Extending the Lemon Model for a Dictionary of Old Occitan Medico-Botanical Terminology

Anja Weingart[1(✉)] and Emiliano Giovannetti[2]

[1] Seminar für Romanische Philologie,
Georg-August-Universität Göttingen, Göttingen, Germany
aweinga@gwdg.de
[2] Istituto Di Linguistica Computazionale,
Consiglio Nazionale delle Ricerche, Pisa, Italy
emiliano.giovannetti@ilc.cnr.it

Abstract. The article presents the adaptation of the lemon model (a model for lexica as RDF data) for a multilingual and multi-alphabetical lexicon of Old Occitan medico-botanical terminology. The lexicon is the core component of an ontology-based information system that will be constructed and implemented within the DFG-funded project "Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan" (DiTMAO). The difficulties for the lemmatization raised by the particularities of the corpus (terms in Latin, Hebrew and Arabic script and corresponding terms in other ancient languages, mostly Hebrew and Arabic) can be perfectly solved by extending the basic properties of lemon and introducing domain specific vocabulary.

Keywords: Lemon model · RDF · Multilingual · Multi-alphabetical · Historical lexicon · Medico-Botanical terminology · Old occitan · Hebrew · Arabic

1 Introduction

The project "Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan" (DiTMAO)[1] aims at constructing an ontology-based information system for Old Occitan medico-botanical terminology. The article shows the application of the lemon model[2] to the lexicon component and focuses on the modelling of the historical, multilingual terminology.

1.1 Aims, Background and Structure of the Article

Old Occitan is the medieval stage of Occitan, the autochthonous Romance language spoken in Southern France, today regional minority language with several dialects.

[1] DiTMAO is a joint project of the PIs Gerrit Bos (Universität zu Köln), Andrea Bozzi (Istituto di Linguistica Computazionale "Antonio Zampolli" of the CNR), Maria Sofia Corradini (Università di Pisa) and Guido Mensching (Georg-August-Universität Göttingen). The project is funded by the Deutsche Forschungsgemeinschaft (DFG).

[2] http://lemon-model.net/ (last access: 30/06/2016).

H. Sack et al. (Eds.): ESWC 2016 Satellite Events, LNCS 9989, pp. 408–421, 2016.
DOI: 10.1007/978-3-319-47602-5_53

During the Middle Ages, the region and its language played a significant role in medical science 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 both in Latin and in Hebrew characters (cf. [3]). The DiTMAO project aims at making this terminology accessible to several scientific communities, such as those of Romance and Semitic studies, as well as that of the history of medicine.

The textual basis[3] of the lexicon, as described in [2, 9, 10], consists of medico-botanical texts in Latin and in Hebrew script. Among the sources in Hebrew script, the most prominent text type are so-called synonym lists, which contain a large amount of Old Occitan medical and botanical terms in Hebrew characters with equivalents or explanations in other languages (also spelled in Hebrew characters), mostly in (Judaeo-)Arabic, but also in Hebrew, Latin, or other Romance languages and sometimes in Greek, Aramaic or Persian. These lists can be described as ancient multilingual dictionaries, which are of particular importance for Old Occitan lexicography for two main reasons: (i) the synonym lists of the Jewish tradition include vernacular (Old Occitan) terms already from the 13th century on, hence these lists contain very early testimonies of Old Occitan technical terms. (ii) The corresponding terms in other ancient languages help to determine the meaning of otherwise opaque Old Occitan terms (cf. [3, 18, 19, 21]). A special difficulty of medieval texts in vernacular languages is that most terms are documented in a large number of variants (reflecting different spellings, dialects, or historical stages of the languages at issue). Thus the dictionary will include all variants of Old Occitan terms, together with the corresponding terms in at least six other ancient languages. Whenever possible, also a translation to modern French and English will be provided. The dictionary aims to be useful not only for users interested in Old Occitan but also in reading the numerous Medieval Hebrew medico-botanical texts written or translated in Southern France, since these texts are full of Occitan terminology and thus partially inaccessible even for readers with a good knowledge of Hebrew (cf. [22]).

After introducing the lemon model and our extensions, the article primarily deals with the lemmatization of simple and multiword terms and their representation in lemon. Furthermore, we will show how the corresponding terms in other ancient languages can be integrated and we will propose a way to resolve polysemy[4].

[3] The corpus consists of 11 texts in Latin script, which are mostly books of prescriptions, herbals and books about medical practices, and nine 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. The corpus of DiTMAO combines already edited manuscripts ([7, 8] for texts in Latin script and [3, 4] for texts in Hebrew script. In addition, terms from several unedited manuscripts will be included.

