information

MDPI

Article

Multilingual and Multiword Phenomena in a *lemon* Old Occitan Medico-Botanical Lexicon

Andrea Bellandi [1,*], Emiliano Giovannetti [1] and Anja Weingart [2,*]

[1] Institute of Computational Linguistics "A. Zampolli", National Research Council (CNR), 56124 Pisa, Italy; emiliano.giovannetti@ilc.cnr.it

[2] Georg-August-Universität Göttingen, Seminar für Romanische Philologie, 37073 Göttingen, Germany

* Correspondences: andrea.bellandi@ilc.cnr.it (A.B.); aweinga@gwdg.de (A.W.)

Received: 31 January 2018; Accepted: 24 February 2018; Published: 28 February 2018

Abstract: This article illustrates the progresses made in representing a multilingual and multi-alphabetical Old Occitan medico-botanical lexicon in the context of the project Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan (DiTMAO). The chosen lexical model of reference is *lemon*, which has been extended accordingly to some specific linguistic and lexical features of the lexicon. In particular, issues and solutions about the modeling of multilingual and multiword phenomena are discussed, as the way they are managed through LexO, a web editor developed in the context of the project.

Keywords: lemon model; RDF; multilingualism; multiwords; multialphabet; historical lexicon; medico-botanical terminology; Old Occitan; Hebrew; Arabic; LexO web editor

1. Introduction

This article presents the multilingual aspects of the DiTMAO ("Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan" (DiTMAO) is a joint project of Gerrit Bos (Universität zu Köln), Maria Sofia Corradini (Università di Pisa) and Guido Mensching (Georg-August-Universität Göttingen). The project is funded by the Deutsche Forschungsgemeinschaft (DFG) (https://www.uni-goettingen.de/en/487498.html) project, which aims at constructing a resource for Old Occitan medico-botanical terminology. It focuses on a multilingual phenomenon found with multiwords and its representation in *lemon*. The textual basis (The corpus of DiTMAO consists of 18 texts in Latin script, which are mostly books of prescriptions, herbals and books about medical practices, and 11 texts in Hebrew or Arabic script, which are mostly synonym lists, anonymous or contained in medico-botanical books. Each text is represented by up to four manuscripts.) of the DiTMAO lexicon contains several mixed terms, that is multiwords that consist of an Old Occitan term and a term in another ancient language, mostly Hebrew. Before presenting the examples in detail, the particularities of the textual sources, the origin of the multilingual phenomena and the main components of the resource are briefly introduced.

Old Occitan is the medieval stage of Occitan, the autochthonous Romance language spoken in Southern France, and is today a regional minority language with several dialects. During the Middle Ages, the region, as shown in the map below, and its language played a significant role in medical science.

The importance of Old Occitan for medical science was due to the medical schools of Toulouse and Montpellier and the strong presence of Jewish physicians and scholars. For this reason, Old Occitan medico-botanical terminology is documented in texts in Latin, Hebrew and Arabic script; cf. [1–4]. The most important sources for multilingual phenomena are so-called synonym lists and the Hebrew translations of medical texts [5]. The synonym lists in Hebrew script contain many Old Occitan medico-botanical terms with equivalents or explanations in other languages (also spelled in Hebrew characters), mostly in (Judeo-)Arabic, but also in Hebrew, Latin or other Romance languages and sometimes in Greek, Aramaic or Persian. The synonym lists can be considered as a sort of ancient multilingual dictionary. These terms will be included in the DiTMAO lexicon as corresponding terms, because they help to determine the meaning of otherwise opaque Old Occitan terms, as described in [6–9]. Another aspect of medieval writing in vernacular languages is that the terms are documented through numerous variants, expressing different spellings, dialects or historical stages of the language. For this reason, the DiTMAO lexicon includes all variants of Old Occitan terms and the corresponding terms in at least six other ancient languages, together with a translation to modern French and English whenever possible. This multilingual lexicon is the core of the resource that consists of three domains:

- the lexicographic domain, including the lemmatized forms (lemma, variants and corresponding terms in other ancient languages) and their linguistic and lexicographic description;
- the documentation domain, giving the information source of each form of a term and its meaning, as well as a complete bibliography of the sources, editions and dictionaries;
- the conceptual domain, describing the meaning of each term by means of subontologies for the fields of botany, zoology, mineralogy, human anatomy, diseases and therapy (medication, medical instruments).

The DiTMAO resource is conceived of to be accessible to and to be shared by several scientific communities, such as those of Romance and Semitic studies and that of the history of medicine; see [1,4]. In this sense, DiTMAO is part of the current trend to publish linguistic and lexical resources in the context of the Semantic Web, as reported in [10–14]. One of the most important aspects of the publication of datasets in RDF (Resource Description Framework) is the use and re-use of models/vocabularies, which allow the explicit encoding of pertinent aspects of the dataset to be modeled. Indeed, the re-use of models, standards and vocabularies is one of the core best practices underpinning the linked open data publishing paradigm. This means that anyone who wants to publish data as linked open data is strongly encouraged, in the interests of interoperability, to check for the availability of already existing vocabularies that fulfill the modeling requirements of the dataset in question. The *lemon* model has been developed as a standard for publishing lexica as RDF data. More precisely, *lemon* should be considered as an ontology-lexicon model for the multilingual Semantic Web (see [15]), and its nature and purpose perfectly satisfy our needs of representing the DiTMAO lexicon and the relative ontologies. Then, as stated in [16], although the publication of language resources as linked (open) data is being seen as increasingly important within the language resources and the ICT community, a look at the LLOD (Linguistic Linked Open Data) cloud (http://linguistic-lod.org/llod-cloud) reveals that there is still a lack of lexical resources in historical languages.

However, *lemon* has been already adopted (and, when needed, extended) in several initiatives and projects. A multilingual lexicon, called DBnary [17], has been built starting from data extracted from Wiktionary and structured in *lemon*. The Parole-Simple-Clips Italian lexicon has been converted into RDF with *lemon* [10]. Starting from UBY, a lexical-semantic resource for natural language processing (NLP) based on the Lexical Markup Framework, a *lemon* version, called lemonUby, has been developed [18]. In the context of the EuroSentiment project, the *lemon* model has been used to represent language resources for sentiment analysis [19]. *lemon* is used to model linguistic annotations in FrameBase, a linked open and heterogeneous knowledge base representing various sources of structured knowledge [20]. A diachronic extension of *lemon*, called lemonDIA, has been described in [21] to model semantic shifts in a lexicon. More recently, in [22], the authors introduce the PreMOn (PREdicate Model for ONtologies) ontology, a *lemon* extension conceived of to homogeneously represent data from various predicate models. Lastly, a module of *lemon* called LIME (LInguistic MEtadata) has been developed to manage linguistic metadata [23].

