LDK 2023

Lexico-Semantic Mapping of a Historical Dictionary: An Automated Approach with DBpedia

Sabine Tittel
Heidelberg Academy of Sciences and Humanities
Karlstraße 3, D – 69117 Heidelberg, Germany
sabine.tittel@hadw-bw.de

Abstract

Modeling lexical resources following the Linked Data paradigm has become a widespread method to contribute to the multilingual web of data. For the modeling of linguistic information such as words and their morphosyntactic aspects, standard vocabularies offer elaborate means to enable cross-resource and cross-domain access to the resources. To establish access to the word senses, it is pivotal to create a mapping of each word sense and its underlying concept to an external, language-independent knowledge base of the Semantic Web such as DBpedia. However, this lexico-semantic mapping is a very time-consuming endeavor and is often neglected. And yet, the problem of how to install time-saving approaches is not resolved. Therefore, we propose a solution for an automated lexico-semantic mapping based on Old French lexicographic data. The quantitative and qualitative evaluations of the outcome show very promising results. Overall, approx. 71% of the word senses can be mapped to a DBpedia entry: approx. 12.7% of semantically accurate mappings and approx. 58.2% of approximate, yet semantically meaningful mappings. These results can be fully extrapolated to our linguistic resource and also transferred to the Linked Data modeling of related resources.

1 Introduction

The last decade has seen many successful attempts to model lexical resources as Linked Open Data (Bizer et al., 2009). RDF (*Resource Description Framework*, Klyne et al. (2004)) is used as the standard format along with W3C-standard vocabularies and ontologies as a means to create a web of interlinked data. Attempts focus on the modeling of words and parts of speech, their graphical realizations, morphological and syntactic aspects, translations into other languages, their role in multi-word expressions, etc. (for an overview of technologies, vocabularies, and methods, see Bosque-Gil et al. (2018), Khan et al. (2022)). The vocabulary most often used for modeling lexical resources is OntoLex-Lemon, Cimiano et al. (2016). While the linguistic structures of the lexical resources can be seamlessly converted to RDF, a challenging aspect of the modeling process is to integrate links from the *senses* of the words (lexemes) and their underlying *concepts*, respectively, to an external knowledge base. We call this the lexico-semantic mapping (in the following, LexSemMapping). The LexSemMapping is pivotal for establishing lexical-semantics-based access to the lexical units (that is, the nexus of a given lexeme and precisely one (of its) senses): Only lexical-semantics-based access makes the lexical units of, for example, a historical dictionary, available for cross-domain and cross-resource access that is, most importantly, independent from the language and language stage of the resource.

For the LexSemMapping, an extra-linguistic resource depicting the things of the world such as Wikidata and DBpedia[1] can serve as an external knowledge base. An illustration of the motivation for a LexSemMapping is as follows: Lexical resources contain numerous designations for, say, clergymen: Old High German *priest* m., *priestar* m., *prêstar* m., Middle High German *priestære* m., and High German *Priester* m. (since 9thc, Grimm[2] 13,2115[2] and DWDS PRIESTER[3]), Old High German *gotmanno* m., High German *Gottesmann* (since ca. 870, Grimm[2] 8,1285; DWDS

[1] https://www.wikidata.org/, https://www.dbpedia.org/; these and all following URLs are accessed on 02-21-2023].

[2] *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*, digital version, https://woerterbuchnetz.de/?sigle=DWB#Priester.

[3] *Digitales Wörterbuch der deutschen Sprache*, https://www.dwds.de/wb/Priester; we note that the DWDS offers a Thesaurus function leading to semantic cognates; however, this is limited to the German lexemes registered within the DWDS.

GOTTESMANN), Old French *pestre* m., *prastre* m., *prebstre* m., *preiste* m. (since the beginning of the 12thc, DEAF*él* PRESTRE[4]), *flame* m., *flamine* m., *archiflame* m. (since 13th / 14thc., DEAF*él* FLAME[25], Italian *flamine* m. (since 1261-1292, TLIO FLÀMINE[6], Old Occitan *flamina* m. (DOM FLAMINA[7]), and many more. The senses of these lexemes represent concepts that are connected to different religions, cultures, times, and connotations. Their investigation is promising not only from a linguistic point of view but also as a linguistic underpinning for studies on expressions of religion through time and space (cp. the article PRIESTER in Bautier et al., 1977-1998, 7,203-208; Richard, 1959; Salisbury, 2015). Creating a connection, for example, from all senses with the concept ⌜Priests⌝ to the DBpedia entry 'Priest', or from all clergymen of all religions to a generic entry 'List_of_religious_titles_and_styles'[8] could establish access through the means of the Semantic Web to all of the lexemes listed above. These are otherwise very difficult to find.

Indeed, OntoLex-Lemon offers classes to model sense definitions (`LexicalSense`) and concepts (`LexicalConcept`[9]) and the predicates (`reference` and `isConceptOf`, respectively[10]) to link these classes to an external knowledge base. Its entities then serve as the objects of the RDF triples for the LexSemMapping.

However, the LexSemMapping, to the best of our knowledge, has rarely performed on a larger scale. We suspect that this is (partly) because such a mapping is a very tedious and time-consuming endeavor. The problem thus arises as to how a LexSemMapping of lexical units can be established in a quicker and more efficient way. In this paper, we propose a solution for this problem by developing methods for an automatic mapping of lexical units to DBpedia.

[4] https://deaf.ub.uni-heidelberg.de/lemme/prestre.

[5] https://deaf.ub.uni-heidelberg.de/lemme/flame2. Hereafter, all Old French lexemes refer to DEAF*él*.

[6] *Tesoro della Lingua Italiana delle Origini*, http://tlio.ovi.cnr.it/voci/025560.htm.

[7] *Dictionnaire de l'occitan médiéval*, http://www.dom-en-ligne.de/.

[8] https://dbpedia.org/page/Priest, https://dbpedia.org/page/List_of_religious_titles_and_styles.

[9] In accordance with the semiotic pentagon, see, e.g., Blank (2001, 9).

[10] https://www.w3.org/2016/05/ontolex/.

The remainder of the paper is divided into an overview of related work (Section 2), a description of the lexical resource that is our use case (Section 3), an assessment of manual LexSemMapping (Section 4), and the development and evaluation of automatic approaches (Section 5). We conclude our paper by presenting the overall result and an outlook (Section 6).

2 Related Work

Establishing data access based on lexical semantics is important for lexical resources, in particular for historical language stages whose lexical units are harder to access than those of modern languages; and yet, the process of LexSemMapping is rarely described in the literature.

