Ceci n'est pas un dictionnaire. Adding and Extending Lexicographical Data of Medieval Romance Languages to and through a Multilingual Lexico-Ontological Project

Sabine Tittel[1]

[1]Heidelberg Academy of Sciences and Humanities, Seminarstraße 3, D–69117 Heidelberg
E-mail: sabine.tittel@hadw-bw.de

Abstract

Historical lexicography of the Romance languages currently finds itself in a difficult place since the funding of some important dictionaries ended. The newly launched project ALMA will contribute to the future of these dictionaries' content. ALMA combines methods of historical lexicography, text philology, corpus linguistics, and the history of sciences with a Linked Data approach and ontology development. It adopts a Pan-Romance perspective focusing on medieval Italian, French, and Occitan / Gascon within two knowledge domains, 'medicine' and 'law'. ALMA's goals include re-using, extending, further processing, and disseminating lexicographical data by integrating it into its work pipeline. This makes for benefits on both sides: Pivotal for the ALMA project is the anchoring of its philological and lexicological work within the framework of the entire languages examined by the dictionaries. The dictionaries, most notably those whose funding ended, profit by seeing their linguistic, textual, and historico-cultural knowledge put into new formats—e.g., Linked Data—, contexts—e.g., Pan-Romance—, and correlations—e.g., through linking to the historicized domain ontologies ALMA will develop. This introduces the valuable dictionary contents to a knowledge circulation that goes beyond their original scope and ensures its long-term re-use in a somewhat concealed way.

Keywords: Historical Lexicography; Medieval Romance Languages; Corpus Linguistics; Ontology; Linked Data

1. Introduction

Dictionaries of historical language stages are at the core of historical lexicology, text philology, and historiography. They provide the means for grounding research on stable knowledge of the language for generations of researchers. For example, to create a scholarly text edition, the permanent consultation of the dictionaries pertinent to the language and language stage of the respective text is vital. For only the development of a text glossary with the inventory of the lexemes and their meanings can enable and further the editor's understanding of the text. The development of this glossary is not only a philological but also lexicological process and thus needs dictionaries: The text represents, albeit recorded in one or several manuscripts and thus potentially modified, the *parole* of the author while the dictionaries analyze the language on the level of the *langue*.[1] The meaning of a word in a text—as part of the *parole*—cannot properly be grasped without its embedding in the semantic scope of the same word whose senses and sub-senses, its uses as metaphor, metonymy, or other figures of speech, are explained in the dictionaries.

[1] Ferdinand de Saussure, *Cours de linguistique générale*, dichotomy of *langue* and *parole*, edition for example by Wunderli (2013).

Hence, dictionaries both are and facilitate foundational research. Within our context of the historical Romance languages, these include the *Französisches Etymologisches Wörterbuch* (FEW, von Wartburg, 1922–) which presents the diachronic development of the French language until present-day, the *Lessico Etimologico Italiano* (LEI, Pfister, 1979–), the *Dictionnaire étymologique de l'ancien français* (DEAF, Baldinger, 1971–2021), the *Dictionnaire de l'occitan médiéval* (DOM, Stempel, 1996–2013), the *Dictionnaire onomasiologique de l'ancien gascon* (DAG, Baldinger, 1975–2021), the *Dictionnaire du moyen français* (DMF[2]). These are comprehensive, long-term, and internationally well-known dictionary projects of the medieval Romance languages that provide synopses of the knowledge of the particular historical language stage and regional specification.

Despite being important resources in the field of lexicography, in 2020 and 2021, the funding of several of these long-standing endeavors has come to an untimely end leading to the situation that the dictionaries «find themselves currently at a difficult juncture», Selig et al. (2023: 296). This concerns, amongst others, the DEAF, the DOM, and the DAG. Fortunately, the data, dictionary writing system, and digital framework of the latter have been adopted by the University of Zurich where it will be merged into the *Dictionnaire étymologique d'ancien gascon* (DEAG).[3] However, the DAG as a printed oeuvre has come to an end. Much earlier already, in 2006, the financial support for the *Dictionnaire onomasiologique de l'ancien occitan* (DAO) also ended (Glessgen & Tittel, 2018) and in 2007, the *Diccionario del español medieval* (DEM, Müller, 1994–2005) was discontinued.[4] Furthermore, the DMF, while still hosted by the ATILF institute[5], has been solely edited in a while by its director Robert Martin after his retirement.

The newly launched long-term project ALMA (*Wissensnetze in der mittelalterlichen Romania / Knowledge Networks in Medieval Romance Speaking Europe*) takes this development into account. ALMA is an inter-institutional project with a funding period of 22 years carried out by the Heidelberg Academy of Sciences and Humanities (HAdW), the Bavarian Academy of Sciences and Humanities (BAdW), and the Academy of Sciences and Literature Mainz (ADW Mainz).[6] One of ALMA's goals is to integrate four lexicographical state-of-the-art dictionaries—DAG, DEAF, DOM, and LEI—into its scientific concept and work pipeline to produce a highly comprehensive and valuable resource that far exceeds the typical consultation and quotation of dictionaries. This benefits both the ALMA project and the dictionaries, especially those whose funding has been cut.

In this paper, we introduce the ALMA project (Chapter 2) and show the manifold relations to lexicography: re-use, extension, further processing, and dissemination (Chapter 3). We continue with an evaluation of these relations that promise benefits for both the ALMA project and the re-used dictionaries through a minimal case study of the Middle French medical term *addicion* f. (Chapter 4) and close with a short conclusion (Chapter 5).