[4] In lemon a lexicon is restricted, by definition, to exactly one language. Besides a lexicon for terms in Old Occitan, labeled ditmao, we define a lexicon for each of the other languages: ditmao_hebrew, ditmao_arabic, ditmao_latin, ditmao_greek, ditmao_aramaic and ditmao_persian.

1.2 The Ontological Conception and the Lemon Model

Current trends in linguistic and lexical resources show a growing interest towards the publishing in the context of the Semantic Web [14–16]. The sharing of lexica in accordance with linked data principles is, nowadays, mandatory: a resource (not only of linguistic nature) that cannot be accessed, shared and reused as a dataset is basically considered unreachable, and, thus, pretty much useless from a semantic web perspective. 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 [11] and its nature and purpose perfectly satisfy our needs of representing the DiTMAO lexicon and the relative ontologies. DiTMAO consists of three main domains: (i) the lexicographic domain, including the lemmatized forms (lemma, variants and corresponding terms in other ancient languages) and their linguistic and lexicographic description. (ii) 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). We aim to complement the onomasiological description, if possible, with a modern scientific classification, for at least most of the plant names, and a medieval classification[5] of plants and other simple drugs. (iii) The documentation domain, giving the source for each form of a term and its meaning. The documentation is indispensable for a historical (diachronic) dictionary.

The lemon model will be extended with a documentation domain and new vocabulary that is necessary for the lemmatization of a historical multilingual and multi-alphabetical dictionary[6].

2 The Lexicographic Component

In the following sections, we describe the lemmatization of simple and multiword terms in Latin and Hebrew script and their representation in lemon. The representation will be illustrated by some representative examples from our corpus. The fact that we use just a few terms should not obscure the fact that our corpus contains about 5800 Old Occitan forms in Latin script and 3200 forms in Hebrew script. Furthermore, the corresponding terms in the other ancient languages amount to 3050 terms.

[5] The medieval classification follows the Galenic system of four basic body humors (blood, yellow bile, black bile and phlegm). The humors are associated with the two primary qualities by cross-combining the pairs HOT–COLD and DRY–WET (cf. [6]) The simple drugs are classified by these quality pairs together with a certain degree of intensity, which varies from one to four (cf. [13]). In order to ensure that the categorization is in conformity with the classification used in medieval Southern France, we will only introduce the classification provided in the texts of our corpus.

[6] The full extension of the lemon model, together with all data (without copyright restrictions) will be published on the project web site: https://www.uni-goettingen.de/en/487498.html (last access: 30/06/2016).

2.1 Lemmatization and Determination of Variants

As a general criterion of lemmatization, it has been decided for DiTMAO that a *lemma* is a term in Latin characters. All forms that differ from the lemma are classified as *variants*. Among the forms in Latin script the lemma is determined following a set of criteria[7] and the form of an Old Occitan lemma is the oblique[8] singular form for nouns, the oblique singular masculine for adjectives, and the infinitive for verbs. For example, the corpus contains the following variants for the word meaning 'hemp seed': *canabo*, *canebe*, *canabos*, and variants in Hebrew characters (represented here together with the transliterated forms[9]): קנבוש/QNBWŠ, קִנַבּוּש/QiNaBWuŠ, קנבונש/QNBWNŠ. The form *canabo* is taken as lemma or leading variant. The form *canabos* is the plural form of the lemma *canabo*. It is classified as morphological variant. The form *canebe* differs with respect to spelling and pronunciation. The form is thus classified as grapho-phonetic variant. As a general definition, the variants in Hebrew characters are all alphabetical variants. The forms קנבוש/QNBWŠ and קִנַבּוּש/QiNaBWuŠ are alphabetical variants of the plural form *canabos*. In this sense they are variants of a variant. The form קִנַבּוּש/QiNaBWuŠ additionally differs with respect to phonology. As indicated by the vowel signs, the initial syllable has to be interpreted as [ki] instead of [ka]. The form קנבונש/QNBWNŠ (read: "canabons") has no corresponding form in Latin script in our corpus. It is thus classified as alphabetical variant of the lemma, and additionally as grapho-phonetic[10] and morphological variant. Furthermore, concerning variants in Latin characters, there are pure graphic variants, where the spelling does not reflect a difference in pronunciation e.g. *alcanna* and *alquana*.

