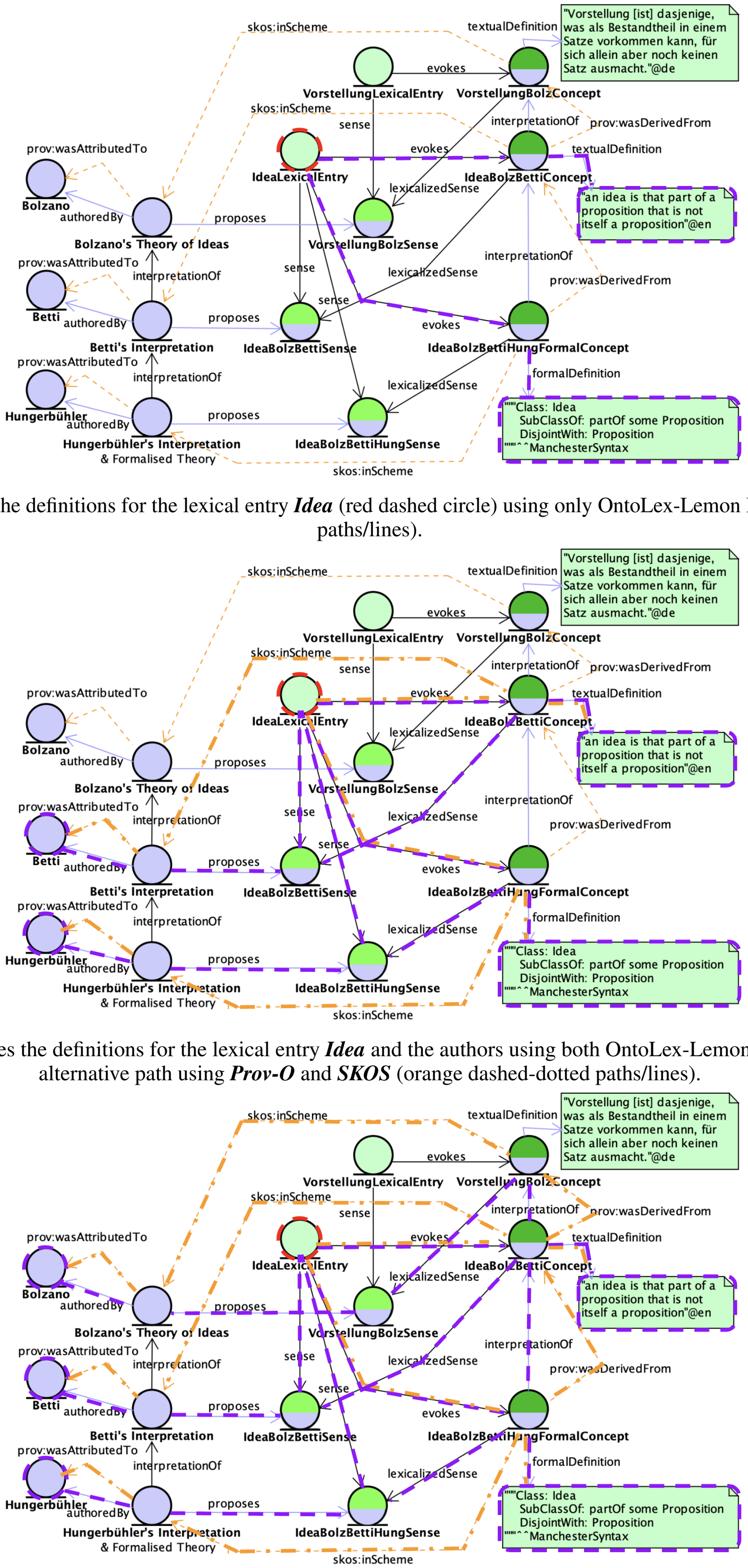
It turns out that a combination with the vocabularies ***Prov-O*** and ***SKOS***[12] can partially simulate the semantics intended in the proposed extension if a *View/Conceptualization* is taken as a `ontolex:ConceptSet`, which is connected to `ontolex:LexicalConcept` via `skos:inScheme`, also if `prov:wasDerivedFrom` connects a `ontolex:ConceptSet` to another one as *interpretation-of*, and if `prov:wasAttributedTo` indicates both the authorship of a `ontolex:ConceptSet` by `prov:Agent` and the coining of `ontolex:LexicalEntry` by a `prov:Agent`. However, the meaning may not be as clear, and therefore we consider the competency question to be partially addressed.

Figure 6 is a variation of the instantiation in Figure 3 including, in orange dashed lines, some of the aforementioned properties as alternatives to the proposed extension. It highlights the paths that could provide answers to the complementary questions CQ1.1, CQ1.2 and C1.3. The dashed purple paths illustrate the paths using the extension, while the dashed-dotted orange paths illustrate the alternative paths. It illustrates that, although similar results can be obtained with existing vocabularies, the proposed extension offers greater domain specificity, making it more suitable for guiding consistent and semantically accurate use.