In parallel with the definition of the adaptation of *lemon* to the DiTMAO needs, we are working on the development of a web editor for termino-ontological resources, called LexO [24]. As a matter of fact, as it emerged from an analysis of the state-of-the-art, none of the currently available tools for the editing of lexica and ontologies appeared suited to our purpose.

The paper has the following structure. Section 2 describes briefly the extensions to the *lemon* model necessary for the representation of a multilingual and multialphabetical historical lexicon. This section is mainly based on [25] and provides the necessary background for the understanding of the multiword phenomena discussed in the subsequent sections. Sections 3 and 4 provide some solutions in the representation of multiword expressions, with a particular emphasis on sublemmata and collocations, and multi-lexicon phrases, respectively. Section 5 briefly presents LexO, and Section 6 summarizes our experience with the modeling of the lexicon in *lemon*, highlighting its potential and its shortcomings. Finally, Section 7 outlines some conclusions and draws the future steps.

2. Background: Representing Multilingual and Multialphabetical Simple Terms in *lemon*

The multilingual and multialphabetical data of DiTMAO required several extensions to the *lemon* model as discussed in [25]. The extensions will be presented by means of an example, which allows us to illustrate all types of extensions with one lexical entry. The corpus contains the following variants of the word meaning 'hemp' : *canabo*, *canebe*, *canabos* and variants in Hebrew characters (represented here together with the transliterated forms): קנבוש /QNBWŠ, קִנַבוּש /QiNaBWuŠ, קנבונש /QNBWNŠ. The form *canabo* is taken, by definition (The criteria for choosing a lemma are hierarchically: (i) the simple term is chosen over the compound term, e.g., *oli* not *oli rossat*; (ii) the form that corresponds to the lemma in most of the standard dictionaries is chosen, e.g., *bleda* is chosen over *bleta* (the form *bleta* is considered a cultism); (iii) the form that is closer to the etymon is chosen, e.g., *oli* not *holi* due to the Latin etymon < OLEUM; (iv) the most frequent form is chosen.), as lemma or leading variant. The form *canabos* is the plural form of the lemma *canabo*. It is classified as a morphological variant. The form *canebe* differs with respect to spelling and pronunciation. The form is thus classified as a grapho-phonetic variant. As a general definition, the variants in Hebrew characters are all alphabetical variants of the lemma. All forms are plural and marked as morphological variants. The forms קִנַבוּש /QiNaBWuŠ, קנבונש /QNBWNŠ additionally differ with respect to phonology. As indicated by the vowel signs, the initial syllable of קִנַבוּש /QiNaBWuŠ has to be interpreted as <ki>instead of <ka>. The form קנבונש /QNBWNŠ (read: "canabons") contains a so-called *n-mobile*, a particular phonological characteristic of Old Occitan; cf. [1]. Thus, the following aspects have been represented in *lemon*:

> *Type of script* is introduced as a specific property `ditmao:hasAlphabet`, which ranges over `Latn`, `Hebr` or `Arab`values of the `lemon:PropertyValue` class.
>
> *Transliteration*: In order to represent the transliteration, we adopted `lexinfo:transliteration`, which is defined as a sub-property of `lemon:representation`. The specific transliteration alphabets are defined as sub-properties of `lexinfo:transliteration`. In addition to a transliteration of Hebrew, there is the need for a transliteration of Arabic. The former is labeled `HebrTransliteration` and the latter `ArabTransliteration`. `HebrTrsl` and `ArabTrsl` have been created as individuals of the class `lemon:PropertyValue` accordingly.
>
> *Types of variants*: We specify all types of variants as values of `ditmao:variant`, defined as a sub-property of `lemon:property`. This sub-property takes the following values `ditmao:alphabeticalVariant`, `ditmao:graphicVariant`, `ditmao:morphologicalVariant` and `ditmao:graphophoneticVariant` that have been created as individuals of the class `lemon:PropertyValue`.

As mentioned in the Introduction, our corpus contains corresponding terms in other ancient languages, which have been considered as synonyms by the authors of the manuscripts. For example, the variant קנבונש /QNBWNŠ figures as a synonym of the Arabic term קנב /QNB and the Hebrew term קינבס / QYNBS in the synonym lists edited in [2]. The meaning of all three terms is documented as 'hemp' (in particular, *Cannabis sativa* L.). However, even if the terms have exactly the same meaning, they should not be considered as synonyms in the modern understanding of the term, because they do not belong to the same language. For each ancient language, a separate `lemon:Lexicon` has been created and a relation established between terms of two different lexica. The ancient synonym relation has been introduced to model corresponding terms.

> *Corresponding terms*: This relation has been modeled in *lemon* as the property `ditmao:correspondence`, defined as a sub-property of `lemon:senseRelation`. It relates two lexical entries of different lexica.

By defining the above elements, we have been able to represent the lemma *canabo* and the variant קנבונש /QNBWNŠ as follows (Please note that the correspondence and the variant in Hebrew script are קנב and קנבונש , respectively.).

```
:canabo a lemon:LexicalEntry;
 lemon:canonicalForm [lemon:writtenRep "canabo"@aoc;
 ditmao:correspondence lemon:writtenRep "correspondence in Hebrew script" @arab ] .

lemon:otherForm [lemon:writtenRep "variant in Hebrew script" @aoc ;
     ditmao:hasAlphabet ditmao:Hebr ;
     ditmao:HebrTransliteration "QNBWNŠ" ;
     ditmao:variant ditmao:alphabeticalVariant ;
     ditmao:variant ditmao:morphologicalVariant ;
     ditmao:variant ditmao:graphophoneticVariant ].
```

At the date of submission, we have lemmatized 1791 terms with 2912 variants and 807 terms in other ancient languages, constituting 60% of the overall terms. So far, the tool encodes a significant number of entries, as shown in Table 1 (Currently, the languages of the lexicon are represented as string values for the language property. We plan to refer, when possible, to the language codes of the ISO 639, the standardized nomenclature used to classify languages.).

Table 1. Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan (DiTMAO) lexicon statistics at the date of submission: 7671 forms. We plan to refer the French and English translations to existing vocabularies (whenever possible) according to the linked-data principles.