A certain difficulty for lemmatization lies in the fact that about 40 % of the terms are only documented in Hebrew characters. Nevertheless, the general criterion for lemmatization (a lemma is a term in Latin script) has been established for two main reasons. First of all, it is not possible to uniquely link a Hebrew character to a Latin character. For example the letter Alef (א - ʾ) may represent different vowels e.g. it stands for /e/in אשפרמא/ʾŠPRMʾ (read: "esperma", 'sperm'), for /a/in ארמולש/ʾRMWLŠ (read "armols", 'orache'). The combinations of initial Alef with Yod or Waw can be interpreted as /i/or /e/like in אינגילש/ʾYNGYLŠ (read: "enguilas", 'eels') or as /o/o /u/like in אורטיגש/ʾWRṬYGŠ (read "ortigas", 'stinging nettles'). Thus, having lemmata in two alphabets would additionally complicate the string search and the display of the

---

[7] The criteria 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* (< Lat. oleum); (iv) the most frequent form is chosen.

[8] Old Occitan preserved the Vulgar Latin two-case system (nominative vs. oblique case) which was lost by the fourteenth century and the nominative forms have been abandoned in favor of the oblique forms (cf. [3]).

[9] For the transliteration of Hebrew characters, we use the system described in [3, 20]: Alef (א - ʾ), Bet (ב - B), Gimel (ג - G), Gimel ('ג- Ǧ), Dalet (ד - D), He (ה - H), Waw (ו - W), Zayin (ז - Z), Ḥet (ח - Ḥ), Ṭet (ט - Ṭ), Yod (י - Y), Kaf (כ ך - K), Lamed (ל - L), Mem (מ ם - M), Nun (נ ן - N), Samekh (ס - S), Ayin (ע - ʿ), Pe (פ ף - P), Ṣade (צ ץ - Ṣ), Qof (ק - Q), Resh (ר - R), Shin (ש - Š), Tav (ת - T).

[10] The form קנבונש/QNBWNŠ contains a so-called *n-mobile*, a particular phonological characteristic of Old Occitan (cf. [3]).

results in alphabetical order. In case a term is only documented in Hebrew characters, a corpus-external lemma, a form documented in other dictionaries, will be included. But in some cases, there is no such corpus-external lemma (so the variant in Hebrew spelling is the only documented form), and we have to introduce a hypothetical or reconstructed form. For example for the term אנאקירד - ʾNʾQYRD (read "anacard"), we introduce the form **anacard* as hypothetical Old Occitan form with the meaning 'marking nut', fruit of Semecarpus anacardium L. . The meaning is documented for the Arabic term בלאדר/BLʾDR that features as its synonym in the lists edited in [4]. Thus, we need to indicate for a lexical entry whether the lemma is corpus-external, a reconstructed or a hypothetical from.

2.2 Modelling the Lemma and Its Variants

A lexicon entry in lemon consists of a `Form` and a `LexicalSense`. For the lemmatization, the class `Form` and its relations with `LexicalEntry` (`lexicalForm` and its subproperties `canononicalForm` and `otherForm`) are relevant. In lemon the lemma *canabo* will have the following shape:

```
:canabo a lemon:LexicalEntry;
lemon:canonicalForm [lemon:writtenRep "canabo" @aoc-Latn;
    lexinfo:partOfSpeech lexinfo:noun ;
    lexinfo:gender lexinfo:masculine ;
    lexinfo:number lexinfo:singular ] .
```

The lemma is represented by the `canonicalForm` of the entry and its realization is the written representation (`writtenRep`). The language, although inferable from the lexicon, will be represented together with the ISO 15924 script code: `Latn` for Latin, `Arab` for Arabic, and `Hebr` for Hebrew. This is an elegant way to avoid the definition of a property specifying the script type. The linguistic information like part of speech, gender and number will be integrated as attribute-value pairs from the Lexinfo ontology[11], an extension of lemon that provides data categories for linguistic annotations. These will be defined as subproperties of the property `lemon:property`. In a similar vein, the labels for corpus-external lemmata and hypothetical and reconstructed forms can be added to the `canonicalForm`.

```
ditmao:lemmaInfo rdfs:subPropertyOf lemon:property .
```

The subproperty `ditmao:lemmaInfo` will have the following values: `ditmao:corpusExternalLemma`, `ditmao:hypotheticalForm` and `ditmao:reconstructedForm`. For the representation of variants, the lemon model only provides the relation `otherForm`. The variant *canabos* has the following entry:

```
lemon:otherForm [lemon:writtenRep "canabos" @aoc-Latn ;
lexinfo:number lexinfo:plural] .
```

[11] http://www.lexinfo.net/ontology/2.0/lexinfo.owl (last access: 30/06/2016).