Herold et al. (2012) describe the attempt to do this for the data of the *Digitales Wörterbuch der Deutschen Sprache – DWDS-Wörterbuch* (DWDSWB)[11]: Through an alignment of this dictionary with the entries of the *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*, Volumes I–XVI, Leipzig 1854–1960 ([1]DWB), a semantic disambiguation shall be achieved. This corresponds to a LexSemMapping, even if the target is not expressed as an RDF triple object. But the challenges due to homonyms, polysemy, and semantic shift led Herold et al. (2012, 42) to conclude that, «Given the huge amount of manual effort needed to complete the alignment between DWDSWB and [1]DWB on the level of lexical entries it seems unfeasible to achieve a mapping for individual senses».

Bozzi (2016) detail their failed attempt to use WordNet for a lexical-semantic networking of data of the *Dictionary of Old Occitan medico-botanical terminology* (DiTMAO). DiTMAO utilizes OntoLex-Lemon as a means to perform a LexSemMapping of the modeled lexemes through external ontologies: «In the next step, the DiTMAO partners will formalize the conceptual domain, describing the fields of botany, zoology, mineralogy, human anatomy, diseases and therapies (medication, medical instruments) [...] to ease the "onomasiological" access to the lexicon», Bellandi et al. (2018, 10-11). However, they do not further elaborate on how to establish a LexSemMapping.

Declerck et al. (2015, 348-350), in sample data of the *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ[12]), link the lexeme Ger-

[11] https://www.dwds.de/d/wb-dwdswb.

[12] https://wboe.oeaw.ac.at/.

man *Trupp* (a squad) to the DBpedia entry 'Social_Group'. They point out the importance of integrating the data into larger semantic contexts, as well as linking to other external resources that also connect to the DBpedia entry given in the example. How this linkage with DBpedia is to be performed, however, remains unresolved: «An issue we would like to consider is the possibility of automatically linking to external resources, those being both of linguistic nature or encyclopedic nature. We do not have an answer to this point for the time being. As a heuristic, while knowing that the Limburg lexical data concerns anatomy, and the reference language is standard Dutch, we can automatically query DBpedia for all entries that have a Dutch word marked with the additional "_(anatomy)" extension, such as for example http://nl.dbpedia.org/page/Hoofd_(anatomie). However, this might only offer a very specific solution», (Declerck et al., 2015, 353).

Cimiano et al. (2013) evaluate possibilities to model the *semantics by reference* implied by OntoLex-Lemon in a more fine-grained method than the connection of `LexicalSense` to an ontology allows, bringing back semantic disambiguation at least partially into the model. Their code samples (Cimiano et al., 2013, 58f.) show DBpedia, among others, as an external knowledge base, but the process of semantic disambiguation itself is not discussed.

Giuliani and Molina Sangüesa (2020) describe the integration of two large historical lexical resources, i.e., the *Tesoro della lingua italiana delle origini* (TLIO[13]) and the *Nuevo Diccionario Histórico del Español* (NDHE, *Real Academia Española*[14]), with the taxonomy of the *Historical Thesaurus of English* (HTE)[15]. Focusing on the domain 'health and illness', they translate HTE's entities into Spanish, extend them to a more fine-grained level, and integrate them into their work infrastructure as an onomasiological backbone. The taxonomy is also converted into an ontology in OWL (Bechhofer et al., 2004) called DHistOntology and the modeling of the two resources in RDF is described as a future goal (Molina Sangüesa, 2023). Their aim is to enhance their workflow by aligning similar concepts in both resources and to streamline sense definitions while editing the dictionary articles with one shared dictionary writing system. This is a promising concept, albeit the lexico-semantic mapping seems to be performed manually.

The historical dictionary *Lessico Etimologico Italiano* (LEI, Pfister 1979–) also examines using the classes of the HTE as a means to establish onomasiological access. The goal is not an integration of the LEI resource into the Linked Data landscape but the creation of a locally used, proprietary feature for the online publication LEI-digitale.[16] As a first step, their approach focuses on the LexSemMapping of the Latin etyma – that serve as the headwords of the LEI articles – and their definitions. The second step is to integrate the lexical units of the articles, i.e., the Italian lexemes and their definitions. The heterogeneity of the latter is significant, including single-word definitions in modern Italian and also Latin, a sequence of modern Italian translations (i.e., of several senses in one definition text), periphrastic definitions, nomenclature adopting the classification by Carl von Linné (we will further discuss Linné in Section 5.1), and more. The mapping is done manually: Concepts are looked up in Wikipedia, and corresponding entities are identified in and linked to the HTE taxonomy. The link is manually integrated into the XML files of the articles.[17] Since the LEI is a very large resource with a great amount of legacy data (and also born-digital data), it seems crucial for the success of their LexSemMapping to integrate automated steps into the process. However, no solution for time-saving automation has been promoted so far.

3 The Linguistic Resource

The motivation for our approach to establishing a more efficient method for LexSemMapping derives from modeling the data of the *Dictionnaire étymologique de l'ancien français* – DEAF (Baldinger, 1971-2020) as Linked Open Data. The DEAF is a comprehensive dictionary of Old French from its first resource 842 AD until ca. 1350 AD, compiled under the aegis of the Heidelberg Academy of Sciences and Humanities until 2020.[18] We have invested in modeling the DEAF articles as Linked Open Data for two reasons: firstly, to make the data of the DEAF accessible beyond the nuanced

[13] http://tlio.ovi.cnr.it/TLIO/.
[14] https://www.rae.es/.
[15] http://historicalthesaurus.arts.gla.ac.uk/.
[16] https://lei-digitale.it/.
[17] Personal communication by Alessandro A. Nannini, LEI, to whom we express our sincere thanks.
[18] https://www.hadw-bw.de/deaf.

yet predefined, and thus limited research functions of its online publication, DEAF*él*[19]; and secondly, to facilitate the usability, queriability, and interpretability of the DEAF data in the global context of the Semantic Web. We describe the vocabularies, e.g., OntoLex-Lemon and OLiA (Chiarcos and Sukhareva, 2015), the concept, outcome, and challenges of the modeling process in Tittel and Chiarcos (2018) and – with further elaboration – in Tittel (forthcoming). In Tittel and Chiarcos (2018), we proposed implementing a semi-automatic process to increase efficiency. In this process, XSLT scripts would model the DEAF data as RDF by integrating the predicate `ontolex:isConceptOf` and a wildcard in place of a link to an extra-linguistic ontology as the object of the RDF triple. This would help prepare for manual mapping. It, of course, does not produce a meaningful statement, and the necessary manual post-processing could not be performed due to the termination of the funding period of the DEAF. However, the RDF data offer a starting point; for example, for Old French *raicele* s.f. "plante vivace de la famille des Violaceae, aux feuilles en rosette et aux fleurs blanches légèrement ou pas parfumées, violette blanche", the concept ⌜White Violet⌝ can now be mapped to the entity of DBpedia 'Viola_alba'[20] in the following way (RDF serialized in Turtle):[21]

```
deaf:raicele_lexConcept
  ontolex:isConceptOf dbr:Viola_alba .
```