[2] Version 2020, ATILF – CNRS & Université de Lorraine, http://www.atilf.fr/dmf/; this and all following web publications have been accessed 2023-05-24.

[3] And integrated into a larger, Pan-Romance dictionary *Lexique étymologique de la Galloromania médiévale* (LEGaMe) Glessgen (2023).

[4] In a more promising development, the *Diccionario del español medieval electrónico*, DEM*el* (Arnold & Langenbacher-Liebgott, 2022–), has resumed the work on medieval Spanish by making the slip inventory (33,000 lemmata with about 900,000 attestations) and prospectively, the printed DEM digitally accessible.

[5] *Analyses et traitement informatique de la langue française*, Nancy, https://www.atilf.fr/.

[6] https://www.hadw-bw.de/alma; directed by Elton Prifti / Wolfgang Schweickard (ADW Mainz), Maria Selig (BAdW), and Sabine Tittel (HAdW).

2. Introducing ALMA

The ALMA project aims to investigate the interaction between language, knowledge and scholarship in the Middle Ages. The field of observation is the Romance cultural sphere that sees the emergence of new knowledge networks expressed in vernacular languages in the time period between around 1100 and 1500 AD. The project traces how medieval Italian, French, Occitan / Gascon[7], and also Catalan and Spanish are developed into languages of knowledge and scholarship within new functional areas of language that are technically and conceptually complex. This will be exemplified by two knowledge domains, namely 'medicine' and 'law'. These technical, 'scientific' languages depicting knowledge and scholarship are a particularly important part of the intellectual and cultural heritage of Europe. Concurrently, the Romance languages are major carriers of a cultural exchange in the Middle Ages that starts to establish the European identity as a knowledge society.

ALMA's concept combines methods of linguistics, text philology, and the history of sciences with technologies of the digital humanities and ontology engineering.

Romance languages have a rich textual tradition.[8] To make—a part of—this textual tradition accessible and to establish its empirical research basis at the same time, ALMA will compile credible, domain-specific corpora for 'medicine' and 'law'. These corpora will consist of newly established text editions and of digitized works covering medieval Italian, Old French, Old Occitan, and Old Gascon, giving access to an important cultural sphere of medieval Romance-speaking Europe. The text selection rests on the substantial, decades-long experience gathered by the dictionaries DAG, DEAF, DOM, and LEI, which directly benefits ALMA.[9] The corpus texts lay the foundation for the reconstruction of the main concepts and concept networks of the two knowledge domains. Applying quantitative methods of corpus linguistics (Hirschmann, 2019)—in particular absolute and relative frequency analyses and co-occurrence analyses—will help carve out these concepts. They provide the basis for the lexical-semantic studies that analyze the internal structure of the concept networks and the depth of their linguistic representation. The lexical-semantic studies also discuss the origin and dissemination of lexical innovations together with the new matters denoted, deepening the knowledge about medieval communication channels. Thus, these studies will evaluate the empirical, quantitative methods of corpus linguistics in a historico-linguistic context combined with the hermeneutical, qualitative approach of historical linguistics. It is at this stage that the lexicographic resources come into play.

3. Relations to Lexicography

The project's relations to lexicography are manifold: ALMA (1) re-uses, (2) extends, (3) processes, and (4) disseminates existing and well-proven dictionary data.

[7] For the long-lasting discussion of the differentiation of Occitan and Gascon, see Glessgen (2021); Selig et al. (2023: 266).

[8] DEAFBiblEl (Möhren, 2022), https://alma.hadw-bw.de/deafbibl/, lists >80 (large and small) texts with medical content and >30 with law-related content for Old French alone. The text corpus of Old French legal documents, *Documents linguistiques galloromans* (DocLing, http://www.rose.uzh.ch/docling), comprises >2,200 medieval French charters (deeds of donation, contracts of purchase, inheritance matter, etc.) dating between 1205 and ca. 1450 AD.

[9] Additionally, the work is supported by the analysis of complementary (Medieval) Latin and vernacular text corpora that are already available for digital research, e.g., the many editions of small legal documents provided by DocLing. This is particularly relevant for the Spanish and the Catalan textual traditions where ALMA will not create its own corpora.

3.1 Re-use

The workflow combining quantitative machine-driven with qualitative competence-linguistic methods is controlled by drawing on the state-of-the-art dictionaries: the lexicography of the Gallo- and Italoromania, i.e., DAG, DEAF, DOM, LEI primarily, but also FEW and DMF, flanked by the dictionaries of the Iberoromania and of (medieval) Latin.

The cognitive step from a given lexeme, its absolute and relative frequency, and its co-occurrences—the result of the corpus analysis—to one or several concepts, is made by analyzing the meaning(s) of the lexeme in all of the text passages and in constant confrontation with the language system documented in the dictionaries. Self-evidently, ALMA follows standards of quoting dictionary entries. Furthermore, it has the unique advantage of being able to re-use—through database access—the published as well as raw data of the DEAF, DOM, LEI, as well as the DAG (depending on its ongoing integration into LEGaMe). This means being able to evaluate the source materials in the *fichiers* (slip inventories) of the dictionary resources, containing millions of paper slips with text references. Since the funding of these dictionaries, apart from the LEI, has recently ended, re-use through ALMA is an excellent means to keep the valuable data alive as part of an innovative workflow.