The fact that *canabos* is a morphological variant can be inferred from the value of `lexinfo:number`. An alphabetical variant can be formalized by adding a script tag to the language tag e.g. `aoc`[12]`-Hebr` or `aoc-Arab`. In order to give the transliteration, we adopted `lexinfo:transliteration` which is defined as a subproperty of `lemon:representation` (the superproperty of `lemon:writtenRep`), in accordance to the Lemon Cookbook [17]. The specific transliteration alphabets are defined as subproperties of `lexinfo:transliteration`. For the DiTMAO, a transliteration of Hebrew and Arabic is needed. The former is labelled `HebrTransliteration` and the latter `ArabTransliteration` with the respective abbreviations `HebrTrsl` and `ArabTrsl`.[13] The entry for קנבונש/ QNBWNŠ (read "canabons") would have the following shape.

```
lemon:otherForm [lemon:writtenRep "קנבונש" @aoc-Hebr ;
ditmao:HebrTransliteration "QNBWNŠ" @aoc-HebrTrsl ;
lexinfo:number lexinfo:plural] .

lexinfo:transliteration  rdfs:subPropertyOf
lemon:representation .

ditmao:HebrTransliteration rdfs:subPropertyOf
lexinfo:transliteration .
```

A problem is the formalization of the graphic and grapho-phonetic variants. Only users who are familiar with Old Occitan phonology and dialectology may distinguish graphic from grapho-phonetic variants. But as the dictionary also wants to reach researchers from other domains, an indication of these types of variants is desired. We propose to specify all types of variants (morphological, alphabetical, grapho-phonetic and graphic variants) as values of `ditmao:variant`, defined as a subproperty of `lemon:property`. This subproperty will take the following values: `ditmao:alphabeticalVariant`, `ditmao:graphicVariant`, `ditmao:morphologicalVariant`, and `ditmao:graphophoneticVariant`. The form *canebe* bears only the value `ditmao:graphophoneticVariant`. Additionally to the marking of the script and grammatical number, the entry קנבונש/QNBWNŠ has the following shape:

[12] For Old Occitan, ISO proposes the language tag *pro*, which is derived from the term Provençal. But Provençal, like Gascon, Limousin, Languedocian, and Auvergnat, has to be considered a dialect of Old Occitan (cf. [1]). Thus, we take the name Old Occitan (French Ancien Occitan) to be the correct hyperonym and define a new language tag *aoc* for DiTMAO.

[13] The alternative option is to label the transliteration alphabet as Latin script, but this would not be correct, because the transliteration alphabets contain special phonetic symbols e.g. the symbols ʾ and ʿ (replacing Alef and Ayin, respectively).

```
lemon:otherForm [lemon:writtenRep "קנבונש" @aoc-Hebr ;
  ditmao:HebrTransliteration "QNBWNŠ" @aoc-HebrTrsl ;
  lexinfo:number lexinfo:plural;
  ditmao:variant ditmao:alphabeticalVariant;
  ditmao:variant ditmao:morphologicalVariant;
  ditmao:variant ditmao:graphophoneticVariant ]
```

The other variants in Hebrew characters have been classified as variants of a variant. The terms קנבוש/QNBWŠ and קִנַבוּש/QiNaBWuŠ are alphabetical variants of the morphological variant *canabos*. In order to represent a relation between two forms of one lexical entry, lemon provides the property `formVariant`. A symmetric subproperty of `formVariant`, `ditmao:varOfVar`, will be defined:

```
ditmao:varOfVar rdfs:subPropertyOf lemon:formVariant .
```

The subproperty `ditmao:varOfVar` will be added to the variant in Hebrew characters. An exemplary entry is shown below for the form קִנַבוּש/QiNaBWuŠ.

```
lemon:otherForm :canabos ;
lemon:otherForm : קִנַבוּש ;

:canabos [lemon:writtenRep "canabos" @aoc-Latn;
  lexinfo:number lexinfo:plural;
  ditmao:variant ditmao:graphophoneticVariant ] .

: קִנַבוּש [lemon:writtenRep "קִנַבוּש" @aoc-Hebr;
  ditmao:HebrTransliteration "QiNaBWuŠ" @aoc-HebrTrsl ;
  lexinfo:number lexinfo:plural ;
  ditmao:varOfVar :canabos ;
  ditmao:variant ditmao:graphophoneticVariant ;
  ditmao:variant ditmao:alphabeticalVariant ] .
```

2.3 Modeling Multiword Expressions

The multiword expressions contained in our corpus are mostly noun-adjective expressions, like *goma arabica*, ‘arabic gum’ or syntagmatic noun-preposition-noun expressions, like *goma de gingibre*, ‘ginger gum’. Multiword terms are classified as sublemma in the sense of a strict alphabetical macrostructure of a dictionary. Both nouns, *goma arabica* and *gomma de ginibre*, are sublemmata of the lemma *goma*. Sublemmata are modeled as a relation between two lexical entries by means of the property `LexicalVariant`. For DiTMAO, a sub-property of `LexicalVariant`, `sublemmaOf`, will be defined. The entry of the term *goma arabica* will have the following entry:

```
:goma_arabica a lemon:LexicalEntry;
  ditmao:sublemmaOf :goma .
```