4 Manual LexSemMapping

A manual LexSemMapping for the DEAF data promises the best results. This is particularly true with respect to the *Historical Semantic Gap* (Tittel and Chiarcos (2018), Giuliani and Molina Sangüesa (2020, 355f.)) that often occurs between a concept represented by a lexeme in a historical (in this case, medieval) language stage and the concept of the same lexeme in the modern language. E.g., medieval concepts of the bloodstream adhere to a metabolism that does not know blood circulation (described only in 1628 by William Harvey, Schipperges (1990, 53)). Therefore, Old French *veine* f., for example, does not denote the blood vessel transporting the blood back to the heart (as part of blood circulation). Instead, *veine* denotes a blood vessel transporting the nourishing blood from the liver to all body parts and then back to the liver. Hence, the concept cannot be mapped to the modern concept of the ⌜vein⌝, as in DBpedia's entry 'Vein'[22] without causing semantic disruption and anachronistic cross-fade.[23] On the other hand, the LexSemMapping is straightforward when the concept to be mapped has the exact same scope and application today as it did in medieval times. This is often the case for plant and animal names, musical instruments, tools, etc., and DBpedia is very well suited for this purpose.

For writing each dictionary article, the lexicographer penetrates the semantic scope of the analyzed lexeme and grasps the concept of each lexical unit in a way that makes possible a seamless integration of an ontology entity into the data. Furthermore, they might analyze several lexemes belonging to a domain at a certain point in time and, in doing so, remain focused on that particular topic. E.g., after editing lexemes occurring in the context of the *veine* (see above), they have internalized medieval metabolic concepts and pneuma theory (Putscher, 1974) to the point of becoming, to a certain extent, an expert which further facilitates the mapping process. We, therefore, argue that a manual LexSemMapping is feasible when done while editing a dictionary article.

The case of legacy data, as is the case for the DEAF dictionary, is different, however. DEAF*él* contains approximately 84,000 lexemes with 92,776 lexical units[24] that must be linked, in hindsight, to an extra-linguistic knowledge base. The dictionary covers all aspects of the language, and hence, a LexSemMapping requires knowledge in all domains of life. For a retrospective mapping of legacy data, this is difficult: While the knowledge of the lexicographer is greatest at the time of the article editing, the person performing the mapping in retrospect must promptly acquire expertise for many domains ad hoc. This is also immensely time-consuming. Estimating 10 min per LexSemMapping adds up to 15.462 hours of work, roughly 200

[19] https://deaf.ub.uni-heidelberg.de/.

[20] https://dbpedia.org/page/Viola_alba.

[21] Namespaces, such as `deaf`, `ontolex`, and `dbr` (DBpedia) in the following code examples are assumed to be defined the usual way.

[22] https://dbpedia.org/page/Vein.

[23] This observation leads to the demand for historicized ontologies that model the historical concepts of a domain of interest. This is not further discussed in this paper. We however indicate that the project *Knowledge Networks in Medieval Romance Speaking Europe* (ALMA, https://www.hadw-bw.de/alma) will develop domain ontologies for medieval medicine and law.

[24] Not counting the lexical units where the sense is marked by '?'.

working days, for the DEAF data — provided that the required entities of a knowledge base do exist.

5 Automatic Approaches to LexSemMapping

To address this problem, we have developed automatic methods involving applying Python scripts for a LexSemMapping of the DEAF data. As an encyclopedic resource, DBpedia only registers (concrete and abstract) things that are described in Wikipedia (from where DBpedia extracts its data[25]). Furthermore, DBpedia shows significant shortcomings with respect to historical concepts. Nonetheless, we focus on DBpedia as a target resource, acknowledging its broad range of entities and its pivotal role as a central node within the web of data.

At this point, we rule out linguistic resources such as WordNet, Open Multilingual Wordnet, and BabelNet[26] because our goal is to semantically map the concepts to an extra-linguistic resource enabling semantic access that is independent of a language representation. For the future expansion of the methodology, we will revisit this decision for the sake of larger interoperability.

5.1 Four Methods for Mapping Nouns

The 92,776 sense definitions of the DEAF are (i) partly defined by following the genus–differentia approach[27], (ii) partly by single French words, and (iii) partly by translations in Modern French, i.e., equivalents of the sense following the genus-differentia definition as the last word of the definition text. Aiming at a maximum of correct hits when linking the definitions to corresponding DBpedia entities, we define four methods for automatically mapping nouns: (i) We establish links using the terminology classified through the *Systema naturae* by Carl von Linné[28] (in the following: LINNÉTERMINUS); (ii) we transform single-word definitions (SINGLEWORD); (iii) we use the Modern French equivalents (LASTWORD); (iv) we extract the genus proximus of a sense definition (GENUSPROXIMUS).

5.1.1 LINNÉTERMINUS Approach

Many definitions include a Linné classification that is utilized in this approach. The standard syntax is: "<definition> (<Latin term> L.)", as in: *faverole* f. t. de botanique "petite plante dicotylédone, de la famille des Plantaginaceae..., véronique des ruisseaux (Veronica beccabunga L.)" (limewort). But we also find definitions (i) with a Latin term enclosed in distinctive parentheses, beginning with an uppercase letter but without the 'L.' marker, (ii) the opposite: with the 'L.' marker but without the parentheses, and (iii) with neither the 'L.' marker nor parentheses. All these cases considered, roughly 200 definitions can be mapped through the LINNÉTERMINUS approach. Although this might not seem a significant contribution to automated mapping, the expected correctness of the results suggests the development of an algorithm that reads Linné classifications.

5.1.2 SINGLEWORD Approach

This approach is straightforward. The algorithm uses the single Modern French word of the definition (filtering out occasional question marks), as in: *lechement* m. "flatterie" (flattery). A database query results in 21,166 such SINGLEWORD definitions. These definitions don't comply with the concept of genus–differentia definitions; they feature in DEAF*pré*, a section of DEAF*él*. DEAF*pré* contains the digitized material of the DEAF card index (with 1.5 million handwritten slips that amount to 12 million attestations of lexemes), structured into preliminary dictionary entries with a provisional semantic analysis.