7 Conclusion and Future Work

We propose a conceptual extension to OntoLex-Lemon with the purpose of representing the provenance of senses with evidence. It allows for ex-

[12] https://www.w3.org/2009/08/skos-reference/skos.html


skos:inScheme
textualDefinition
"Vorstellung [ist] dasjenige, was als Bestandtheil in einem Satze vorkommen kann, für sich allein aber noch keinen Satz ausmacht."@de
evokes
VorstellungLexicalEntry
VorstellungBolzConcept
sense
interpretationOf
prov:wasDerivedFrom
prov:wasAttributedTo
IdeaLexicalEntry
IdeaBolzBettiConcept
Bolzano
authoredBy
lexicalizedSense
"an idea is that part of a proposition that is not itself a proposition"@en
proposes
Bolzano's Theory of Ideas
VorstellungBolzSense
Betti
Betti's Interpretation
IdeaBolzBettiSense
IdeaBolzBettiHungFormalConcept
formalDefinition
Hungerbühler
Hungerbühler's Interpretation & Formalised Theory
IdeaBolzBettiHungSense
"""Class: Idea SubClassOf: partOf some Proposition DisjointWith: Proposition """^^ManchesterSyntax

(a) CQ1.1 Achieves the definitions for the lexical entry ***Idea*** (red dashed circle) using only OntoLex-Lemon Base (purple dashed paths/lines).

(b) CQ1.2 Achieves the definitions for the lexical entry ***Idea*** and the authors using both OntoLex-Lemon extended and an alternative path using ***Prov-O*** and ***SKOS*** (orange dashed-dotted paths/lines).

(c) CQ1.3 Achieves the definitions for the lexical entry ***Idea*** and the direct and indirect authors using both OntoLex-Lemon extended and an alternative path using ***Prov-O*** and ***SKOS*** .

Figure 6: Visualization highlighting paths providing answers to the complementary questions CQ1.1, 1.2 and 1.3.

pressing conceptual setups as well as their interpretations, as well as expressing the textual or formal definitions of the concepts, accompanied by annotations leading to the excerpt of original text where the definition is provided. The current proposal addresses Competency Questions 1 to 5. Competency Questions 6 and 7 are challenging regarding evolution of concepts and comparison among them and will be addressed in future work, as well as new competency quesitons.

As our aim is to outline, at a conceptual level, how the OntoLex-Lemon model could be extended to address the proposed competency questions, the implementation is still to be investigated. For that we would consider the reuse of existing vocabularies, such as Prov-O, SKOS and DC-Terms[13], as well as structured representations like nanopublications[14]. It is also important to further investigate the connections to other modules and extensions of OntoLex. Next we will conduct a practical evaluation of our proposal by applying it not only to extended versions of our case studies but to related cases from the literature.

We furthermore plan to investigate the alignment of our proposal with existing models to address upcoming challenges. One promising direction is the integration with the Linguistic Annotation Scheme GRaSP (van Son et al., 2016), a framework that adopts a multilayered approach in four layers, namely events, attribution, factuality, and opinion. We aim to explore how our proposed model for representing definitions and interpretations can be integrated with GRaSP. In particular, the opinion layer offers a promising space to explore differing theoretical perspectives, conceptual interpretations, and scholarly disagreements.

Another relevant direction is the alignment with the LMM (Linguistic Meta-Model) (Picca et al., 2008), meant for representing heterogeneous lexical knowledge, providing a semiotic-cognitive representation of linguistic knowledge grounded in DOLCE foundational ontology (Gangemi et al., 2002). In particular, it considers different ways of assigning meaning to an expression, expliciting the ontological nature of the "meaning definitions" and the relations between them, which can be relevant for understanding how to relate and compare the definitions of lexical concepts.

In terms of modelling concept evolution, we also want to explore geographical and temporal dimensions because the meanings and the interpretations of concepts can vary across time and geography. Geographical factors can influence how a concept is understood and used in a specific place. For example, the VOC might be perceived differently in the former Dutch colonies from a postcolonial perspective than in the Netherlands from a perspective of Dutch national history.[15] Temporal dimensions, including historical periods, cultural-societal shifts, and technological advancements, also reveal how concepts evolve over time and how their interpretations change. For example, the concept of "privacy" has undergone significant transformation in the digital age, evolving from Warren and Brandeis's 1890 conception of "the right to be let alone" to contemporary debates between individual autonomy-based approaches versus social relational frameworks that "surpass the perspective of the individual" (Becker, 2019). Survey data demonstrates measurable temporal shifts in privacy attitudes, with older adults more concerned about their online security and privacy compared to the younger generations, reflecting broader cultural-societal shifts in how privacy is conceptualized in digital contexts (Holmes, 2022). Highlighting the geographical and temporal contexts from which a concept or interpretation emerges is likely to promote historiographical practices, and representing geographical and temporal information along with lexical information can contribute to this advancement. The documentation of our extension can be found at `https://github.com/trifecta-project/lexical-sense-definition`.

