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

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

Abstract

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

2. Lemma Banks

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

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

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

3. Related Work

3.1. LiLa Lemma Bank

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

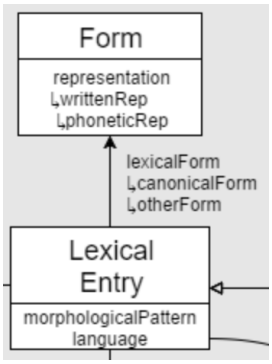

Figure 1: A section of the OntoLex-Lemon core model, specifically the relationship between the classes `ontolex:LexicalEntry` and `ontolex:Form`.

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

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

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

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

3.2. Large Language Models for Knowledge Graph Construction

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

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

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

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

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

4. Methodology

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

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

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

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

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

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

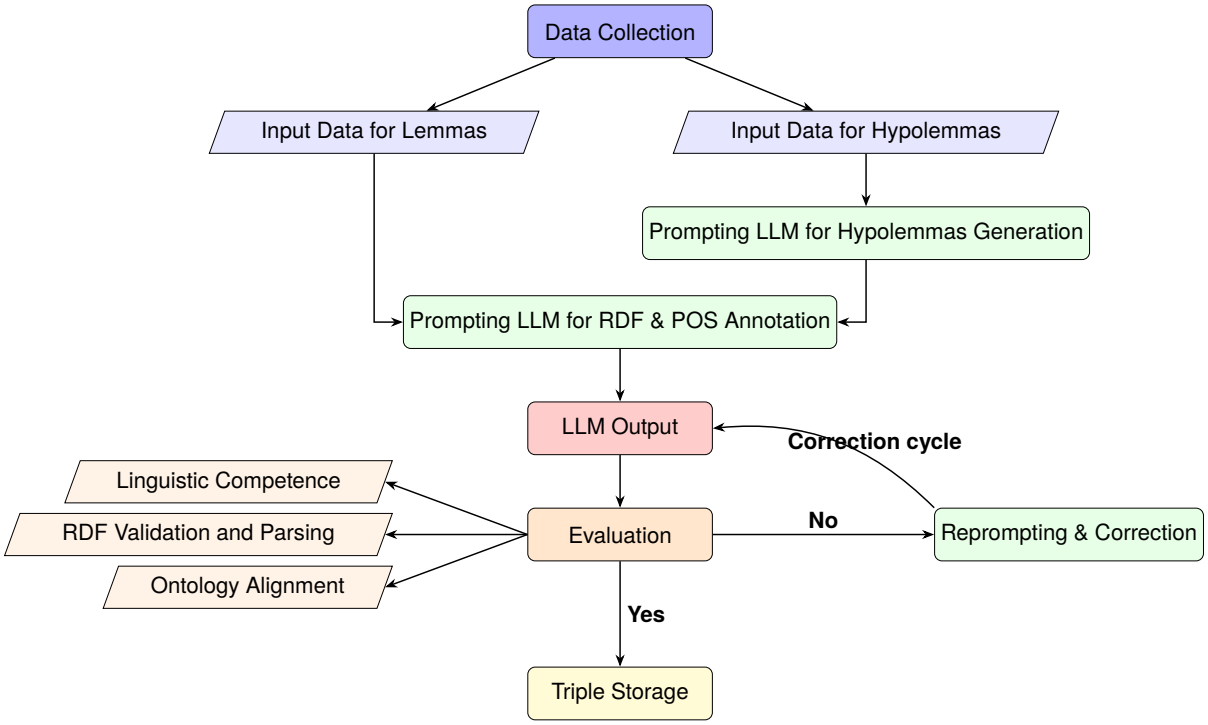

Figure 2: Workflow diagram of the proposed neuro-symbolic pipeline towards the construction of LinkEn.

4.1. Lexical Data

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

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

book -> booked,booking,books

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

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

4.2. Ontological and modeling choices

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

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

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

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

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

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

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

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

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

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

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

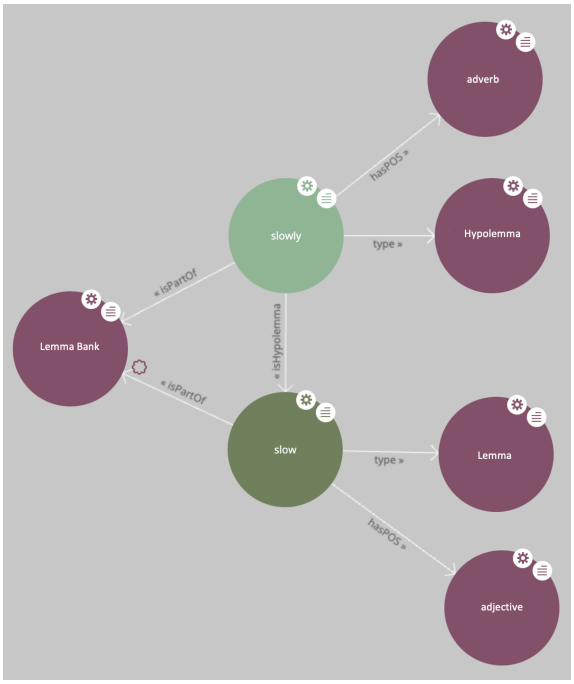

Figure 3: LodLive visualization of the deadjectival adverb *slowly*,[19] hypolemma of the base adjective *slow*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

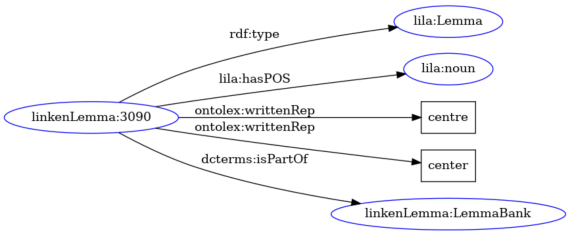

Figure 4: Visualization of the lemma *centre* with the `ontolex:writtenRep` data property pointing to two different strings.

4.3. Prompting

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

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

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

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

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

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

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

5. Evaluation and Results

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

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

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

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

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

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

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

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

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

Apart from the one related to PoS, another er-

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

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

6. The LinkEn Lemma Bank

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

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

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

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

7. Future Work

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

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

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

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

8. Bibliographical References

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

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

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

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

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

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

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

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

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

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

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

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

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

Marco Carlo Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Maria Gabriella Litta Modignani Picozzi, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

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

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

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

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

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

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

9. Language Resource References

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

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

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