Language	Lexical Entries			Sense Relations		
	Word	Phrase		*Synonym*	*Translation*	*Correspondence*
		Collocation	*Sublemma*			
Old Occitan (main language)	1451	113	227	1424	2587	1225
French	717	294	-	-	1483	-
English	520	214	-	-	1104	-
Hebrew	225	109	-	-	-	379
Arabic	475	157	-	-	-	702
Aramaic	27	5	-	-	-	60
Latin	80	21	6	15	-	84
total	4641 (+ 3030 *other forms*)			5251 (inverse properties are not counted)		

3. Representing Phrases in *lemon*: Sublemmata and Collocations

In [25], we introduced the modeling of multiword expressions in *lemon* focusing on the representation of the internal phrase structure. Using the `lemon:componentList`, we showed how an adjective noun compound can be decomposed into its parts. Each part is related to a lexical entry of the lexicon and to a position in a tree structure. Further, the decomposition function of *lemon* allows for representing so-called mixed terms, consisting of an Old Occitan element and a Hebrew element. These specific cases will be discussed in more detail in the next section. In [25], all multiwords have been defined as sublemmata. As a result, the sublemma relation has been defined as a sub-property of `lemon:lexicalVariant`, which is a formal relation between two lexical entries. This section addresses the lemmatization of multiword expressions, necessary for the understanding of multilingual terms, and shows that the definition of the sublemma relation, as proposed in [25], needs to be revised according to the morphological and semantic properties of the multiword terms.

As a general (lexicographic) criterion, multiword expressions are only considered as sublemmata if their meaning is non-compositional and additionally, in our case, if they can be considered as a technical term. Drawing a line between technical terms and commonly-used terms is not unproblematic, particularly for medieval vernacular terminology. For example, the word *bescueig*, meaning 'cookie', is, in our modern understanding, certainly not a medical technical term. However, it has to be considered as such in the Middle Ages because nutrition was essential in medical treatments, and a cookie can be classified as a form of administration of medicine. Regarding the criteria of non-compositional meaning (The question whether or to what extent the meaning of a compound or a sentence is compositional can not be addressed here. We assume that some version of the principle of compositionality holds for compound terms.), only compound terms with a non-compositional or lexicalized meaning should be part of the lexicon. For example, the meaning of the term *goma arabica* cannot be derived from the meaning of its parts, 'gum' and 'arabic'. However, the meaning of a multiword term like *goma de gingibre* is derivable from its parts, 'gum', 'of' and 'ginger'. Its meaning is thus not lexicalized. From a morphological point of view, the multiword is a syntagmatic compound or collocation. Standardly, collocations are not included into a dictionary. However, as our corpus, in particular the synonym lists, contains many collocations designating mostly pharmaceutical substances, we decided to include these terms, because they have to be considered as medical technical terms. In order to mark this difference, the lexical model has been extended accordingly by the introduction of two new classes: `ditmao:Sublemma` and `ditmao:Collocation`. Sublemma was defined as a subclass of collocation, which, in turn, has been defined as a subclass of the *lemon* class `lemon:Phrase`. The resulting classification of the lexical entries is schematized in Figure 1.

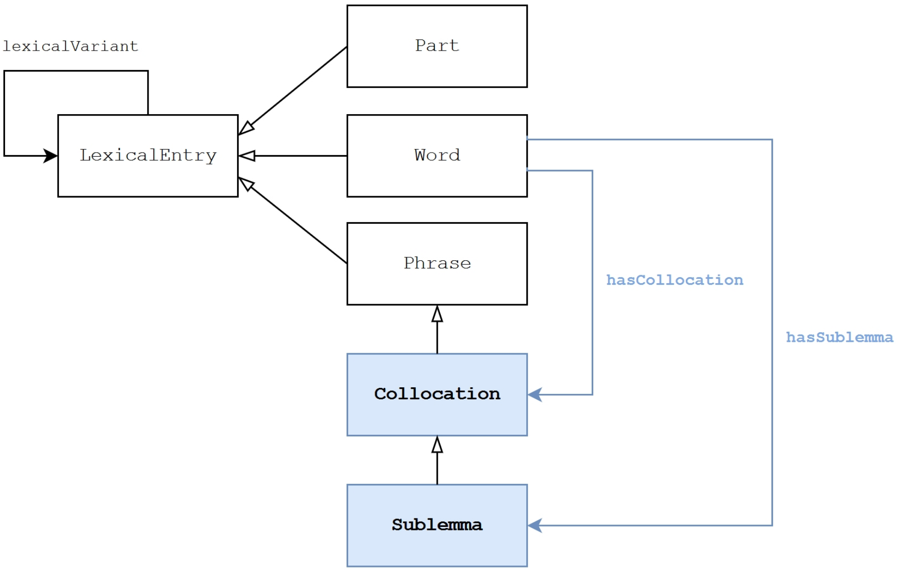


Figure 1. Definition of sublemma and collocation.

A new property, `ditmao:hasSublemma`, has been defined as a sub-property of `lemon:LexicalVariant`, holding between word and sublemma (e.g., *goma* `hasSublemma` *goma arabica*). Similarly, a property `ditmao:hasCollocation` has been defined as a sibling sub-property, holding between `Word` and `Collocation` (e.g., *goma* `hasCollocation` *goma de gingibre*).

Another aspect not discussed in [25] is that the sublemma relation is essentially two-fold. On the one hand it is formal, in the sense that the head noun of the multiword expression determines its lemma, and on the other hand, it expresses a semantic relation between the lemma and sublemma, mostly a hypernym-hyponym relation. In about 80% of the multiword expressions, both criteria coincide. These terms are formally and semantically endocentric. For example, *febre*, meaning 'fever' has the sublemmata *febre cartana*, *febre contunia* and *febre cotidiana*, which designate certain subtypes of fever. However there are cases where the head noun of the multiword expression and its lemma are not directly related semantically. For example, *sanc de dragon*, which is literally translated as 'blood of dragon', is not, in our texts, a kind of blood, but a (red) resin derived from various plants of the Liliaceae family (*Dracaena draco* Willd. or *Dracaena draco* L.). The literal, compositional meaning of the compound is nevertheless available (Similar cases are the terms *cap de monge* with the literal meaning 'head of monk' and a second meaning designating a plant, *Taraxacum officinale* F.H. Wigg., and *lenga de buou* with the literal meaning 'tongue of cow' and a second meaning designating a plant *Anchusa officinalis L.*). As the term *sanc* is the head of the compound and the prepositional phrase (*de dragon*) is its modifier, it will be the corresponding lemma. In order to capture the fact that *sanc* and *sanc de dragon* are related metaphorically, we decided to introduce a metaphorical lexical sense for *sanc*. This sense states something like 'a substances that resembles blood in, e.g., color and consistency'. Doing so, the semantic relation between a sublemma and the lemma may be maintained as a hypernym-hyponym relation (for example, by means of the `skos:narrower` relation). The representation of the formal and semantic aspects of the sublemma relation and the collocation relation in *lemon* is given in Figure 2.