A second significant aspect of dictionary re-use concerns the bibliographical supplements of DEAF (DEAFBiblEl), DOM (DOMBibl[10]), and LEI (*Bibliografia Generale online* / BiG[11]), all of which are reference works with immense value for studies on historical linguistics of the Romance languages and text philology. Based on the DEAFBiblEl model, which became the state-of-the-art work used by many monographs, journals, and other dictionaries, ALMA will create a critical research bibliography assessing primary literature—for and beyond the corpus texts—, secondary literature, and dictionaries. This will serve as the bibliographical groundwork entangled with the corpus texts and the lexical-semantic analyses, and also facilitate validation and enrichment of corpus text information. It will also be published as a stand-alone research instrument. While the original bibliographical works will be preserved as such, the pertinent data of DEAFBiblEl (for 'medicine' and 'law' in Old French) and of DOMBibl (Old Occitan / Gascon) will be fully integrated into the ALMA database and extended therein. The comprehensive LEI BiG (Italian) will be closely interlinked on the level of each mentioned siglum (through APIs for database communication), but will remain an external, independent resource since it is a vital module of an ongoing dictionary project.

3.2 Extension

As mentioned earlier, the lexical-semantic studies carried out by ALMA build on corpus material that has only been partly considered by dictionaries. Here, the ALMA corpora will enlarge the material basis for lexicography in a significant way. An example is the text edition of the *Chirurgia magna* by Gui de Chauliac, written after 1363 and translated into many, Romance and non-Romance languages (Tittel, 2004: 17-29). This key text of the field of medicine provided the foundation for didactic surgery and became very influential until the 17th c. The first treatise of the text in the oldest French manuscript (Montpellier, Ecole de Médecine H 184 [2nd third 15th c], f^{os} 14v^{o}-36v^{o}) is accessible through

[10] http://www.dom-en-ligne.de/.
[11] https://lei-digitale.it/.

GuiChaulmT (Tittel, 2004) and its terminology found its way into DEAF, DMF, and FEW. Despite its importance, the French text as a whole (with 254 f^{os}) still lacks an edition (as well as translations). ALMA will fill this gap and in doing so, provide valuable data with great potential for research into the development of the language of medicine.[12]

Within the two domains 'medicine' and 'law', the lexical-semantic studies will add to, advance, or even replace entries of the four dictionaries DAG, DEAF, DOM, and LEI: The confrontation of the dictionary data with the new, comprehensive material accessible through the corpora will substantially extend the lexicographical knowledge documented so far. Merging corpus texts and dictionaries' information will realize a vital communication between the *parole* of a text (of many texts, respectively) and the description of the *langue* by the lexicographical resources. Also, the dictionary data that is typically focused on a single language will be put into a multilingual, Pan-Romance context. This will shed new light on the terminology: the semantic scope of the lexemes in each language, the history of the lexemes and their etymology (*histoire du mot*), and the history of the designated concepts (*histoire du concept*) across the languages. It will thus enhance the comprehension of the inter-relatedness of the medieval languages stages of Italian, French, Occitan, and Gascon, which is hitherto scattered among the individual dictionary publications.

The lexical-semantic studies have many features similar to a dictionary:

1. (Multilingual) lexemes as the heads of lexical-semantic analyses; lemmatized in each language,
2. Registration of senses and sub-senses in a hierarchical, tree-like structure reflecting semantic shift,
3. State-of-the-art genus-differentia definitions of the senses following Möhren (2015: 407–417),
4. An apparatus—separated from the semantics—documenting the dated and regionally classified graphical realizations of the lexemes,
5. Contexts (taken from the corpus material) for encyclopedic illustration of the senses,
6. Discussion of the etymology and *histoire du mot* typical for many historical dictionaries,
7. Close-knit interlinking with other lexicographic resources and text corpora.

Since the ALMA project will have access to the online publications of DEAF, DOM, LEI, and potentially DAG (DEAG, respectively), it will be possible to indicate within these publications that a given dictionary entry is incorporated and advanced within a lexical-semantic study; we will return to this with our case study in Chapter 4.2.

3.3 Further Processing

An innovation of the ALMA project is the combination of the philological and linguistic approach with Semantic Web technologies. ALMA's goals include modeling the project's results as Linked Open Data (LOD[13]) in *Resource Description Framework* (RDF, Cyganiak et al., 2004–2014) using standard vocabularies such as OntoLex-Lemon (Cimiano et al., 2016). The advantages of modeling data as LOD comprise structural and conceptual

[12] See Tittel (2004: 53-58) for an evaluation of the findings of the first, French treatise.
[13] https://www.w3.org/DesignIssues/LinkedData.html.

interoperability (through same format and shared vocabularies such as OntoLex-Lemon), accessibility (via standard Web protocols), and resource integration (through interlinking data), resulting in cross-resource access (Chiarcos et al., 2013). A pivotal aspect for establishing cross-language access to the content of historical linguistic resources—to words and their meanings—is lexico-semantic mapping: the mapping of concepts (of things) expressed through representations in historical languages (words) to an entity of an external, language-independent knowledge base of the Semantic Web (Tittel, accepted). To enhance this lexico-semantic mapping, ALMA will develop domain-specific ontologies for medicine and law. These ontologies will be historicized, taking into account the specificity of medieval explanation patterns. This bridges the historical semantic gap between historical concepts and entities of modern ontologies—for example, of modern physiology that differs significantly from the medieval humoral pathology and doctrine of pneumata—and prevents anachronistic classifications.