Acknowledgments

This work was funded by the European Union under grant agreement No. 101088548 (TRIFECTA), and by the Dutch Research Council (NWO) through the projects Culturally Aware AI (KIVI.2019.005), e-Ideas (VICI grant 277-20-007), and Counting the Bees (Open Competition grant 406.22.FHR.031). Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We acknowledge our

[13] `www.dublincore.org/specifications/dublin-core/dcmi-terms/`

[14] `https://nanopub.net/`

[15] `https://internationaleonline.org/contributions/the-dutch-voc-mentality-cultural-policy-as-a-business-model/`

colleagues Manjusha Kuruppath for the input information to the VOC case study and Teresa Paccosi for valuable discussions about interpretations and translations, and the reviewers for their insightful comments and constructive criticism.

Author Contributions

Author contributions (by author initials) are listed according to the Contributor Roles Taxonomy (CRediT). Conceptualization: VZ, AB, JZ, MvE; Formal Analysis: VZ; Methodology: VZ, AB, JZ, MvE; Project administration: AB, MvE; Visualization: VZ, MvE; Writing (original draft): VZ, JZ; Writing (review and editing): VZ, JZ, MvE, AB

References

Melanie Becker. 2019. Privacy in the digital age: comparing and contrasting individual versus social approaches towards privacy. *Ethics and Information Technology*, 21(4):307–317.

A. Betti. 2012. *Bolzano's Universe: Truth, Logic and Metaphysics*, pages 167–190. Oxford Univ. press.

Leonard Blussé. 1986. *Strange Company: Chinese Settlers, Mestizo Women and the Dutch in VOC Batavia*. KITLV Press, Leiden.

Bernard Bolzano. 1837. *Dr. B. Bolzanos Wissenschaftslehre: Versuch einer ausführlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter. Zweiter Band*, volume 2. in der JE v. Seidelschen Buchhandlung.

Pieter C. Emmer. 2003. *The Dutch in the Atlantic Economy, 1580–1880: Trade, Slavery and Emancipation*. Ashgate, Aldershot.

Femme Gaastra. 2003. *The Dutch East India Company: Expansion and Decline*. Walburg Pers, Zutphen.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. 2002. Sweetening ontologies with dolce. In *International conference on knowledge engineering and knowledge management*, pages 166–181. Springer.

Emiliano Giovannetti, Davide Albanesi, Andrea Bellandi, Simone Marchi, Mafalda Papini, and Flavia Sciolette. 2024. Maia: an open collaborative platform for text annotation, e-lexicography, and lexical linking. *Umanistica Digitale*, 8(18):27–52.

Dagmar Gromann, Elena S. Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian O. Truica, Andrius Utka, and Giedre V. Oleskeviciene. 2024. Multilinguality and llod: A survey across linguistic description levels. *Semantic Web*, 1.

Lisa Holmes. 2022. Generational attitudes and actions around data privacy. *Euromonitor International*. Data from Euromonitor International's annual Voice of the Consumer: Digital Survey.

Silvan Hungerbühler. 2018. A computational method for philosophical interpretation. Master's thesis, University of Amsterdam.

Jonathan Israel. 1989. *Dutch Primacy in World Trade, 1585–1740*. Clarendon Press, Oxford.

Anas F. Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P. Mccrae, Émilie Pagé-Perron, Marco Passarotti, Salvador R. Muñoz, and Ciprian O. Truica. 2022. When linguistics meets web technologies. recent advances in modelling linguistic linked data. *Semantic Web*, 13:987–1050.

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*.

Erik Odegard. 2020. *The Company Fortress: Military Engineering and the Dutch East India Company in South Asia, 1638-1795*. Leiden University Press.

Yvette Oortwijn, Hein van den Berg, and Arianna Betti. 2021. Ground truths for the humanities. *Preprint*, arXiv:2103.12841.

Davide Picca, Alfio Gliozzo, and Aldo Gangemi. 2008. Lmm: an owl-dl metamodel to represent heterogeneous lexical knowledge. In *Proc. of the Int. Conference on Language Resources and Evaluation*.

Alicia Schrikker. 2007. *Dutch and British Colonial Intervention in Sri Lanka, 1780–1815: Expansion and Reform*, volume 3 of *EMI (European Expansion and Indigenous Response)*. Brill, Leiden.

Philip Stern. 2011. *The Company-State: Corporate Sovereignty and the Early Modern Foundations of the British Empire in India*. Oxford University Press.

Sabine Tittel. 2023. Ceci n'est pas un dictionnaire. adding and extending lexicographical data of medieval romance languages to and through a multilingual lexico-ontological project. *Electronic lexicography in the 21st century*, pages 39–52.

Marieke van Erp. 2023. Unflattening knowledge graphs. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, page 223–224, New York, NY, USA. Association for Computing Machinery.

Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. 2016. Grasp: A multilayered annotation scheme for perspectives. In *Tenth Int. Conference on Language Resources and Evaluation*, pages 1177–1184.

Arthur Weststeijn. 2014. The voc as a company-state: Debating seventeenth-century dutch colonial expansion. *Itinerario*, 38(1):13–34.

Author Index