5.1.3 LASTWORD Approach

A further approach is a method of reading the Modern French translation typically given as an equivalent of the sense at the end of the definition. This approach is based on the syntax: "<definition>, <Modern French word>", as in: *figuier* m. "arbre qui produit la figue, figuier", the fig tree. However, this approach has several drawbacks. The algorithm accurately reads a single word between the last comma and the closing quotation marks of the definition text (filtering out question marks). How-

[25] See https://www.dbpedia.org/resources/linked-data/.

[26] https://wordnet.princeton.edu/, https://omwn.org/, https://babelnet.org/.

[27] A genus–differentia definition is the state-of-the-art definition of a sense consisting of a generic term (*genus*, e.g., 'plant') and specifications of that term (*differentia*, e.g., 'perennial', 'with rosette-shaped leaves', 'with lightly scented white flowers', cp. the above mentioned White Violet.

[28] Editio princeps Leiden [Lugdunum Batavorum] (Theodor Haak) 1735.—The systems by Carl Gottlob Rafn (https://viaf.org/viaf/106965171/) and Georges Léopold Chrétien Frédéric Dagobert, Baron de Cuvier (https://viaf.org/viaf/4981028/), are alternatives; in the DEAF, however, we do not see them used in a sense definition.

ever, the hit ratio is influenced by many cases in which that particular single word is not a Modern French equivalent, but part of an enumeration that belongs to the periphrastic definition itself. An example is: *dachete* f. “sorte de petit clou à la tête particulièrement grande et à la tige angulaire, adapté aux besoins de cordonniers, tapissiers, etc.”. In this case, following the rules, the algorithm finds that *etc.* is the last word after the last comma; this can be filtered out. Consequently, *tapissiers* (tapestry weavers) is the word to be used by the algorithm for LexSemMapping. Sure enough, the tapestry weavers are only an example (together with *cordonniers*, shoemakers) for professional groups that use the *dachete* (a type of small nail). Nevertheless, this approach is highly relevant for automatic LexSemMapping due to its numerous occurrences.

5.1.4 GENUSPROXIMUS Approach

While the first three approaches aim at the LexSemMapping of the specific meaning of the word, this approach uses the genus proximus of the sense definition for an approximate mapping, i.e., of the meaning’s core. It relies on the periphrastic definitions in accordance with the syntax: “sorte de / sorte d’ / espèce de / espèce d’ <genus> <differentiae>”, e.g.: *tideman* m. “espèce de douanier qui attend la marée haute pour faire les bâteaux arrivant acquitter les impôts”. Although *tideman* denotes a very particular tollkeeper, the generic tollkeeper (*douanier*) is the concept that will be mapped by the GENUSPROXIMUS approach. Oftentimes, the genus proximus is preceded by an adjective, such as ‘small’ or ‘large’; this will be considered by the algorithm. A database query results in 3,870 such GENUSPROXIMUS definitions.

5.1.5 Proof of concept with manually created data sample and English Translations

The mapping process to DBpedia is based on the fact that for each Wikipedia entry, a DBpedia entry can be assumed: «For each Wikipedia page, DBpedia has an entity following the same pattern: http://en.wikipedia.org/wiki/Berlin → http://dbpedia.org/resource/Berlin», see https://www.dbpedia.org/resources/linked-data/ [accessed 02-17-2023]. To query Wikipedia’s data, e.g. for article entries, the Python script imports an API provided by Wikipedia (see ‘Wikipedia API’ at https://pypi.org/project/Wikipedia-API/).

To test feasibility, we conduct a Proof of concept (PoC): We implement a semi-automatic approach by manually preparing a data sample (`data_poc`). This sample consists of a list of lexemes, definitions, and keywords to be mapped for LINNÉTERMINUS, SINGLEWORD, LASTWORD, and GENUSPROXIMUS, each including 30 examples. The DEAF sense definitions are written in Modern French. Therefore, we provide English translations of the keywords to facilitate the detection of corresponding entries in the English Wikipedia for the algorithm. A list entry is structured as follows, with ‘lexeme’, ‘definition’, and ‘English keyword’, respectively:

```
['zecharr', 'espèce de faucon', 'falcon']
```

The pseudocode for our PoC reads as follows:

```
1  IMPORT wikipediaapi
2  SET wiki_wiki TO wikipediaapi.Wikipedia('en')
3
4  DEFINE FUNCTION concat(text):
5    RETURN str(text).replace(' ', '_')
6      .replace('œ', 'oe').replace('æ', 'ae')
7      .replace('?', '')
8
9  DEFINE FUNCTION map(data_poc):
10   SET entries_to_dbr TO data_poc
11   FOR row IN data_poc[1:]:
12     SET keyword TO concat(row[2])
13     SET page_py TO wiki_wiki.page(keyword)
14     IF page_py.exists():
15       SET url TO page_py.fullurl
16       SET url_db TO str(url).replace('https://
17         en.wikipedia.org/wiki/',
18         'https://dbpedia.org/resource/')
19       row.append(url_db)
20     ELSE:
21       SET keyword TO 'unknown_entry'
22       row.append(keyword)
23   RETURN entries_to_dbr
```

The function `concat` (lines 4-7) replaces spaces with underscores, French ligatures, and question marks. The function `map` (lines 9-23) iterates over the lines of the sample data, requests Wikipedia entries and their URLs, and converts them into DBpedia URLs. If no entry is found, a message is printed. The result is saved to a JSON file; an extract is shown in Fig. 1.

```
[
    "anemoine",
    "sorte de renonculacées à fleurs violettes, dite aussi coquelourde,
      passe-fleur ou pulsatille",
    "anemone pulsatilla",
    "https://dbpedia.org/resource/Pulsatilla_vulgaris"
],
[
    "zecharr",
    "espèce de faucon",
    "falcon",
    "https://dbpedia.org/resource/Falcon"
]
```

Figure 1: Mapping result: LINNÉTERMINUS (extract).

Evaluation of the PoC The mapping result is promising, despite the fact that five mappings are

nonsense. E.g., Old French *lecherant* (lickspittle), falsely leads to https://dbpedia.org/page/Licker: «a fictional creature from Capcom's Resident Evil series». *Datil* (date [fruit]) maps to a disambiguation page with person and place names, double dates, etc.; the correct mapping would be the entry 'Date_(fruit)' which in turn leads to the entry 'Date_palm', which again is wrong. Furthermore, one keyword could not be mapped by the script: *feve* "plante aquatique de la famille des Nélumbonacées [...], fève d'Égypte, Lotus sacré ou Lotus d'Orient (Nelumbo nucifera, Nymphæa Nelumbo L.); la graine de cette plante". In our test data set, we select the second Linnaean term, Nymphæa Nelumbo (Indian lotus), as the keyword to be mapped. However, the English Wikipedia does not list the Indian lotus under 'Nymphæa_Nelumbo' but instead under the first term, 'Nelumbo_nucifera' (the German Wikipedia redirects from one to another; the English site does not). All the other keywords, i.e., 114 out of the possible 120, have been correctly mapped.