The LOD modeling covers the lexical-semantic studies (as well as text editions and bibliographical data), including the incorporated material of the four dictionaries. Also, the project takes a step forward in that it extends modeling to the original dictionary articles in their entirety, thus feeding full DOM, DEAF, LEI (and possibly DAG) entries as RDF resources into the Semantic Web. Preparatory work on the RDF-modeling of DEAF and DAG (Tittel & Chiarcos, 2018; Tittel, forthcoming) and LEI (Nannini, forthcoming) has already been successfully performed. The concepts represented by the lexical units of the dictionaries will be mapped to the entities of the historicized domain ontologies for medicine and law developed by ALMA.

The following code example shows an extract of a DEAF entry as LOD/RDF serialized in Turtle (Prud'hommeaux & Carothers, 2014) and automatically created from XML with XSLT and Python scripts:[14]

```
@prefix dbr:      <https://dbpedia.org/resource/> .
@prefix dct:      <http://purl.org/dc/terms/> .
@prefix deaf:     <https://deaf.ub.uni-heidelberg.de/lemme/> .
@prefix decomp:   <http://www.w3.org/ns/lemon/decomp#> .
@prefix lexinfo:  <https://lexinfo.net/ontology/3.0/lexinfo#> .
@prefix olia:     <http://purl.org/olia/olia.owl#> .
@prefix ontolex:  <http://www.w3.org/ns/lemon/ontolex#> .
@prefix rdfs:     <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos:     <http://www.w3.org/2004/02/skos/core#> .
@prefix vartrans: <http://www.w3.org/ns/lemon/vartrans#> .

# --- lexical entry ---------------------------------
deaf:fiel a ontolex:LexicalEntry , ontolex:Word ;
  lexinfo:partOfSpeech "m."@fr ,
  lexinfo:Noun ;
  lexinfo:gender lexinfo:masculine ;
  ontolex:canonicalForm deaf:fiel_form_fiel .
deaf:fiel_form_fiel a ontolex:Form ;
  ontolex:writtenRep "fiel"@fro .

# graphical variant
deaf:fiel ontolex:otherForm deaf:fiel_form_fel .
deaf:fiel_form_fel a ontolex:Form ;
  ontolex:writtenRep "fel"@fro .

```

[14] Examples of complete DEAF entries modeled as LOD in RDF can be found on GitHub, https://github.com/SabineTittel/LexSemMapping/tree/main/results.

```
# collocation "fiel de terre", sense sense1.h
deaf:fiel_de_terre a ontolex:LexicalEntry , ontolex:MultiwordExpression ;
  decomp:subterm deaf:fiel ;
  vartrans:lexicalRel lexinfo:collocation ;
  rdfs:label "fiel de terre"@fro .

deaf:fiel_de_terre ontolex:sense deaf:fiel_sense1.h ;
  ontolex:evokes deaf:fiel_sense1.h_lexConcept .

deaf:fiel_sense1.h a ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf deaf:fiel_sense1.h_lexConcept ;
  ontolex:usage dbr:Metonymy ;
  olia:hasRegister olia:TechnicalRegister ;
  dct:subject dbr:Botany .

deaf:fiel_sense1.h_lexConcept a ontolex:LexicalConcept ;
  skos:definition "plante herbacée [...], petite centaurée"@fr ;
  ontolex:isConceptOf dbr:Centaurium_erythraea ;
  ontolex:lexicalizedSense deaf:fiel_sense1.h .
```

3.4 Dissemination

ALMA will disseminate the pertinent dictionary articles in the form of RDF resources and make them accessible for semantic research driven by Semantic Web technologies, a contribution well beyond ALMA's core focus.

4. Evaluation: More than Old Wine in New Bottles

We see contributions leading to significant advancements on both sides: ALMA and lexicography.

4.1 Contribution of the Lexicographical Data to ALMA

4.1.1 Reflections on Corpus Integration and why Dictionaries Help

Limiting lexicological research to the material of a self-contained corpus is a problematic approach. All results generated by the analysis of a corpus, irrespective of its composition and size can only be relevant to the subset of the language represented in that corpus. This is because creating a corpus means drawing corpus borders which generate absences. E.g., studying a corpus of the works of a single, particular author, such as Chrétien de Troyes, the famous French poet and founder of the textual genre of the chivalric romance[15] is interesting but does not reveal how his language differs from that of other authors. The insights gained from such a study will be limited (cp. Filatkina, 2009: 79).