For a description of the internal structure of multiword expressions, lemon provides a phrase structure module. Multiword terms can be decomposed into their components by means of an ordered list, the `lemon:componentList`. A list consists of components, which are linked by means of the property `lemon:element` to the lexical entries. Each component can be associated to a leaf of a tree structure, representing the internal structure of the phrases *goma arabica* and *goma de gingibre*. The determinatum *goma* is the head of the noun phrase and the determinans is the adjective phrase or the prepositional phrase, which are themselves decomposed into an adjective and a preposition + noun phrase. Each component is linked to its lemma, which is unproblematic for the noun *goma de gingibre*, because the components correspond to the canonical form of the lemmata at issue. However, the term *arabica* is inflected for feminine and the canonical form of an adjective is, per definition, the masculine singular form. For relating such components, we cannot use the `lemon:element` property since it is defined to have the class `LexicalEntry` as range. For this reason, we chose to define a specific property, whose range is the `lemon:Form`:

```
ditmao:formElement rdfs:subPropertyOf lemon:property .
```

The decomposition of *goma arabica* is shown in Fig. 1.

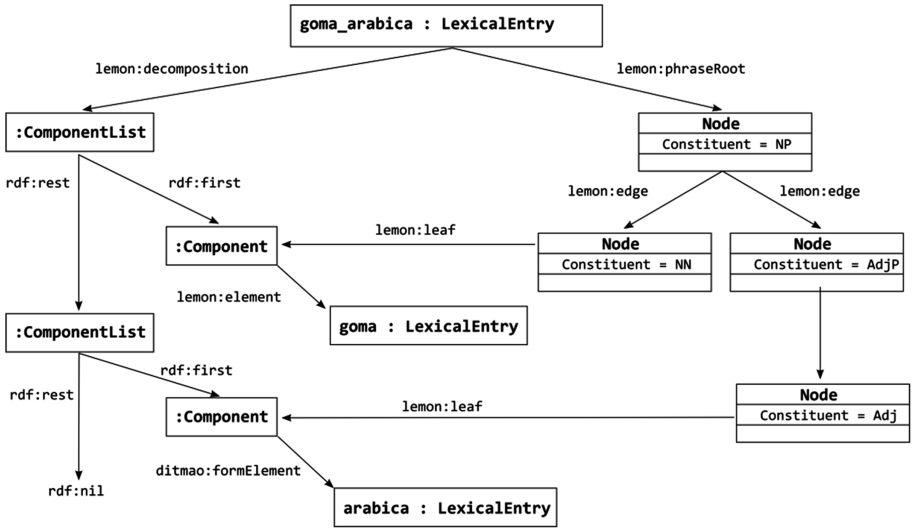

Fig. 1. Decompositon of *goma arabica*

A particularity of our corpus is multiword expressions, consisting of an Old Occitan and a Hebrew word e.g. בול חתום/ BWL ḤTWM meaning 'sealed clay/earth' and אגוז מושקאדא/ʾGWZ MWŠQʾDʾ, meaning 'nutmeg'. The former consists of an Old Occitan head noun, בול/BWL, an alphabetical variant of the term *bol*, followed by a Hebrew participle passive *ḥatum*. The latter has a Hebrew head noun אגוז/ʾGWZ, meaning 'nut', followed by an alphabetical variant of the Old Occitan adjective *muscada*. These mixed terms mostly occur in Hebrew prose texts or in Hebrew translations

and should be considered as foreign technical terms of Jewish physicians living in the Southern France. As for the lemmatization, the terms are taken to be lexical entries of the ditmao_hebrew lexicon, irrespective of the language of the head noun. Due to the decomposition function of lemon, we can preserve the information that the components בול/BWL and מושקאדא/MWŠQʾDʾ are variants of the Old Occitan terms *bol* and *muscat*[14], respectively. How these terms can be represented in lemon will be discussed in the following subsection.