5.1.6 Implementation

Use of French Wikipedia entries. The following steps aim to use the French originals and avoid the manual English translation of the keywords that we performed for the PoC. We test two ways to do this: First, we direct the algorithm to use the French Wikipedia instead of the English: `wikipediaapi.Wikipedia('fr')` (line 2 of the code above) but don't change the URL-replacement process. The algorithm produces 117 mappings. However, since DBpedia models the English Wikipedia entries, many of the produced mappings are incorrect. E.g., French *bois*, the woods, produces a link to the DBpedia entry 'Bois'[29], which is, however, a disambiguation page with person and place names. The correct hit would have been the entry 'Wood'.

Use of English Wikipedia equivalents. Next, the algorithm queries the Wikipedia API for French Wikipedia entries and, at the same time, for their English equivalents. `langlinks` is appended to the Python function `map` to test whether an English equivalent exists and if so, use its URL to generate the DBpedia URL (lines 6-15):

```
1  DEFINE FUNCTION map(data_poc):
2    SET entries_to_dbr TO data_poc
3    FOR row IN data_poc[1:]:
4      SET keyword TO concat(row[2])
5      SET page_py TO wiki_wiki.page(keyword)
6      SET langlinks TO page_py.langlinks
7      IF page_py.exists():
8        FOR k IN sorted(langlinks):
9          IF k EQUALS 'en':
10           SET url_en TO langlinks[k].fullurl
11       SET url TO page_py.fullurl
12       SET url_db TO str(url_en).replace('https://
13         en.wikipedia.org/wiki/',
14         'https://dbpedia.org/resource/')
15       row.append(url_db)
16     ELSE:
17       SET keyword TO 'unknown_entry'
18       row.append(keyword)
19   RETURN entries_to_dbr
```

Although this also produces incorrect mappings (e.g., when an English equivalent is missing[30] or when Wikipedia falsely allocates an English equivalent), the hit ratio is better than the first attempt.

Automatically identified keywords. We then implement solutions for automatically identifying the keywords to be mapped by the algorithm. Here, we work with a manually created test data set of 236 lexical units in the form of RDF data, e.g.:

```
deaf:ebenus skos:definition
    "bois de l'ébénier, ébène"@fr .
deaf:pivernaus skos:definition
    "goutte"@fr .
deaf:fie skos:definition
    "fruit du figuier (Ficus carica L.),
    comestible et de couleur violette,
    ..., figue"@fr .
```

Many sense definitions offer keywords for several approaches simultaneously, for example, a keyword for LINNÉTERMINUS and for GENUSPROXIMUS. Thus, we order the approaches by the expected mapping accurateness of their performance. E.g., LINNÉTERMINUS is more accurate than GENUSPROXIMUS and, consequently, the algorithm prefers the first method to the second.

The pseudocode (extract) reads as follows[31]:

```
SET linne TO re.compile(r'\(.* L\.\)')
SET linne_unobvious TO re.compile(r'\(([A-Z]
\w+|[A-Z]\w+\ \w+)\)')
SET linne_cap TO re.compile(r'([A-Z]\w+\
\w+(\ L.))')
SET linne_cap_single TO re.compile(r'([A-Z]
\w+(\ L.))')
SET linne_cap_unobvious TO re.compile(r'([A-Z]\w+\
\w+)')
SET linne_cap_single_unobvious TO re.compile
(r'([A-Z]\w+)')
SET last_word TO re.compile(r'(\,[^\,\r\n]|\;
[^\,\r\n])(\w+\ ?\w+)(\ et sim.|
,\ et sim.){0,1}(\??)(\ \(?\))?$')
SET single_word TO re.compile(r'^(\w+\ ?\w+)\??$')
SET sorte TO "sorte de"
SET sorte_apostr TO "sorte d'"
SET espece TO "espèce de"
SET espece_apostr TO "espèce d'"
```

[29] https://dbpedia.org/page/Wood.

[30] This is the case for ten keywords: 'Lèchefrite', baking sheet, 'Amertume', bitterness, 'Machine de guerre', apparatus belli, etc.

[31] The complete Python script and RDF data can be found on GitHub, https://github.com/SabineTittel/LexSemMapping.

```
20
21 DEFINE FUNCTION map_rdf(graph):
22   FOR s, p, o IN graph:
23     IF p EQUALS (skos + 'definition')
24     and type(o) EQUALS rdflib.term.Literal:
25       IF linne.search(o):
26         SET keyword TO concat(re.sub('.*\((.*)
27         (\ L\.)\).*', r'\1', o))
28         SET page_py TO wiki_wiki.page(keyword)
29         IF page_py.exists():
30           make_langlinks(s, page_py)
31           continue
32         IF linne_cap.search(o):
33           SET keyword TO concat(normalize(re.sub
34           ('(.*\ )([A-Z]\w+\ \w+)(\ L.)(.*)',
35           r'\2', o)))
36           SET page_py TO wiki_wiki.page(keyword)
37           IF page_py.exists():
38             make_langlinks(s, page_py)
39             continue
40         # all other keyword queries follow
41
42         ELSE:
43           graph.add((s, ontolex + 'isConceptOf',
44           Literal('to be mapped')))
45
46 DEFINE FUNCTION make_langlinks(s, page_py):
47   SET langlinks TO page_py.langlinks
48   IF langlinks:
49     FOR k IN sorted(langlinks):
50       IF 'en' IN sorted(langlinks):
51         IF k EQUALS 'en':
52           SET url_en TO langlinks[k].fullurl
53           SET url_dbr TO str(url_en).replace
54           ('https://en.wikipedia.org/wiki/', '')
55           graph.add((s, ontolex + 'isConceptOf',
56           dbr + url_dbr))
57         ELSE:
58           graph.add((s, ontolex + 'isConceptOf',
59           Literal('missing English equivalent to
60           French Wiki entry')))
61   ELSE:
62     graph.add((s, ontolex + 'isConceptOf', Literal
63     ('no equivalents to French Wiki entry')))
```

To find the keywords, the algorithm uses regular expressions and looks for pre-defined strings: catchwords (lines 1-15). The function `map_rdf` iterates over the parameter for the argument `graph` (line 21): subject, predicate, and object of the triples of the imported RDF data set (with the 236 lexical units). For all literal objects that follow the predicate `skos:definition` (line 23f.), the algorithm checks for the existence of keywords (line 25ff). For each keyword, the algorithm searches for entries in the French and English Wikipedia respectively and generates DBpedia URLs as described. It then adds a triple to the lexeme with `ontolox:isConceptOf` and the DBpedia URL respectively, or generates a message in case the mapping is unsuccessful (lines 59f., 63).