ALMA chooses a discourse tradition—technical texts—as the constitutive feature of its corpus because it is necessary to limit the research material. And yet, it is incorrect to presume that only texts that are assumed to belong to this discourse tradition are relevant for the study of the terminology in question. In order to capture the essence of a term, the entire literature (texts of other technical domains, historiography, belles

[15] https://viaf.org/viaf/87681171/. For his language, see the *Dictionnaire Électronique de Chrétien de Troyes*, http://www.atilf.fr/dect/.

lettres, etc.) is relevant in shedding light on the quality of the word in the language as a whole. For this reason, a corpus that included all—but only—the technical texts of the given domain, per see, would not be able to make valid statements about the existence and meaning of technical vocabulary. An example is the Old French nomenclature of navigation that occurs in the *Vie seint Edmond le rei*, an Anglo-Norman hagiographic poetry from ca. 1193: *dromunz*, *chalanz*, *esnekes*, *hallos* etc., Kjellman (1935: 2029-2034), all designations for particular ships. On the other hand, the internal differentiation of the corpus must be supported by comparison with other literary or technical texts that are accessible through dictionaries. E.g., deeds are more often about sales than is a *chanson de geste*, the medieval epic poem. This naturally leads to a more frequent use of lexemes like Old French *achat* "purchase" or *vente* "sale" in deeds than in a *chanson de geste*. However, this does not imply that they necessarily have to be diaphasically bound, Glessgen (2005: 226).[16]

Consequently, interpreting corpus material must be done with great caution, and conclusions about the language must be made only after a recontextualization within the language as a whole. This can be achieved by matching them against broader lexicographical works.[17]

4.1.2 Lemmatization

A crucial part of corpus constitution is the tokenization of the text involving lemmatization and part-of-speech annotation. This process can adapt the models already established by the dictionaries building on the standards of the disciplines. For Old French, for example, this is the lemma list of the DEAF that accords to current rules of lemmatization.

4.1.3 Anchoring the Lexical-Semantic Studies

The lexical-semantic studies will be written based on lexicographical grounding. While ALMA focuses on the languages of two particular domains, the dictionaries examine and describe the languages as a system with all functional areas, beyond the technical vocabulary in question. Thus, lexicography enables anchoring the findings within the framework of entire languages. This is crucial for a proper grasp not only of the technical terms to be analyzed but also of the words of the textual contexts, as well as of the contexts of each term. If the technical texts often reveal that a lexeme is of special interest because it designates a special thing and thus its sense definition makes for a new sense in the dictionaries, the consultation of the lexicographical resources might show the opposite: the history of the concept and of the lexeme with its etymology often makes clear that it is really only one concept and no new sense: «Comme réflexion de contrôle face à un 'nouveau' sens, on peut se dire que tout sens insolite est un sens erroné», Möhren (2015: 416).

[16] See Coseriu (1980) for diaphasical, diastratical, and diatopical aspects of language.

[17] Cp. Kabatek (2016: 4): «El corpus contiene lengua, naturalmente, pero el corpus no contiene *la lengua*, ni como objeto abstracto, ni como objeto concreto y mental. El corpus […] nos ofrece una ventana que permite acceder a una parte de esta, pero no al todo, y deja, por tanto, abierta la especulación acerca de lo que no se puede ver». Also Oesterreicher (2006: 485-490) for examples of how, for 16th-century American Spanish, corpus linguistic research yields results that do not stand up to competence-linguistic scrutiny.

4.1.4 Shedding Light on Cross-Domain Relations

The lexical-semantic studies will also concentrate on identifying connections between the various knowledge domains. For medicine, the connection to the domain of astronomy can be seen, for example, in the polysemy of Old French *mirac* and its cognates: The lexemes represent the concept ⌜abdominal wall⌝ and also designate the star Beta Andromedae. In the metabolic-pathological field particularly, such connections can often be observed and must be considered when analyzing the concepts of 'healthy' and 'sick'. The comparison with lexicographical data from all knowledge domains helps clarify such connections.[18]

4.2 Contribution of ALMA to Lexicography

We argue that the ALMA project will make two major contributions to the lexicographical resources mentioned above: (1) The enlargement of the lexeme inventory and the enhancement of existing entries through a multilingual perspective and through new findings in new texts that are, furthermore, exploited in a way supported by the machine, (2) the modeling as LOD.

4.2.1 Enlargement of the Lexeme Inventory and Enhancement of Entries

In the following, we describe a minimal case study for our first argument, the impact on lexeme inventory and entries of the dictionaries: *Addicion* f. is—to the best of our current knowledge—a Middle French medical terminus expressing the concept ⌜protuberance of an osseous or cartilaginous structure⌝. We find the terminus attested in Middle French GuiChaulmT 316; 318; 390; etc., defined as "éminence à la surface d'une structure osseuse ou cartilagineuse" in Tittel (2004: 285). Presuming that ALMA will study this concept, we look into the four dictionaries DAG, DEAF, DOM, and LEI:

DAG The dictionary was founded in 1955 by Kurt Baldinger[19] (who also founded the DEAF) and was printed from 1975 to 2021; the preparation of the online version DAGél began in 2014 (Glessgen & Tittel, 2018: 805-808). The DAG has a checkered history due to changes in finances with several concept shifts and alterations in its material base with respect to the time span treated in the dictionary (originally the Gascon from ca. 1100 AD until the end of the 16th c, then until ca. 1300, then—for DAGél—until ca. 1500). DAGél was never made open to the public and is—since 2021—being turned into the DEAG and integrated into the Pan-Romance endeavor LEGaMe. As concerns our data quest for *addicion*, the data of the DAGél does not include a Gascon cognate of this term.[20] However, the medieval Gascon scripturality is almost exclusively limited to the text genre of documents (testaments, charters, court records, etc.). Hence, it is unsurprising to not find the term with a medical sense. Nevertheless, it is notable that no attestation at all (with whatever meaning) can be found.

[18] The study on *mirac* also shows the difficulty in capturing the precise scope of a concept in a knowledge area that is still under development: the different authors use the lexeme *mirac* (in its realizations in the individual languages, respectively) in slightly different ways. Comparing the preliminary findings with dictionary knowledge will be helpful.