The term אגוז מושקאדא/ʾGWZ MWŠQʾDʾ is a sublemma of the Hebrew entry אגוז/ʾGWZ, meaning 'nut'. The adjective מושקאדא/MWŠQʾDʾ is the alphabetical variant of the feminine form *muscada*, hence a variant of a variant.

```
:ditmao_hebrew lemon:entry :אגוז_מושקאדא
ditmao:sublemmaOf :אגוז
:אגוז_מושקאדא lemon:canonicalForm [lemon:writtenRep "אגוז
מושקאדא" @heb-Hebr ;
    ditmao:HebrTransliteration "ʾGWZ MWŠQʾDʾ" @hebr-
HebrTrsl ].
```

In order to decompose the term, the relations lemon:element and ditmao:formElement are needed, because the head noun corresponds to a lemma of the ditmao_hebrew lexicon and the adjective is a variant.

```
lemon:decomposition ( [lemon:element :אגוז ]
        [ditmao:formElement :מושקאדא ]  ) .
```

The representations for the terms אגוז/ʾGWZ and מושקאדא/MWŠQʾDʾ have the following shape.

```
:ditmao_hebrew lemon:entry :אגוז
:אגוז lemon:canonicalForm [lemon:writtenRep "אגוז" @heb-Hebr
;
ditmao:HebrTransliteration "ʾGWZ" @heb-HebrTrsl ].

:ditmao lemon:otherForm :מושקאדא
:  מושקאדא lemon:otherForm [lemon:writtenRep        @aoc-Heb
;
ditmao:HebrTransliteration "מושקאדא" @aoc-HebrTrsl ;
  ditmao:varOfVar :muscada ;
  ditmao:variant ditmao:alphabeticalVariant ] .
```

As for the term בול חתום/BWL ḤTWM, which consists of an Old Occitan head noun and a Hebrew participle passive, lemmatization is more problematic. In order to define a sublemma relation we would need to assume, contrary to fact, that the simple term בול/BWL was a Hebrew medical term. An equally undesired solution would be to allow

[14] מושקאדא / MWŠQʾDʾ is an alphabetical variant of the feminine singular form *muscada*. The lemma of adjectives is per definition the the masculin singular form, here *muscat*.

the sublemma relation to be valid across the lexica. Thus, multiword terms with an Old Occitan head noun will not be lemmatized with respect to the sublemma relation, but the information that the word בול/BWL is an alphabetical variant of the Old Occitan term *bol* may be preserved, due to the decomposition, as shown below.

```
:ditmao_hebrew lemon:entry :בול_חתום
:חתום_לובם lemon:canonicalForm [lemon:writtenRep "בול_חתום"
@heb-Hebr ;
ditmao:HebrTransliteration "BWL ḤTWM" @hebr-HebrTrsl ].

lemon:decomposition ( [lemon:element :בול ]
[ditmao:formElement :חתום ]  ) .

:ditmao lemon:otherForm :בול
:בול lemon:otherForm [lemon:writtenRep "בול" @aoc-Hebr
ditmao:HebrTransliteration "BWL" @aoc-HebrTrsl;
ditmao:varOfVar :bol ; ditmao:variant
ditmao:alphabeticalVariant  ] .
```

Further we preserve the information that the word חתום/ḤTWM is a morphological variant of the lemma חתם/ḤTM

```
:ditmao_hebrew lemon:otherForm :חתום
:חתום lemon:otherForm [lemon:writtenRep "חתום" @hebr-Hebr
;
ditmao:HebrTransliteration "ḤTWM"@hebr-HebrTrsl;
ditmao:variant ditmao:morpholgicalVariant ].
```

In some cases a mixed term is documented in a synonym list together with a term in Old Occitan. The mixed term will be classified as corresponding term, in the same way as simple terms or other monolingual multiword expressions. E.g. the term אגוז מושקאדא /ʾGWZ MWŠQʾDʾ appears together with the Old Occitan term נוץ מושקאדא/NWṢ MWŠQʾDʾ (read: “noz muscada”) and the Arabic term גוז בוי/GWZ BWY (read: “ğawz bawwā”). The mixed term and the Arabic term will be linked as correspondence to the sublemma in Latin script: *noze moscada*. How these corresponding terms are modeled in lemon will be discussed in the next section.

2.4 Corresponding Terms and Other Sense Relations

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 term ליטוגא/LYṬWGʾ (a variant of *laytugua*) figures as synonym of the Aramaic term חסא/ḤSʾ and the Arabic term כס/KS in the synonym lists

edited in [3]. The meaning of all three terms is documented[15] as 'lettuce' (in particular Lactuca sativa L.). But 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 (cf. [5]). In order to model this relation in lemon, we propose the property `ditmao:correspondence`, as a subproperty of `senseRelation`. It links the senses of two lexical entries that belong to distinct lexica of ancient languages. In order to give a corresponding term in modern French and modern English, the subproperty `lemon:translationOf` will be used. The relations have to be kept apart for mainly two reasons: corresponding terms and translations belong to different historical stages and to different registers. The former are medieval technical terms and the latter are modern common names. Furthermore, the corpus contains Old Occitan terms that are synonyms in the modern understanding of the term, e.g. the terms *litargia* and *mal de dormir* have the meaning: 'fatigue'. The corresponding `LexicalSense` of both terms is linked via the subproperty `lemon:equivalent`. The relations are represented in Fig. 2.