Sanc de dragon is a sublemma of *sanc* and related to the metaphorical sense of *sanc*. The collocation *sanc de fain*, meaning 'blood of a marten', is related to the non-metaphorical sense of *sanc*.

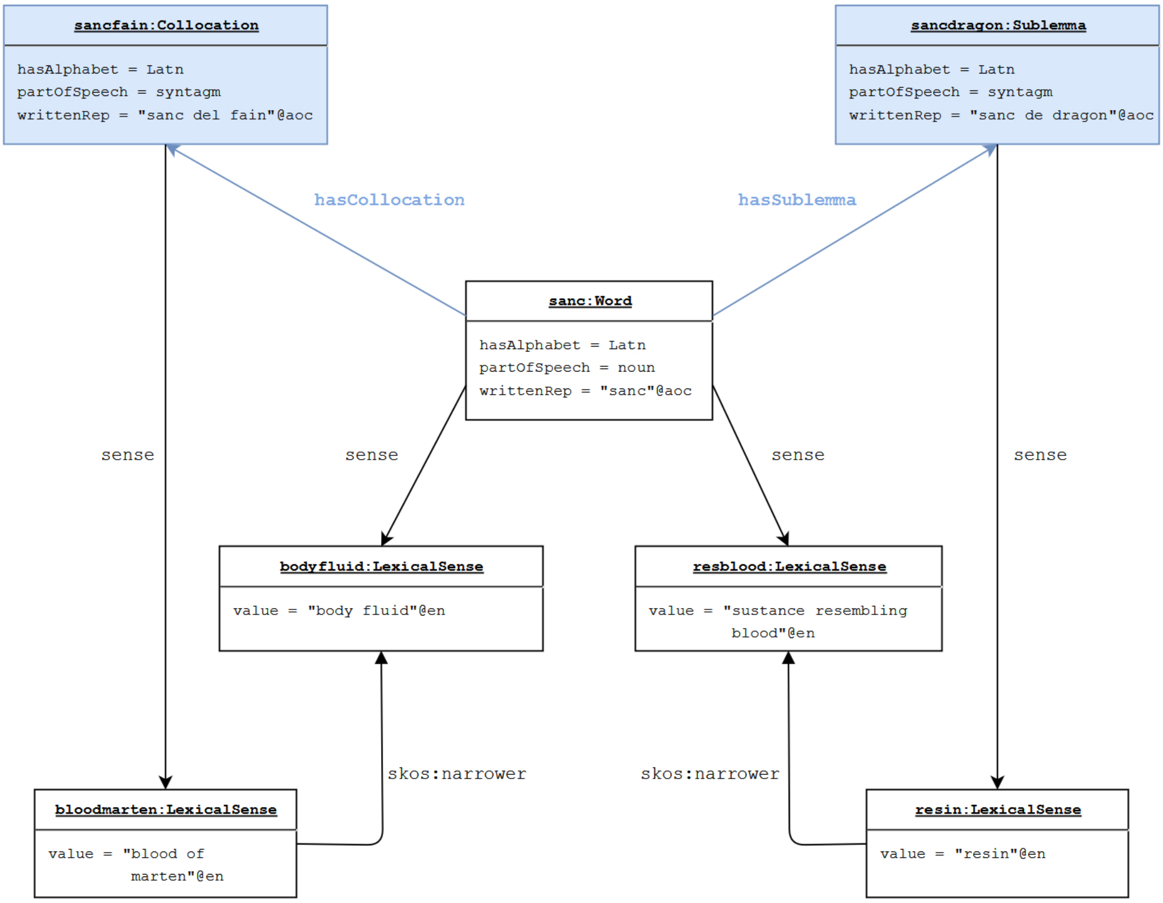


Figure 2. Formal and semantic relations between words, sublemmata and collocations.

4. Multi-Lexicon Phrases

In *lemon*, a lexicon is, by definition, restricted to one language. A challenge for this restriction is the classification of so-called mixed terms. As mentioned above, these multiword terms consist of a Hebrew and an Old Occitan word. For example, the term ארישטולוגיאה ארתכה /’RYŠTWLWGY’H ’RWKH (read: “aristologia aruka”) consists of the Old Occitan term ‘RYŠTWLWGY’H, an alphabetical variant of *aristologia* and the Hebrew adjective ארתכה /’RWKH, meaning ‘long’, which is a Hebrew translation of the Old Occitan adjective *longa*. However, this term is not isolated. The texts contain also the complete Old Occitan forms, here *aristologia longa*, both in Latin and in Hebrew script. In [25], we argued that these terms should be classified as belonging to the Hebrew lexicon, because the terms mostly occur in Hebrew prose texts or in Hebrew translations. This is taken as an indication that the mixed term was part of the technical vocabulary used by Jewish physicians living in Southern France and the Hebrew, and Old Occitan part was most likely transparent for these physicians. Shortcomings of this solution are: (i) the vocabulary of Jewish physicians living in Southern France should not be equated to the Hebrew lexicon; (ii) the Old Occitan part should not be part of the Hebrew lexicon. These terms thus do not belong to either of the two languages, but the *lemon* model requires the language to be unique for a lexicon. The solution we propose is to introduce a bilingual lexicon, here an Old Occitan-Hebrew lexicon, but only for mixed terms that are not incorporated in one of the languages. As shown in Figure 3 below, ארישטולוגיאה ארתכה /’RYŠTWLWGY’H ’RWKH is part of the bilingual lexicon “aocheb”.