[19] https://viaf.org/viaf/109932631/.

[20] We thank M. Glessgen for the database search.

DEAF The DEAF was edited between the 1960s and 2021 and published in print (since 1974) and online (since 2010). The online edition DEAF*él* consists of a two-tier system consisting of DEAF*plus*, the scholarly dictionary edited for decades but then limited to letters D-K (approx. 10,000 entries), and DEAF*pré*, the dictionary's raw data of the remaining letters of the alphabet which is published only online: 1.5 million slips with references to more than 10 million attestations, pre-structured into >70,000 preliminary dictionary entries with maximum assistance by the machine. Since the letter A is not part of DEAF*plus*, we turn to DEAF*pré* with its valuable yet unverified material which has: ADDICION, https://deaf.ub.uni-heidelberg.de/lemme/addition1[21]: "action d'ajouter qch., ce qu'on ajoute à qch; accroissement". This 'sense definition' is developed from the slip material, the word family, and the etymon, and does not comply with good definition rules (accounting for the preliminary nature and very limited time spent on editing this DEAF*pré* entry), cp. above. A sub-sense denoting the anatomical concept attested in GuiChaulmT[22] is missing.

DOM The DOM, whose preparatory work began in the 1960s, was published from 1996 to 2013 in printed form; since then, it has only been accessible online as DOMél. DOMél integrates its own research and existing dictionaries into its publication to create a lemma list covering the whole alphabet. The entry production of DOMél follows the concept of «cumulative development» (editing the dictionary) and «incremental functionality» (creating access), Klein (2004: 28f.). As of 2023, DOMél combines 1,845 digitized DOM entries, 37,998 entries uniting entries of two other dictionaries under DOM lemmata, and 9,509 *mots nouveaux*, that is, lemmata not previously recorded in lexicography (Selig et al., 2023: 267). The latter represents a significant expansion of Old Occitan vocabulary in both quantitative and qualitative terms (ib. 270). It is obvious that, e.g., significant words of the domain of medicine such as *arteria* f., *arterial* adj., *vẹna* f., etc. are not treated (properly) by DOMél due to its time-frame, concept, and its focus mainly on troubadour lyrics and the pragmatic scripturality of the legal-administrative domain. In DOM 1, 175b ADICIO̲N f.[23], we find several sense definitions of which all but one are attested primarily in one medical text.[24] However, none of the listed senses corresponds to the concept we find in GuiChaulmT.

LEI *LEI digitale* is the digital representation of the LEI, advanced through the benefits typically derived from the digital processing and concerning, e.g., entry editing and publication versioning (Prifti, 2022). In LEI 1, 627 ADDITIO[25], *addizione* f., we find a sense definition compiled of four approximate translations in modern Italian («équivalents (ou gloses traductives)», Möhren, 2015: 408) depicting two concepts, ⌜addition⌝ and ⌜supplement⌝ ("aggiunta, complemento; supplemento, integrazione"). The sense denoted by our *addicion* is not listed.

With *addicion* as a minimal case study and based on the current situations of the dictionaries, we argue that a significant enhancement of all dictionaries in question will be achieved by ALMA's contributions, both with respect to improving existing entries and to

[21] Nota bene: DEAF*él* is currently moving to this address and will be accessible shortly.
[22] And possibly in Middle French (1365) AmphYpL^2 360 (Lafeuille, 1964), see DEAF*pré*: to be examined.
[23] See also online on http://www.dom-en-ligne.de/dom.php?lhid=3f9nCHBVSBMrNwNkGSfoMr.
[24] The lexeme is also attested in the *Leys d'Amor*, a treatise of Toulouse poetry from the middle of the 14thc, and in a document from Auvergne.
[25] See also on https://online.lei-digitale.it/ without an entry-specific URL.

filling gaps in their lexeme inventory. In this case, this will be achieved by conducting a lexical-semantic study of the concept ⌜protuberance of an osseous or cartilaginous structure⌝ represented by Old French *addicion* and possibly its cognates. Beyond ALMA's own publication channels, the new findings can either be indicated by inserting a link into the respective dictionary entries (e.g., into the DEAF*pré* article ADDITION) pointing to the lexical-semantic study published by ALMA at https://alma.hadw-bw.de, or by integrating the research results with respect to each language directly into an extended version of the respective dictionary article.[26]

ALMA foresees lexical-semantic studies for approx. 1,000 lemmata with all senses relevant for the two domains and a large number of graphical realizations in the four languages. Studies will be comprehensive: philological, linguistic, lexicological, lexicographical, and concatenated with the entities of the extralinguistic ontologies. Therefore, they have great potential for new findings and for significant enhancement of the lexicographical resources. A substantial part of the lexeme inventory covered by DEAF*pré* and DOMél will be expanded into valuable and well-researched articles; lexemes, senses, and more data will be added to DAG (DEAG, respectively) and LEI. All will be linked through ALMA.