Evaluation. The four methods for mapping nouns achieve varying hit rates, with the LINNÉTERMINUS approach producing different results according to the syntax of the definition text described in chap. 5.1.1. Fig. 2 shows an extract of the results in the form of the RDF triples, and fig. 3 summarizes the results achieved for the data set with 236 DEAF entries.

```
deaf:wodlark skos:definition "espèce d'oiseaux. Comme toutes
  les alouettes elle appartient à la famille des Alaudidae,
  alouette lulu (Lullula arborea L.)"@fr ;
  ontolex:isConceptOf dbr:Woodlark .

deaf:zecharr skos:definition "espèce de faucon"@fr ;
  ontolex:isConceptOf dbr:Hawk .

deaf:abenlie skos:definition "sorte de tente?"@fr ;
  ontolex:isConceptOf dbr:Tent .

deaf:turguet skos:definition "plante, sous-espèce de céréale,
  amidonnier (Triticum turgidum L.)"@fr ;
  ontolex:isConceptOf "missing english equivalent to French Wiki entry" .

deaf:pere skos:definition "père"@fr ;
      ontolex:isConceptOf dbr:Father .
```

Figure 2: Result (extract) of automatic keyword search.

Lexical Units	Linné	Single Word	Last Word	GenusProximus *sorte de*	GenusProximus *espèce de*
236 overall	86	60	60	20	10
				overall: 30	
mapped	82	37	51	18	8
no equivalence	0	5	2	1	0
no Engl. equivalence	4	6	5	0	2
not mapped	0	12	2	1	0
mapping rate	**95.3%**	**61.7%**	**85%**	**90%**	**80%**
correct hits	77	32	43	13	8
disambiguation pages	5	2	7	4	0
incorrect hits	0	3	1	1	0
hit rate	**94%**	**86.5%**	**84.3%**	**72.2%**	**100%**
mapping overall					**194**
mapping rate overall					**82.4%**
hits overall					**173**
hit rate overall					**87.4%**

Figure 3: Evaluation of the mapping of 236 entries.

Interpretation of the results and extrapolation. The methods produce promising mapping rates and hit rates. The highest mapping rate shows the LINNÉTERMINUS method with 95.3% mappings and also a very accurate hit rate with 94%. The SINGLEWORD method achieves the lowest mapping rate with 61.7%. The highest hit rate is achieved by the GENUSPROXIMUS method with the catchword 'espèce de' with 100%; albeit, this result needs to be interpreted with the caveat that the absolute number of mappings for 'espèce de' is only eight – with 77 for the LINNÉTERMINUS method. This must also be considered for the low hit rate of (72.2%) achieved by the GENUSPROXIMUS method with the catchword 'sorte de'. As expected, the 84.3% hit rate of the LASTWORD method is rather low for the reasons explained above.

The overall result for all four methods is a mapping rate of 82,4% (194 out of 236) with 87,4% correct hits (173).

We see that 18 mappings lead to disambiguation pages in DBpedia, a result we cannot influence. E.g., *pié* m. "pied" maps to 'Pied_(disambiguation)' (with proper names, the Pied Piper of Hamelin, etc.) without redirection to

'Foot' (the correct DBpedia entry). Encouragingly, the number of semantically incorrect hits is low, with three for the SINGLEWORD method and one for both the LASTWORD and GENUSPROXIMUS methods. E.g., *diacalamant* m. "sorte de confection dont la base était le calament" wrongly maps to 'Sewing' (from the polysemic French terme *confection*); however, it is a concoction using calamint, a plant of the mint family. We consider the results (mapping rate and hit rate) to be satisfactory and thus extrapolate them to the DEAF totals: out of the 92,776 lexical units, 30,065.6 are, thus, potential mappings, and – out of these – 25,423,4 are potential hits. This equals 27,4% hits overall.

5.2 A Method for Non-Nouns

This method maps lexical units of lexemes that are not nouns (but also include nouns that have not been reached by the approaches described above), i.e., adjectives, adverbs, verbs: roughly 70% of the DEAF entries. The algorithm processes keywords in the definitions that can be mapped to entities of DBpedia. This aims at grasping the significant core elements from the sense of a given lexeme. Of course, this is only an approximation to the respective sense. Nevertheless, it represents a rough but automatic placement of the sense within the structure of an external knowledge base. To do this, the algorithm applies what we call the 'splitting method' (SPLITTING) where it tokenizes the definition texts, iterates over the tokens, and looks for those that can be mapped. The pseudocode is the following:

```
IF (re.findall('\w+', o)):
  FOR word IN (re.findall('\w+', o)):
    SET page_py TO wiki_wiki.page(word)
    IF page_py.exists():
      make_langlinks(s, page_py)
    ELSE:
      graph.add((s, ontolex + 'isConceptOf',
      Literal('to be mapped')))
```

Nota bene: We apply `re.findall` instead of `re.split` to avoid having to define identification rules for split perimeters.

A model case for this method is the adjective *lovin* adj. "a la manière d'un loup" (wolflike), with the tokenized result being `['à', 'la', 'manière', 'd', 'un', 'loup']`. From these tokens, the algorithm produces:

```
deaf:lou#lovin
  skos:definition "à la manière d'un loup"@fr ;
  ontolex:isConceptOf
    <https://dbpedia.org/resource/%C3%80>,
    dbr:D_(disambiguation),
    dbr:La,
    dbr:UN_(disambiguation),
    dbr:Wolf,
    "no equivalents to French wikipedia entry" .
```

We can interpret the result as follows:

- 'À' ([%C3%80], letter) (line 4),
- 'D_(disambiguation)' is a disambiguation page with 'D' representing 'differential equation', 'Delaware', 'Desktop Environment', etc. (line 5),
- 'La' equally, representing 'Louisiana', 'LucasArts' (a subsidiary company of LucasFilm Ltd.), a type of moth, etc. (line 6),
- 'UN_(disambiguation)' representing 'United Nations', a Korean music band, etc. (line 7);
- the only mapping with semantic value is `dbr:Wolf` (line 8);
- 'manière' is an entry in the French Wikipedia without an equivalent in the English Wikipedia (line 9).