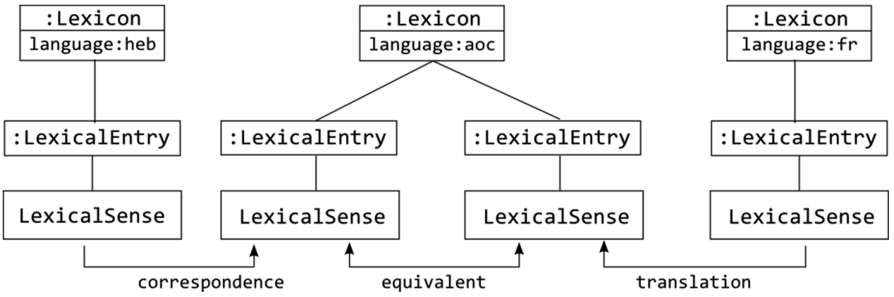

Fig. 2. Relating lexical senses in DiTMAO

But about 20 % of the lemmata in our corpus have more than one meaning. For example, we often find polysemic plant names which designate several species of a genus, e.g. the term *laureola* is documented with the names for the species Daphne oleoides Schreb., Daphne gnidium L., and Daphne sericea Vahl. In lemon, polysemy will be formalized as follows: a `LexicalEntry` has several instances of `LexicalSense`. The Arabic and Hebrew corresponding terms that feature in the synonym lists, give an additional meaning: Daphne mezereum L. The entry of *laureola* has four instances of `LexicalSense`. Each `LexicalSense` has a translation into modern French and English and the `LexicalSense` referring to Daphne mezereum L. will be linked via `ditmao:correspondence` to the respective Arabic and Hebrew entries. Furthermore, each `LexicalSense` of *laureola* has a referent in the botanical branch of the ontology, giving a general description of the plant e.g. that it is a kind of shrub. These entities are linked to the modern classification, here the binominal plant names,

[15] For a complete documentation see pp. 225/226 of [3].

and to a medieval classification. The term *laureola* is described as HOT and DRY in the third degree (see [12] and fn. 5). The general division of the conceptual subontology into an onomasiological subontology, a medieval and a modern classification system allows us to provide a description of the term´s concepts independently from a modern or a medieval classification. This division is necessary for terms that designate e.g. medical instruments or substances whose composition is uncertain.

3 Conclusion and Outlook

We have shown how the lemon model can be adapted to the needs of historical lexicography, by defining subproperties of the basic lemon properties: `lemon:senseRelation`, `lemon:formVariant`, `lemon:element` and `lemon:property`. Furthermore, we introduced our own, domain specific, vocabulary for the description of form variants. In the spirit of lemon, and, in general, of the Semantic Web, we plan to link the dictionaries to other resources. However, at the moment the most important resource related to Old Occitan (i.e. DOM[16]) is a database and it's not exposed as a linked data. Among the resources we are planning to use to provide the conceptual references of lexical senses we cite DBpedia[17], Wikidata[18] and more domain-specific datasets, such as TDWG[19] or the Biological Taxonomy Vocabulary[20].

To ease the process of modelling of the various lexica in lemon and the construction of the ontologies of reference, we are also working on a web editor. As a matter of fact, none of the currently available tools for the editing of lexica and ontologies appears suited to our purpose. Protégé[21], probably the most used tool for the construction of ontological resources, is general enough to allow the building of lemon resources. However, the process can be quite tedious, requiring the manual construction of instances of entries, senses, forms and relations among them. In addition, it is a stand-alone tool which cannot be used collaboratively by a team of users (its Web version[22] has several limitations, as the lack of support for reasoning mechanisms and plug-in extensions). We also plan to develop a controlled natural language querying interface to ease the access to the resources.

[16] Dictionnaire de l´occitan médiéval - http://www.dom.badw.de/indexde.htm (last access: 30/06/2016).

[17] http://wiki.dbpedia.org/ (last access: 30/06/2016).

[18] http://www.wikidata.org (last access: 30/06/2016).

[19] http://rs.tdwg.org/dwc/rdf/dwctermshistory.rdf (last access: 30/06/2016).

[20] http://lov.okfn.org/dataset/lov/vocabs/biol (last access: 30/06/2016).

[21] http://protege.stanford.edu/ (last access: 30/06/2016).

[22] http://webprotege.stanford.edu (last access: 30/06/2016).