4.2.2 Modeling as Linked Open Data

The modeling of the lexicographical resources as LOD creates a new publication channel for printed and digitally published works. We envisage the modeling of those dictionaries' articles that are relevant for the lexical-semantic studies of ALMA (cp. Chapter 3.3). However, once data models and modeling workflows have been installed, one could consider extending the modeling to more dictionary articles. Thus, ALMA's contribution could go beyond the scope of its own lexeme inventory, and dictionaries in their entirety could benefit from this approach. Offering the lexicographical data as LOD, the linguistic, textual, and historico-cultural knowledge documented therein will be placed within new contexts and correlations, and the dictionary contents will be introduced to a knowledge circulation wider than that of historical lexicography and linguistics. Naturally, the transfer of the dictionary contents to the new formats also includes linking to the extralinguistic, historicized domain ontologies developed by ALMA, as mentioned above. Through the lexico-semantic mapping to the ontologies, the dictionaries will be extended by an onomasiological-ontological component and the availability of their content will be improved by semantic access options. Publishing the dictionary resources as LOD will allow for their exploitation with the benefits of LOD. This is a significant enhancement of their visibility and re-usability within a global research context independent from their original publication form and place, language, and language stage. Overall, the LOD approach will create a frame-like architecture of historical Semantic Web resources fostering the prospective ALMA and dictionary LOD resources and simultaneously enforcing the historical resources of the LOD landscape.

[26] Following the example of the entries of 'DEAF*pré* - Version révisée', e.g., https://deaf.ub.uni-heidelberg.de/lemme/alcothedem.

5. Conclusion

Ceci n'est pas un dictionnaire, ALMA is *not a real dictionary* but it can be interpreted as a particular—*real*—representation of a dictionary, much like Magritte's pipe[27] but on another abstraction level. The project's scope includes an elaboration of a dictionary in the form of multilingual, concept-driven lexical-semantic studies—a quasi-monography for each concept—that is deeply rooted in long-approved approaches to lexicography. ALMA is less and more than a dictionary at the same time: It is less because it focuses only on a part of the language covered by the comprehensive dictionaries, i.e., on medical and juridical terminology; and it is more because the conceptual entanglement significantly benefits from the combination of re-using well-tried dictionary knowledge, with the addition of new corpus material, integrating machine-driven methods, and introducing a Pan-Romance perspective. This combination allows statements about (1) the extent to which communication spheres in the vernaculars had already been developed for expert cultures and (2) the extent to which these are connected to the Latin-dominated knowledge networks.

A further substantial and expanding aspect is the extralinguistic facet of the studies: the *histoire du concept* next to the *histoire*(*s*) *du mot*. This is not only expressed through textual discussion within the lexico-semantic studies but also through Linked Data modeling and ontology engineering. Two interdependent elements are crucial: (i) the development of hitherto non-existent historicized ontologies for medicine and law and (ii) the lexico-semantic mapping to entities of these and other extra-linguistic knowledge bases of the Semantic Web landscape.

6. Acknowledgements

We would like to thank the anonymous reviewers and Ragini Menon, Maria Selig, and Wolfgang Schweickard (ALMA) for helpful comments and feedback.

7. References

Arnold, R. & Langenbacher-Liebgott, J. (2022–). Diccionario del Español Medieval *electrónico* (DEM*el*). Directed by Rafael Arnold and Jutta Langenbacher-Liebgott on the basis of the Fichero of the Diccionario del español medieval by Bodo Müller (Heidelberg), in collaboration with Anna-Susan Franke, Karsten Labahn, Caroline Müller, Martin Reiter, Stefan Serafin, and Robert Stephan. University of Rostock und University of Paderborn. URL https://demel.uni-rostock.de.

Baldinger, K. (1971–2021). *DEAF.* Dictionnaire étymologique de l'ancien français, *founded by Kurt Baldinger, continued by Frankwalt Möhren and Thomas Städtler.* Québec / Tübingen / Berlin: Presses de L'Université Laval / Niemeyer / De Gruyter. DEAF*él*: https://deaf.ub.uni-heidelberg.de.

Baldinger, K. (1975–2021). *DAG. Dictionnaire onomasiologique de l'ancien gascon*, founded by Kurt Baldinger, directed in collaboration with Inge Popelar, Nicoline Winkler, continued by Martin Glessgen. Tübingen / Berlin: Niemeyer / De Gruyter.

Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. (2013). Towards Open Data for Linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen, L. Qin & E. Hovy (eds.)

[27] https://collections.lacma.org/node/239578.

New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems. Berlin, Heidelberg: Springer, pp. 7–25.

Cimiano, P., McCrae, J. & Buitelaar, P. (2016). Lexicon Model for Ontologies: Community Report, 10 May 2016. URL https://www.w3.org/2016/05/ontolex/.

Coseriu, E. (1980). 'Historische Sprache' und 'Dialekt'. In J. Göschel (ed.) *Dialekt und Dialektologie. Ergebnisse des Internationalen Symposions "Zur Theorie des Dialekts". Marburg/Lahn, 5.–10. Sept. 1977*. Franz Steiner Verlag, pp. 106–122.

Cyganiak, R., Wood, D. & Lanthaler, M. (2004–2014). RDF 1.1. concepts and abstract syntax: W3C recommendation 25 February 2014. URL https://www.w3.org/TR/rdf11-concepts/.

Filatkina, N. (2009). Historische formelhafte Sprache als "harte Nuss" der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt. *Linguistik online*, 39, 3, pp. 75–95.

Glessgen, M.D. (2005). Diskurstraditionen zwischen pragmatischen Vorgaben und sprachlichen Varietäten. Methodische Überlegungen zur historischen Korpuslinguistik. In A. Schrott & H. Völker (eds.) *Historische Pragmatik und historische Varietätenlinguistik in den romanischen Sprachen*. Göttingen: Universitätsverlag, pp. 207–228.