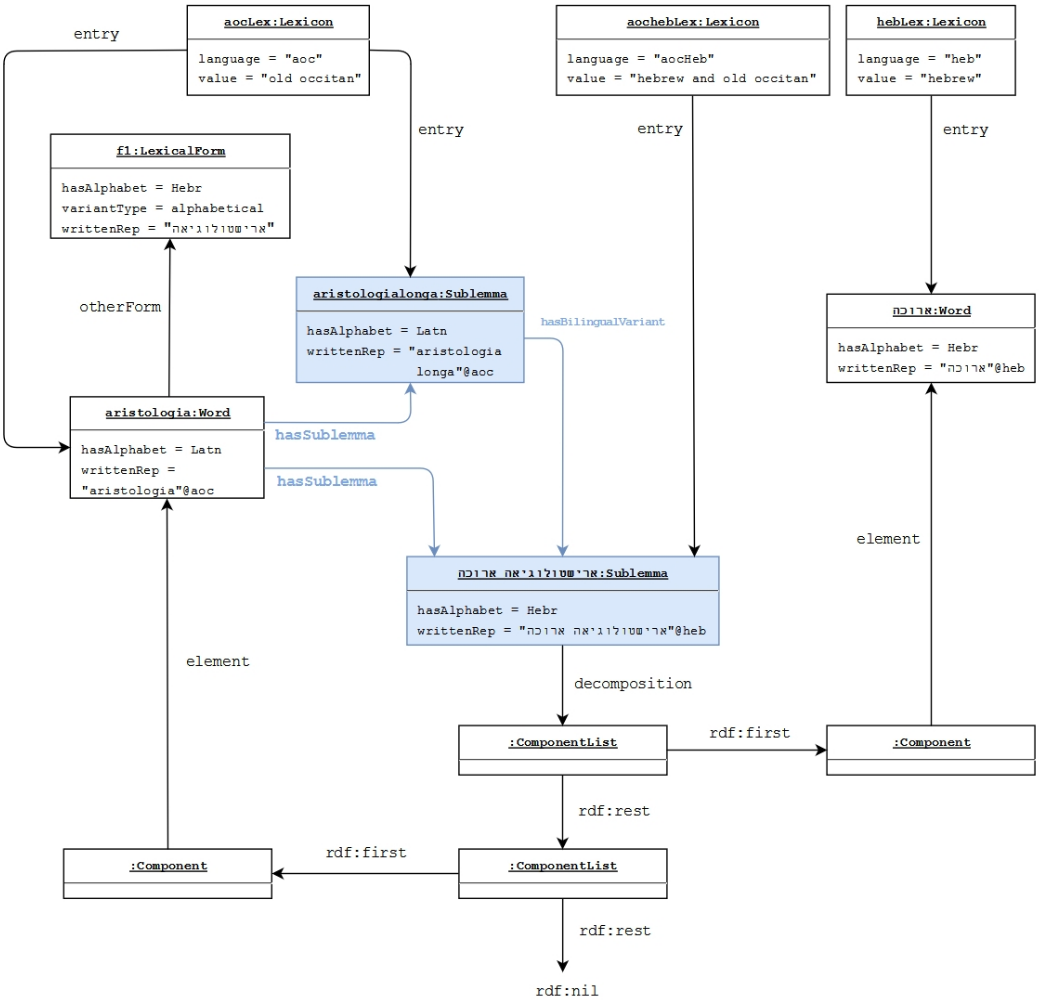


Figure 3. Modeling of bilingual variants.

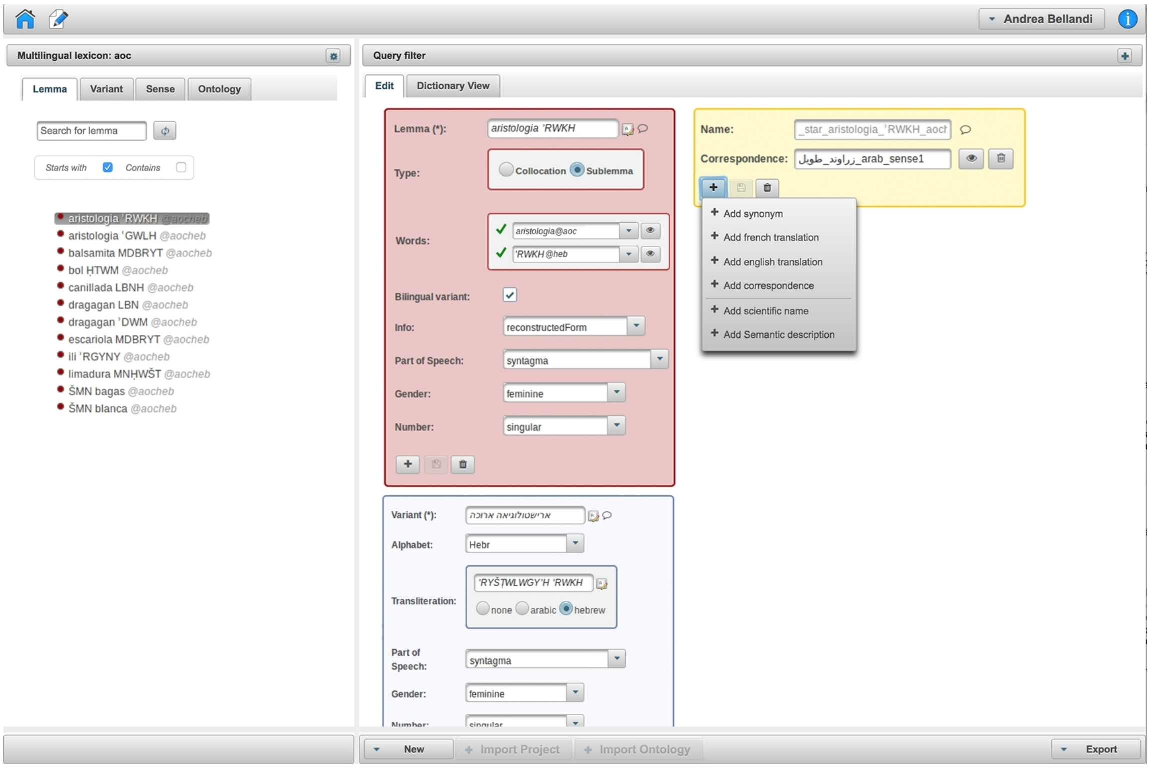
The mixed term is related by means of two relations to the Old Occitan lexicon. First, it is a sublemma of *aristologia*, and it is a bilingual variant of Old Occitan *aristologia longa*. The property `ditmao:hasBilingualVariant` has been introduced as a sub-property of `lemon:LexicalVariant`. By means of the decomposition function provided by *lemon*, the parts of the mixed terms are related to the lemmata of the corresponding lexica: ארתכה /'RWKH to the Hebrew lexicon and ארישטולוגיאה /'RYŠTWLWGY'H to the Old Occitan lexicon. Further, we can state the fact that ארישטולוגיאה /'RYŠTWLWGY'H is an alphabetical variant of *aristologia*. Every part is aligned with the correct lexicon, and we are able to conceive of the mixed term as part of the Old Occitan medico-botanical vocabulary via the `ditmao:sublemma` and `ditmao:hasBilingualVariant` relations without classifying the term as belonging to the Old Occitan lexicon. A related issue concerns other terms like *fumus terre* or *agnus castus* that are morpho-phonologically Latin terms. The term *fumus terre* occurs along with its Old Occitan equivalent *fumtere* in Old Occitan medico-botanical prose texts. This means that the term was commonly known to Old Occitan-speaking physicians and should therefore be considered as part of the Old Occitan medico-botanical vocabulary, but it belongs to the Latin lexicon. The term *agnus castus* has no Old Occitan equivalent and should be considered as a foreign term used in Old Occitan medico-botanical vocabulary, but belonging to the Latin lexicon. These terms show that a distinction between a lexicon defined by morpho-phonological properties and a vocabulary/lexicon that contains foreign or loanwords and that reflects different degrees of incorporation in a certain language is necessary for an accurate representation of a multilingual, historical dictionary. In order to propose a formal representation in *lemon*, further research is necessary.