Evaluating a larger number of such examples, we learn that the many incorrect hits must be limited. For this purpose, we create a list of words to be generally ignored by the algorithm, i.e., articles, pronouns, prepositions, and the like. We also include words that occur in many definitions but lead to false results such as:

- *manière* (see in the example above),
- *changeant*, present participle of *changer* (to change), which maps to 'List_of_Star_Trek_aliens#Changeling', a fictitious species of the Star-Trek universe,
- *référant*, present participle of *référer* (to refer to), which maps to 'HTTP_referer',
- and the adjective *sérieux* (serious) which maps to 'Paul_Sérieux', a French psychiatrist.

We import this list into the Python script.

Implementation. To test our method we create a data set with 100 entries: lexical units for 20 adjectives, 20 adverbs, and 20 verbs; we add 40 nouns that cannot be computed with the four methods, as described in chap. 5.1. A first test with the existing algorithm (without the SPLITTING method) confirms that all 100 entries cannot be mapped. With the algorithm using the SPLITTING method, however, the results are as shown in fig. 4.

The mapping rates of 55% up to 77.5% yield an average of 65%. We give an example of the

Lexical Units	Adv.	Adj.	Verb	Nouns	overall
number	20	20	20	40	100
mapped	12	11	11	31	65
no equivalence	1	6	7	10	24
no Engl. equivalence	6	5	5	12	28
not mapped	7	14	14	25	60
mapping rate	**60%**	**55%**	**55%**	**77.5%**	**65%**

Figure 4: Quantitative evaluation of SPLITTING method.

outcome for *efimere* adj. (a fever or a pain that lasts for about a day), which shows both successful mappings and a miss:

```
1 deaf:efimere skos:definition
2     "qui dure un jour ou peu plus (dit
3     de la fièvre, de la peine)"@fr ;
4     ontolex:isConceptOf
5         dbr:Day,
6         dbr:Fever,
7         "missing English equivalent to
8         French Wiki entry" .
```

Evaluation. To assess the quality of the mapping result of the SPLITTING method, we conduct an evaluation of each mapping for each lexical unit. For *efimere*, for example, the mapping to the entities 'Day' and 'Fever' are meaningful; the keyword 'peine' (pain) produces a result in the French Wikipedia but no English equivalent (lines 7-8).

Extrapolation to the DEAF data, all methods included. We extrapolate these results to the DEAF data. The total number of the DEAF lexical units that can be mapped by the SPLITTING method, i.e., that are not reached by the four methods LINNÉTERMINUS, SINGLEWORD, LASTWORD, and GENUSPROXIMUS (total 30,065.6, see above) is: $92,776 - 30,065.6 = 62,710.4$. With a mapping rate of overall 65% (see fig. 4), the SPLITTING method, therefore, has the potential to generate 40,761.76 mappings.

Together with the 25,423.4 semantically correct mappings of nouns, this results in an approximate amount of 66,185 semantically mapped lexical units. This corresponds to 71.34% of the total set of 92,776 lexical units.

5.3 Applying the Algorithm to the RDF Data Sets of the DEAF

As a litmus test for the validity of the extrapolation, we exclude the manually prepared test scenarios and apply the algorithm to actual RDF data: We use the results of automatic routines modeling the DEAF entries as Linked Open Data in RDF. We apply the algorithm to 300 datasets with 617 lexical units overall, including all parts of speech. The result is a mapping rate of 71.03%. Compared with the extrapolated rate of 71.34% mapped lexical units within our test scenario, we conclude that the validity of the extrapolation is confirmed. This is important for future applications of the methods to the 92,776 lexical units of the DEAF.

Evaluation. Following the example given for *efimere* adj. (see above), we manually assess the quality of each of the 617 mappings with respect to the sense of the mapped lexical unit. Examples of the quality evaluation and the overall findings are shown in fig. 5.

DEAF entry	Def.	≠ Mapp.	Mapp. ✓✓	Mapp. ✓	Mapping overall Abs.	Mapping overall Hit Ratio
fable	22	7	1	14	15	68.2%
faraon	3	0	1	2	3	100%
faucille	10	0	0	10	10	100%
fece	1	0	0	1	1	100%
festele	31	11	10	10	20	64.5%
festre	12	1	3	8	11	91.7%
fiel	28	6	10	12	22	78.6%
fievre	31	0	3	29	32	100%
figure	60	24	3	33	36	60%
flajol	31	11	6	13	19	61.3%
flamesche	1	0	0	1	1	100%
flaüte	17	6	1	10	11	64.7%
gratifier	1	1	0	0	0	0
guihale	1	0	0	1	1	100%
guimauve	1	0	1	0	1	100%
guindas	2	0	0	2	2	100%
guinlechier	2	0	0	2	2	100%
halstre	2	2	0	0	0	0
harigoter	7	2	0	5	5	71.4%
hart	35	18	0	15	15	42.9%
...	...	...	...	...	...	...
overall	617	169	77	368	445	
percentage		28.6%	12.7%	58.2%		71%

Figure 5: DEAF RDF data with LexSemMapping.

Explanation of the table columns:

- **DEAF entry**: entry name of an article,
- **Def.**: number of lexical units in the entry,
- ≠ **Mapp.**: no mapping, i.e., the total amount of the messages 'to be mapped' respectively, 'no equivalents to French Wiki entry', and 'missing English equivalent to French Wiki entry'; we also add the number of mappings that are semantically nonsense (the result of our qualitative evaluation),
- **Mapp.** ✓✓: number of semantically precise and correct mappings using the LINNÉTERMINUS, SINGLEWORD, and the LASTWORD methods,
- **Mapp.** ✓: number of the mappings through the GENUSPROXIMUS or the SPLITTING method that are semantically correct in an approximate way.

The qualitative evaluation of the mappings shows that 12,7% of the mappings produce semantically precise and correct hits, and 58,2% of the mappings produce approximately correct hits.[32] The latter are able to assign the lexical units to an extra-linguistic entity in the form of a first and rough classification; at the same time, it lays an excellent foundation for a manual and more precise elaboration of the mapping for these lexical units.

6 Result and Outlook

As an overall result, we can state the following: Due to the heterogeneity of the sense definitions, achieving 100% correctness in the LexSemMapping of all 92,776 lexical units of the DEAF to DBpedia is not realistic. However, the methods we have developed (LINNÉTERMINUS, SINGLEWORD, LASTWORD, GENUSPROXIMUS, SPLITTING) clearly approach our goal: the automatic LexSemMapping of lexical units of the DEAF dictionary. Our methods are able to successfully map large portions of the total set of lexical units; approx. 71% of the lexical units (= 53,996) can be mapped: approx. 12.7% (= 11,783) will be mapped accurately in terms of semantic content, and approx. 58.2% will be mapped in an approximate, yet meaningful way.