References

1. Bolduc, M.: Occitan Studies. In: Classen, A. (ed.) Handbook of medieval studies: terms, methods trends, vol. 2, pp. 1023–1038. De Gruyter, Berlin (2010)
2. Bos, G., Corradini, M.S., Mensching, G.: Le DiTMAO (Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan): caractères et organisation des données lexicales. In: Proceedings of the XIen Congrès de l'Asociacion Internacionala d'Estudis Occitans (AIEO) (in press)
3. Bos, G., Hussein, M., Mensching, G., Savelsberg, F.: Medical synonym lists from Medieval provence: Shem Tov ben Isaak of Tortosa: Sefer ha-Shimmush. Book 29, Part1: Edition and Commentary of List 1 (Hebrew-Arabic-Romance/Latin). Brill, Leiden (2011)
4. Bos, G., Kley, J., Mensching, G., Savelsberg, F.: Medical synonym lists from medieval provence: Shem Tov ben Isaak of Tortosa: Sefer ha-Shimmush. Book 29, Part2: Edition and Commentary of List 2 (Romance/Latin-Arabic-Hebrew). Brill, Leiden (in prep)
5. Bos, G., Mensching, G.: Arabic-romance medico-botanical glossaries in hebrew manuscripts from the Iberian Peninsula and Italy. In: Aleph, vol. 15.1, pp. 9–61. Indiana University Press (2015)
6. Bruchhausen, W., Schott, H.: Geschichte, Theorie und Ethik der Medizin. Vandenhoeck/Ruprecht, Göttingen (2008)
7. Corradini, M.S.: Ricettari medico-farmaceutici medievali nella Francia meridionale. Olschki, Florence (1997)
8. Corradini, M.S.: Per l'edizione del corpus delle opere mediche in occitanico e in catalano: nuovo bilancia della tradizione manoscritta e analisi linguistica dei testi. In: Rivista di Studi Testuali, vol III, pp. 127–195. Università di Torino (2001)
9. 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: Iliescu, M., Siller-Runggaldier, H., Danler, P. (eds.) Actes du XXVe Congrès International de Linguistique et de Philologie Romanes, vol. 6, pp. 87–96. Max Niemeyer, Tübingen (2010)
10. 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: Herrero, E.C., Rigual, C.C. (eds.) Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas, vol. VIII, pp. 113–124. Max Niemeyer, Tübingen (2013)
11. Declerck, T., Buitelaar, P., Wunner, T., McCrae, J.P., Montiel-Ponsoda, E., de Cea, G.A.: Lemon: an ontology-lexicon model for the multilingual semantic web. In: W3C Workshop: The Multilingual Web - Where Are We? Madrid, 26/10/2010–27/10/2010
12. González, A.G.: Alphita. Edición crítica y comentario, SISMEL-Edizioni del Galluzzo, Florence (2007)
13. Gerabek, W.: Enzyklopädie Medizingeschichte. Walter de Gruyter, Berlin (2005)
14. 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), vol. 775, pp. 62–67. Bonn (2011)
15. Hellmann, S., McCrae, J.P., Del Gratta, R., Frontini, F., Khan, F., Monachini, M.: Converting the parole simple clips lexicon into RDF with lemon. Semant. Web **6**(4), 387–392 (2015)
16. 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. Program **47**(3), 263–281 (2013)

17. McCrae, J.P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Pérez, A.G., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Lemon Cookbook. http://lemon-model.net/lemon-cookbook/index.html
18. Mensching, G.: Per la terminologia medico-botanica occitana nei testi ebraici: le liste di sinonimi di Shem Tov Ben Isaac di Tortosa. In: Corradini, M.S., Periñán, B. (eds.) Atti del convegno internazionale: Giornate di studio di lessicografia romanza, pp. 93–109. ETS, Pisa (2006)
19. Mensching, G.: Listes de synonymes hébraïques-occitanes du domaine médico-botanique au Moyen Âge. In: Latry, G. (ed.) La voix occitane. Actes du VIIIe Congrès Internationale d'Études Occitanes, vol. I, pp. 509–526. Presses universitaires de Bordeaux, Bordeaux (2009)
20. Mensching, G.: Éléments lexicaux et textes occitans en caractères hébreux ». In: Trotter, D. (ed.) Manuel de la philologie de l'édition, pp. 237–264. De Gruyter, Berlin (2015)
21. 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, pp. 69–81. Universidad de Barcelona (2004). http://www.institutmonjuic.googlepages.com/2.ACTESPDF.pdf
22. 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 wa-qut al-hadir (XIIIe). In: Garabato, C.A., Torreilles, C., Verny, M.-J. (eds.) Los que fan viure e tresluire l'occitan (AIEO 2011), pp. 226–236. Lambert-Lucas, Limoges (2014)