5. LexO: Work in Progress

In order to support the humanistic partners of the project in the creation of the multilingual lexicon, we developed LexO. Through this editor, the scholars can formalize the lexical knowledge without being familiar with the model and the language underlying the representation.

To date, few attempts have been made to give humanists an easy way to encode a lexicon in *lemon*. In [26], the authors use ontology design patterns [27] for defining how certain lexico-semantic phenomena should be modeled. In [28], a platform called *lemon source* is presented. It supports the creation of linked lexical data, and it builds on the concept of a semantic wiki to enable collaborative editing of the resources. We also cite [29], an editor with custom forms to support the construction of *lemon* lexica. It is an extension of VocBench, a web-based collaborative thesaurus editing and workflow system, natively supporting Semantic Web standards such as RDF, OWL (Web Ontology Language) and SKOS(-XL) (Simple Knowledge Organization System eXtension for Labels).

LexO's interface is composed of two main sections, as shown in Figure 4. On the left, a column shows, depending on the selected tab, the list of lemmas composing the resource, the forms, the lexical senses and the concepts belonging to the ontology (or ontologies) of reference (this part is still in development).



Andrea Bellandi

Multilingual lexicon: aoc

Lemma | Variant | Sense | Ontology

Search for lemma

Starts with ☑ Contains □

- aristologia ʾRWKH @aocheb
- aristologia ʿGWLH @aocheb
- balsamita MDBRYT @aocheb
- bol ḤTWM @aocheb
- canillada LBNH @aocheb
- dragagan LBN @aocheb
- dragagan ʾDWM @aocheb
- escariola MDBRYT @aocheb
- ili ʾRGYNY @aocheb
- limadura MNḤWŠT @aocheb
- ŠMN bagas @aocheb
- ŠMN blanca @aocheb

Query filter

Edit | Dictionary View

Lemma (*): aristologia ʾRWKH

Type: ○ Collocation ● Sublemma

Words: ✓ aristologia@aoc ✓ ʾRWKH@heb

Bilingual variant: ☑

Info: reconstructedForm

Part of Speech: syntagma

Gender: feminine

Number: singular

Variant (*): אריסטולוגיאה ארוכה

Alphabet: Hebr

Transliteration: ʾRYŠṬWLWGYʾH ʾRWKH ○ none ○ arabic ● hebrew

Part of Speech: syntagma

Gender: feminine

Number: singular

Name: _star_aristologia_ʾRWKH_aoch

Correspondence: زراوند_طويل_arab_sense1

Figure 4. LexO's interface. The *aristologia ʾRWKH* multi-lexicon phrase (see Section 4) is composed of the *aristologia* entry, from the old Occitan lexicon, and the *ʾRWKH* entry from the Hebrew lexicon.

If the resource is multilingual, lemmas, forms and senses can be filtered by language. The information related to the selected entry is shown in the central panel where the lemma (red box) appears in the upper part of the leftmost column on top of the relative forms (blue boxes). On the right, the lexical senses are shown (yellow boxes). A user can add both lexico-semantic relations between senses (such as synonymy and translation) and associative relations between lemmas (e.g., sublemma and collocation). For example, Figure 4 shows the details of the multi-lexicon phrase *aristologia ʾRWKH* (aristologia longa).

6. Discussion about the Model

The *lemon* model was originally developed to enrich a given ontology with a lexical layer. However, the conversion of the lexicographic aspects of the Old Occitan medico-botanical lexicon to *lemon* was not always straightforward. Indeed, the model is quite general, and it is meant to be agnostic to the representation of a particular resource. Specific model extensions must be introduced to correctly represent linguistic, historic and scientific facets of a resource.

Understanding what kind of information the model must represent is not simple. However, even if we believe that the needs summarized in Section 2 may be more specific for our use case, we have found some more general phenomena that, in our opinion, the *lemon* model is not able to manage yet. In the following, we describe a series of issues we identified in the encoding of the Old Occitan resource. We distinguish them with NH, which stands for a phenomenon not handled by the model, and with WH, which stands for a phenomenon not handled in a totally correct way.

NH.1: sense-form association is not possible. Due to the multilingual corpus, in particular the synonym lists, there are some cases in which a (alphabetical) variant is associated with an additional meaning. For example גוטא /GWTʾ (read: "gota") is an alphabetical variant of the Old Occitan term *gota*, meaning 'gout'. The variant features the Hebrew and Arabic terms ךפיונ /KYPWN and ץראע /SR'', respectively meaning also 'epilepsy'. In order to represent exactly the given state of documentation, we should be able to relate the variant to an additional meaning not given for other variants and the lemma.

NH.2: sublemma and collocation phrases are not representable. As shown in Section 3, a comprehensive construction of the DiTMAO lexicon needed the inclusion of two novel subtypes of the phrase, namely the sublemma and the collocation; in addition to the description of the parts that compose them (which can be done using *lemon*'s decomposition system), there is the need to specify the formal and semantic relations holding between them and the respective lemmata.

WH.1: phrases' decomposition should involve senses and forms instead of whole lexical entries. The expression *cap de monge*, mentioned in Section 3, designates a plant, and its meaning is non-compositional. However, the simple term *cap* with the meaning 'head (of a human or animal)' is documented. Thus, the decomposition would relate *cap* from a non-compositional plant name to the lexical entry *cap* with the meaning 'head', although the relation between the two occurrences of *cap* is a purely formal one. In this case, the decomposition should not relate to the whole lexical entry, but just to one of its senses.

These issues, not yet considered, for example, in [30], may serve as an input to start a discussion on the integration of more historically-oriented lexicographic aspects in the *lemon* model. In this work, we have chosen to adopt the *lemon* original core model, since it appeared adequate for our purposes. However, the W3C Ontology Lexicon Community Group (https://www.w3.org/community/ontolex/) proposed an evolution of that component, called Ontolex (https://www.w3.org/community/ontolex/wiki/Final_Model_Specification), where some of the limitations of the original model have been overcome. In the near future, we plan to migrate to Ontolex and update the LexO editor accordingly, both to be aligned with the community working on e-lexicography and, at the same time, to be able to exploit all the new features exposed by the new model.