Based on this extrapolation, we reason that applying the algorithm to the RDF data sets of the DEAF is able to enhance the RDF data in a significant way. It establishes semantics-based, language-independent access to potentially almost 65,800 lexical units of the dictionary by linking to DBpedia. The RDF data of the DEAF will be released under Public Domain in a triple store by the Heidelberg Academy of Sciences and Humanities (HAdW) or on `https://lod.academy/`, a hub for Linked Open Data and Graph Technologies run by the Academy of Sciences and Literature Mainz and the HAdW.

With the achieved result, we deduce that approx. 29% of the lexical units still need to be mapped manually. With the estimated 10 min per mapping, this still adds up to roughly 65 days of work. What comes to mind are methods utilizing artificial intelligence to interact with the sense definitions of the DEAF. Our first impression, however, was not very promising because the definition texts seemed too heterogeneous for an AI model to identify patterns that could lay the foundation for a successful approach. Nonetheless, recent developments in this sector such as the emergence of ChatGPT[33] for instance, suggest considering the topic anew.

Furthermore, we utilized the automatic matching of French Wikipedia entries with corresponding English entries offered by the Wikipedia API. To bypass this error-prone step, it could be worthwhile to test integrating a machine-driven translation from French into English recurring to external services such as the DeepL API.[34]

Possible generalization of the approach. Lexicographic resources typically contain lexical units—words and their senses, the latter being defined through translations into a (modern) language, through genus-differentia definitions or other methods. We know how time consuming a manual lexico-semantic mapping of the lexical units is. With (i) its specific solutions for different kinds of definitions, (ii) the possibility to feed varying languages into the algorithm (adapting the query to the Wikipedia API to the particular language) and (iii) given the hit rate of the algorithm, we conclude that a generalization of our LexSemMapping approach is promising: It can be re-used both for the semantic enhancement of already existing RDF resources and for newly approached Linked-Data modeling of (historical) linguistic resources. Also, related approaches could benefit, e.g., the aforementioned endeavor of the LEI to install an onomasiological structure and where DBpedia entities could be added to the HTE taxonomy to establish interoperability within the Linked-Data landscape.

References

Kurt Baldinger. 1971-2020. *Dictionnaire étymologique de l'ancien français – DEAF*. Presses de L'Université Laval / Niemeyer / De Gruyter, Québec, Canada / Tübingen / Berlin, Germany. [Kurt Baldinger (founder), continued by Frankwalt Möhren and Thomas Städtler; electronic version DEAF*él*: `https://deaf.ub.uni-heidelberg.de`].

Robert-Henri Bautier, Robert Auty, and Norbert Angermann. 1977-1998. *Lexikon des Mittelalters*. Artemis, München.

[32] Examples of RDF data sets with mapped lexical units can also be found at GitHub: `festre_mapped.ttl`, `fiel_mapped.ttl`, etc.

[33] `https://openai.com/blog/chatgpt/`.

[34] `https://www.deepl.com/pro-api?cta=header-pro-api`.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language. Reference. W3C Recommendation 10 February 2004. URL: https://www.w3.org/TR/2004/REC-owl-ref-20040210/.

Andrea Bellandi, Emiliano Giovannetti, and Anja Weingart. 2018. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information*, 9 (3), 52.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5:1–22.

Andreas Blank. 2001. *Einführung in die lexikalische Semantik*. Niemeyer, Tübingen.

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2018. Models to represent linguistic linked data. *Natural Language Engineering*, 24(6):811–859.

Andrea Bozzi. 2016. Un'ontologia per il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*). In David Trotter, Andrea Bozzi, and Cédric Fairon, editors, *Actes du XXVII^{e} Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16: Projets en cours; ressources et outils nouveaux*, pages 55–63. ATILF, Nancy.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 518:379–386.

Philipp Cimiano, John McCrae, Paul Buitelaar, and Elena Montiel-Ponsoda. 2013. On the Role of Senses in the Ontology-Lexicon. In Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing*, pages 43–62. Springer, Berlin / Heidelberg.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Final Community Group Report 10 May 2016. https://www.w3.org/2016/05/ontolex/.

Thierry Declerck, Eveline Wandl-Vogt, and Karlheinz Mörth. 2015. Towards a Pan European Lexicography by Means of Linked (Open) Data. In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of eLex 2015, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, pages 342–355, Ljubljana/Brighton. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Mariafrancesca Giuliani and Itziar Molina Sangüesa. 2020. Hacia Una Taxonomía Integrada En La Redacción y Revisión De Diccionarios Históricos. *Bollettino Dell'Opera Del Vocabolario Italiano*, 25:325–374.

Axel Herold, Lothar Lemnitzer, and Alexander Geyken. 2012. Integrating Lexical Resources Through an Aligned Lemma List. In Christian Chiarcos, editor, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 35–44. Springer, Berlin / Heidelberg.

Anas Khan, Christian Chiarcos, and Thierry Declerck et al. 2022. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web*, 13:1–64. DOI: 10.3233/SW-222859.

Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. URL: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

Itziar Molina Sangüesa. 2023. Diseño de una ontología aplicada a la lexicografía histórica digital. *Círculo de Lingüística Aplicada a la Comunicación*, 93:229–242. DOI: 10.5209/clac.72654.

Max Pfister. 1979–. *LEI. Lessico Etimologico Italiano*, founded by Max Pfister, directed by Elton Prifti and Wolfgang Schweickard. Reichert, Wiesbaden.

Marielene Putscher. 1974. *Pneuma, Spiritus, Geist. Vorstellungen vom Lebensantrieb in ihren geschichtlichen Wandlungen*. Steiner, Wiesbaden.

Willy Richard. 1959. *Untersuchungen zur Genesis der reformierten Kirchenterminologie der Westschweiz und Frankreichs: mit besonderer Berücksichtigung der Namengebung*. Francke, Bern.

Matthew Cheung Salisbury. 2015. *The secular liturgical office in late medieval England*. Brepols, Turnhout.

Heinrich Schipperges. 1990. *Geschichte der Medizin in Schlaglichtern*. Meyers Lexikonverlag, Mannheim.

Sabine Tittel. forthcoming. *Integration von historischer lexikalischer Semantik und Ontologien in den Digital Humanities.* Heidelberg.

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 LREC 2018. GLOBALEX Workshop (GLOBALEX-2018), Miyazaki, Japan, 2018*, pages 58–66, Paris. ELRA.