Glessgen, M.D. (2021). Pour une histoire textuelle du gascon médiéval. *Revue de Linguistique Romane*, 85(339-340), pp. 325–384.

Glessgen, M.D. (2023). Les Documents et analyses linguistiques de la Galloromania médiévale (GallRom): structure et potentiel interprétatif. In D. Corbella, Josefa Dorta & Rafael Padrón (eds.) *Perspectives de recherche en linguistique et philologie romanes*. Strasbourg: EliPhi, pp. 1025–1044.

Glessgen, M.D. & Tittel, S. (2018). Le Dictionnaire d'ancien gascon électronique (DAGél). In *Atti del XXVIII Congresso internazionale di linguistica e filologia romanza (Roma, 18-23 luglio 2016)*. Strasbourg: ELiPi, pp. 805–818.

Hirschmann, H. (2019). *Korpuslinguistik. Eine Einführung*. Metzler.

Kabatek, J. (2016). *Lingüística de corpus y lingüística histórica iberorrománica*. Berlin: De Gruyter.

Kjellman, H. (1935). *La Vie seint Edmund le rei*. Elander.

Klein, W. (2004). Vom Wörterbuch zum Digitalen Lexikalischen System. *Zeitschrift für Literaturwissenschaft und Linguistik*, 34, pp. 10–55.

Lafeuille, G. (1964). *Les commentaires de Martin de Saint-Gille sur les Amphorismes Ypocras*. Droz.

Möhren, F. (2015). L'art du glossaire d'édition. In D. Trotter (ed.) *Manuel de la philologie de l'édition*. Berlin: De Gruyter, pp. 397–437.

Möhren, F. (2022). *Complément bibliographique 2021*. Berlin / Boston: De Gruyter Akademie Forschung.

Müller, B. (1994–2005). *Diccionario del español medieval*, volume 1-3. Winter: Heidelberg.

Nannini, A. (forthcoming). La mappatura semantica del Lessico Etimologico Italiano (LEI). Possibilità, metodi e prospettive. In E. Prifti, L. Becker, J. Kuhn, C. Ossenkop & C. Polzin-Haumann (eds.) *Digitale romanistische Sprachwissenschaft. Stand und Perspektiven. Romanistisches Kolloqium XXXIV, Wien, November 2019*. Berlin: De Gruyter.

Oesterreicher, W. (2006). Korpuslinguistik und diachronische Lexikologie. Fallbeispiel aus dem amerikanischen Spanisch des 16. Jahrhunderts. In W. Dietrich, U. Hoinkes, B. Roviró & M. Warnecke (eds.) *Lexikalische Semantik und Korpuslinguistik*. Tübingen: Narr, pp. 479–498.

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

Prifti, E. (2022). Il LEI digitale. Un resoconto, con particolare attenzione alla dialettologia. In M. Cortelazzo, S. Morgana & M. Prada (eds.) *Lessicografia storica dialettale e regionale. Atti del XIV Convegno ASLI (Associazione per la Storia della Lingua Italiana) (Milano, 5-7 novembre 2020).* Firenze: Franco Cesati Editore, pp. 293–314.

Prud'hommeaux, E. & Carothers, G. (2014). RDF 1.1 Turtle: Terse RDF Triple Language. URL: http://www.w3.org/TR/turtle/.

Selig, M., Reichle, E. & Schöffel, M. (2023). New Entries in the *Dictionnaire de l'ancien occitan*: some preliminary remarks on methodological and historical aspects. In N. Pomino, E.M. Remberger & J. Zwink (eds.) *From Formal Linguistic Theory to the Art of Historical Editions. The Multifaceted Dimensions of Romance Linguistics.* Göttingen: Brill / Vandenhoeck & Ruprecht Verlage, pp. 263–280.

Stempel, W.D. (1996–2013). *DOM.* Dictionnaire de l'occitan medieval*, founded by Wolf-Dieter Stempel, continued by Maria Selig.* Berlin [i. a.]: De Gruyter.

Tittel, S. (2004). *Die Anathomie in der Grande Chirurgie des Gui de Chauliac: Wort- und sachgeschichtliche Untersuchungen und Edition.* Tübingen: Niemeyer.

Tittel, S. (accepted). Lexico-Semantic Mapping of a Historical Dictionary: An Automated Approach with DBpedia, *Proceedings of 4th Conference on Language, Data and Knowledge (LDK 2023).*

Tittel, S. (forthcoming). *Integration von historischer lexikalischer Semantik und Ontologien in den Digital Humanities.*

Tittel, S. & Chiarcos, C. (2018). Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the *Dictionnaire étymologique de l'ancien français* with OntoLex-Lemon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). GLOBALEX Workshop (GLOBALEX-2018).* Paris: ELRA, pp. 58–66.

von Wartburg, W. (1922–). *Französisches Etymologisches Wörterbuch (FEW).* Bonn, Heidelberg, Leipzig/Berlin, Basel: ATILF. [Continued by O. Jänicke, C. T. Gossen, J.-P. Chambon, J.-P. Chauveau, and Yan Greub].

Wunderli, P. (2013). *Ferdinand de Saussure: Cours de linguistique générale. Zweisprachige Ausgabe französisch-deutsch mit Einleitung, Anmerkungen und Kommentar.* Tübingen: Narr.