7. Conclusions and Future Work

The DiTMAO project aims at the construction of a resource for Old Occitan medico-botanical terminology. The lexicon is based on a corpus composed of medico-botanical texts in Latin and in Hebrew script. In order to make the resource accessible to and shared with all the scientific communities of reference (such as those of Romance and Semitic studies and that of the history of medicine) as much as possible, we modeled the lexicon according to the linked data paradigm. The chosen lexical model of reference is *lemon*, which has been extended accordingly to some specific linguistic and lexical features of the lexicon. We have shown how the lexicographic phenomena of sublemma and collocation and the relative formal and semantic properties have been modeled within *lemon*. Mixed terms have also been represented with the inclusion of a bilingual lexicon. Extensions like these are necessary to make a terminological resource, such as the DiTMAO lexicon, published and shared on the (Semantic) Web and, at the same time, to make explicit and preserve its many linguistic, historical and scientific facets. From a more lexicological point of view, the next steps of this work will include the modeling of the last lexical phenomena that remain to be represented for DiTMAO, such as loanwords and etymology, taking into account the latest emergent works on these topics, such as [31,32].

However, in order to be useful for an interdisciplinary research community, a term should not only be accessible via the lemmata, but also via the meaning of the terms. Indeed, in onomasiological dictionaries, the terms are grouped according to their meaning and conceptual relations. The *lemon* model, by virtue of the explicit separation of the lexical and conceptual layers, naturally allows a resource to be classified according both to formal, linguistic criteria and according to the semantics of the terms structured in an external ontology. 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). For this concern, we will finish the development of LexO by developing both a module for the management of thesauri/taxonomies and a controlled natural language interface, as in [33], to ease the "onomasiological" access to the lexicon.

The last aspect we will deal with is the documentation of the source texts attesting to each form of a term and its meaning. On the one hand, some models for representing documents in linked data have been already proposed, offering an opportunity for high quality bibliographic data to be exposed to the Semantic Web, such as FRBR (Functional Requirements for Bibliographic Records) [34] or Bibframe (Bibliographic Framework) [35]. On the other hand, a few works about the modeling of the attestation of a term in a document have been done [32,36]. We consider this aspect as crucial, especially in the representation of a historical lexicon with a diachronic dimension.

Acknowledgments: This research has been conducted in the context of the cooperation agreement between Guido Mensching, director of the DiTMAO project at the Seminar für Romanische Philologie of the Georg-August-Universität Göttingen, and the Istituto di Linguistica Computazionale "A. Zampolli" of the Italian National Research Council (29 August 2016).

Author Contributions: Andrea Bellandi, Emiliano Giovannetti and Anja Weingart (together with the DiTMAO partners) conceived and designed the *lemon* extensions and the interface of LexO. Andrea Bellandi implemented LexO.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bos, G.; Corradini, M.S.; Mensching, G. Le DiTMAO (Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan): caractéristiques et organisation des données lexicales. In Proceedings of the XIen Congrès de l'Asociacion Internacionala d'Estudis Occitans (AIEO 2014), Universitat de Lleida, Lleida, Spain, 16–21 June 2014.
2. Bos, G.; Hussein, M.; Mensching, G.; Savelsberg, F. Medical Synonym Lists from Medieval Provence: Shem Tov ben Isaak of Tortosa: Sefer ha-Shimmush. In *Book 29. Part 1: Edition and Commentary of List 1 (Hebrew-Arabic-Romance/Latin)*; Brill: Leiden, The Netherlands, 2011.
3. Corradini, M.S.; Mensching, G. Les méthodologies et les outils pour la rédaction d'un lexique de la terminologie médico-botanique de l'occitan du Moyen Âge. In *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes*; Iliescu, M., Siller-Runggaldier, H., Danler, P., Eds.; Max Niemeyer: Tübingen, Germany, 2010; Volume 6, pp. 87–96. (In French)
4. Corradini, M.S.; Mensching, G. Nuovi aspetti relativi al Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan (DiTMAO): Creazione di una base di dati integrata con organizzazione onomasiologica. In *Actas del XXVI Con-greso Internacional de Lingüística y de Filología Románicas*; Casanova Herrero, E., Calvo Rigual, C., Eds.; Max Niemeyer: Tübingen, Germany, 2013; Volume VIII, pp. 113–124. (In Italian)
5. Mensching, G.; Zwink, J. L'ancien occitan en tant que langage scientifique de la médecine. Termes vernaculaires dans la traduction hébraique du Zad al-musafir waqut alhadir (XIIIe). In *Los Que Fan Viure Etresluire L'occitan (AIEO 2011)*; Alén Garabato, C., Torreilles, C., Verny, M.J., Eds.; Lambert-Lucas: Limoges, France, 2014; pp. 226–236. (In French)
6. Bos, G.; Mensching, G. Arabic-Romance Medico-Botanical Glossaries in Hebrew Manuscripts from the Iberian Peninsula and Italy. In *Aleph*; Indiana University Press: Bloomington, IN, USA, 2015; Volume 15.1, pp. 9–61.
7. Mensching, G. Per la terminologia medico-botanica occitana nei testi ebraici: Le liste di sinonimi di Shem Tov Ben Isaac di Tortosa. In *Atti Del Convegno Internazionale: Giornate Di Studio Di Lessicografia Romanza*; Corradini , M.S., Periñán, B., Eds.; ETS: Pisa, Italy, 2006; pp. 93–109. (In Italian)
8. Mensching, G. Listes de synonymes hébraïques-occitanes du domaine médico-botanique au Moyen Âge. In *La Voix Occitane. Actes Du VIIIe CongrèS Internationale D'ÉTudes Occitanes*; Latry, G., Ed.; Presses Universitaires De Bordeaux: Bordeaux, France, 2009; Volume I, pp. 509–526. (In French)
9. Mensching, G.; Savelsberg, F. Reconstrucció de la terminologia mèdica occitanocatalana dels segles XIII i XIV a través de llistats de sinònims en lletres hebrees. In *Actes Del CongréS Per a L'estudi Dels Jueus en Territori De Llengua Catalana*; Universidad de Barcelona: Barcelona, Spain, 2004; pp. 69–81. (In Catalan)
10. Del Gratta, R.; Frontini, F.; Khan, F.; Monachini, M. Converting the parole simple clips lexicon into RDF with lemon. *Semant. Web* **2015**, *6*, 387–392.

11. Hayashi, Y. Direct and Indirect Linking of Lexical Objects for Evolving Lexical Linked Data. In Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW Volume 775), Bonn, Germany, 23 October 2011; pp. 62–67.
12. Lezcano, L.; Sánchez-Alonso, S.; Roa-Valverde, A.J. A survey on the exchange of linguistic resources: Publishing linguistic linked open data on the web. In *Program*; Emerald Group Publishing Limited: Bingley, UK, 2013; Volume 47, pp. 263–281.
13. McCrae, J.P.; Fellbaum, C.; Cimiano, P. Publishing and linking WordNet using RDF and lemon. In Proceedings of the Third Workshop on Linked Data in Linguistics, Reykjavik, Iceland, 27 May 2014.
14. McCrae, J.; Spohr, D.; Cimiano, P. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In Proceedings of the 8th Extended Semantic Web Conference (ESWC-11), Heraklion, Greece, 29 May–2 June 2011; Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J., Eds.; Springer: Heidelberg, Germany, 2011; pp. 245–259.
15. Declerck, T.; Buitelaar, P.; Wunner, T.; McCrae, J.P.; Montiel-Ponsoda, E.; Aguado de Cea, G. Lemon: An Ontology Lexicon model for the Multilingual Semantic Web. In Proceedings of the W3C Workshop: The Multilingual Web—Where Are We? Madrid, Spain, 26–27 October 2010.
16. Khan, F.; Bellandi, A.; Boschetti, F.; Monachini, M. The Challenges of Converting Legacy Lexical Resources to Linked Open Data using OntoLex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets, Galway, Ireland, 18 June 2017; pp. 1–8.
17. Gilles, S. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semant. Web* **2015**, *6*, 355–361.
18. Eckle-Kohler, J.; McCrae, J.; Chiarcos, C. LemonUby-A large, interlinked, syntactically-rich lexical resource for ontologies. *Semant. Web* **2015**, *6*, 371–378.
19. Sánchez Rada, J.F.; Vulcu, G.; Iglesias Fernandez, C.A.; Buitelaar, P. EUROSENTIMENT: Linked data sentiment analysis. In Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, 19–23 October 2014; pp. 145–148.
20. Rouces, J.; de Melo, G.; Katja, H. FrameBase: Enabling integration of heterogeneous knowledge. *Semant. Web* **2017**, *8*, 817–850.
21. Khan, F.; Federico, B.; Francesca, F. Using lemon to model lexical semantic shift in diachronic lexical resources. In Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language, Reykjavik, Iceland, 27 May 2014.
22. Corcoglioniti, F.; Rospocher, M.; Aprosio, A.P.; Tonelli, S. PreMOn: A Lemon Extension for Exposing Predicate Models as Linked Data. In Proceedings of the LREC-16th International Conference on Language Resources and Evaluation, Portorož, Slovenia, 23–28 May 2016.
23. Fiorelli, M.; Stellato, A.; Mccrae, J.P.; Cimiano, P.; Pazienza, M.T. LIME: The metadata module for OntoLex. In Proceedings of the European Semantic Web Conference, Portorož, Slovenia, 28 May–1 June 2017; Springer: Cham, Switzerlands, 2017; pp. 321–336
24. Bellandi, A.; Giovannetti, E.; Piccini, S.; Weingart, A. Developing LexO: A Collaborative Editor of Multilingual Lexica and Termino-ontological Resources in the Humanities. In Proceedings of the Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017), Montpellier, France, 19–22 September 2017.
25. Weingart, A.; Giovannetti, E. Extending the lemon Model for a Dictionary of Old Occitan Medico-Botanical Terminology. In *The Semantic Web. ESWC 2016. Lecture Notes in Computer Science*; Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C., Eds.; Springer: Cham, Switzerlands, 2016; Volume 9989, pp. 408–421.
26. McCrae, J.P.; Unger, C. Design patterns for engineering the ontology-lexicon interface. In *Towards the Multilingual Semantic Web*; Paul, B., Philipp, C., Eds.; Springer: Berlin, Germany, 2014; pp. 15–30.
27. Gangemi, A. Ontology design patterns for semantic web content. In Proceedings of the International Semantic Web Conference, Galway, Ireland, 6–10 November 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 262–276.
28. McCrae, J.; Montiel-Ponsoda, E.; Cimiano, P. Collaborative semantic editing of linked data lexica. In Proceedings of the 2012 International Conference on Language Resource and Evaluation (2012), Istanbul, Turkey, 21–27 May 2012.

29. Fiorelli, M.; Lorenzetti, T.; Pazienza, M.T.; Stellato, A. Assessing VocBench Custom Forms in Supporting Editing of Lemon Datasets. In Proceedings of the International Conference on Language, Data and Knowledge, Galway, Ireland, 19–20 June 2017; pp. 237–252.
30. Bosque-Gil, J.; Jorge, G.; Elena, M.-P. Towards a Module for Lexicography in OntoLex. In Proceedings of the 1st Workshop on the OntoLex Model (OntoLex-2017), Galway, Ireland, 18 June 2017.
31. Bowers, J.; Laurent, R. Deep Encoding of Etymological Information in TEI. Availible online: http://jtei.revues.org/1643 (accessed on 24 February 2018).
32. Khan, F.; Jack, B.; Francesca, F. Situating Word Senses in their Historical Context with Linked Data. In Proceedings of the IWCS 2017—12th International Conference on Computational Semantics—Short papers, Kissimmee, FL, USA, 8–11 October 2017.
33. Piccini, S.; Bellandi, A.; Benotto, G. Formalizing and Querying a Diachronic Termino-Ontological Resource: The CLAVIUS Case Study. In Proceedings of the From Digitization to Knowledge 2016 Workshop (D2K), Krakow, Poland, 11 July 2016.
34. Carlyle, A. Understanding FRBR as a conceptual model: FRBR and the bibliographic universe. *Bull. Am. Soc. Inf. Sci. Technol.* **2007**, *33*, 12–16.
35. Kroeger, A. The road to BIBFRAME: The evolution of the idea of bibliographic transition into a post-MARC Future. *Cat. Classif. Q.* **2013**, *41*, 873–890.
36. Bellandi, A.; Boschetti, F.; Khan, F.; Del Grosso, A.M.; Monachini, M. Provando e riprovando modelli di dizionario storico digitale: Collegare voci, citazioni, interpretazioni. In Proceedings of the AIUCD 2017 Book of Abstracts, Rome, Italy, 24–28 January 2017. (In Italian)