Challenges for the Representation of Morphology in Ontology Lexicons

Bettina Klimek[1], John P. McCrae[2], Julia Bosque-Gil[3], Maxim Ionov[4], James K. Tauber[5], Christian Chiarcos[4]

[1] Institute for Applied Informatics (InfAI), Leipzig University
[2] Data Science Institute, National University of Ireland Galway
[3] Ontology Engineering Group, Universidad Politécnica de Madrid
[4] Goethe-Universität Frankfurt am Main
[5] Open Greek and Latin Project

Abstract

Recent years have experienced a growing trend in the publication of language resources as Linguistic Linked Data (LLD) to enhance their discovery, reuse and the interoperability of tools that consume language data. To this aim, the OntoLex-*lemon* model has emerged as a *de facto* standard to represent lexical data on the Web. However, traditional dictionaries contain a considerable amount of morphological information which is not straightforwardly representable as LLD within the current model. In order to fill this gap a new Morphology Module of OntoLex-*lemon* is currently being developed. This paper presents the results of this model as on-going work as well as the underlying challenges that emerged during the module development. Based on the MMoOn Core ontology, it aims to account for a wide range of morphological information, ranging from endings to derive whole paradigms to the decomposition and generation of lexical entries which is in compliance to other OntoLex-*lemon* modules and facilitates the encoding of complex morphological data in ontology lexicons.

Keywords: morphology; RDF; OntoLex-*lemon*; MmoOn; inflection; derivation

1. Introduction

Morphology is a vital and, in many languages, very sophisticated part of language, and as such it has been an important part of the work of lexicographers. In the traditional print form, morphological information is provided in brief abbreviated terms that can only be deciphered with significant knowledge of the language, however with the transformation of the dictionary to an electronic resource a re-imagining of the morphology information in a dictionary is certainly due. We base our work within the framework of the ontology-lexicon (McCrae et al., 2012; Cimiano et al., 2014) and in particular in that of the OntoLex-*lemon* model. This model has been used not only for the conversion of existing dictionaries (Khan et al., 2017; Borin et al., 2014; Bosque-Gil et al., 2015) but also for the development of new dictionaries (Gracia et al., 2017) as Linked Data (Chiarcos et al., 2013).

In this paper, we present the current modelling as well as the underlying challenges within the development of the Morphology Module for OntoLex-*lemon*, which extends the existing work by providing modelling for representing the morphology that is associated with the entries. In many cases, morphology is an important part of the language, for example in both German and Irish noun plurals are irregular and cannot be predicted from the stem alone, so many dictionaries, especially learners' dictionaries, list these irregular forms for most or all of the entries. Further, for languages such as the Romance ones, verbs may have many forms that are frequently irregularly or semi-irregularly derived, and learners' dictionaries for these languages also list many forms. However, as electronic dictionaries become of use not only to humans but also machines, it is necessary to provide all forms in a manner that can be readily processed by the latter. To this end, the Morphology Module covers not only the description of some forms of a lemma, but also allows the generation of all forms through morphological patterns, which corresponds to the idea of declensions or conjugations of an entry. Further, we base our model on the MMoOn Core ontology (Klimek, 2017), which has been designed to more generally represent morphology as a linguistic domain, and as such this module can handle a wide range of linguistic phenomena including distinctions between derivational and inflectional morphology, allomorphy, suppletion, simulfixes and transfixes among others. Moreover, this module is, as its name suggests, part of the overall model of OntoLex-*lemon* and as such can be integrated well with other parts of OntoLex-*lemon* and is consistent with its other semantic and syntactic modules.

The rest of this paper is structured as follows. In Section 2, we provide an example based illustration of the shortcomings of morphological data representation in traditional dictionaries. In Section 3 we provide background of the OntoLex-*lemon* model for readers, who are not familiar with it, which is followed by an overview of related work in Section 4. We then present the challenges of representing morphology within the OntoLex-*lemon* framework in Section 5 before presenting the current modelling state of our proposed model in Section 6. Finally we look into the further improvements that we plan for the module in Section 7, and present some conclusions in Section 8.

2. Morphological data in dictionaries and lexical databases

The treatment of morphology in dictionaries is a complex topic which is related to the lexicographic selection process (or lemma selection) (Schierholz, 2015), and the definition of the micro-structure of entries, i.e., the data model upon which the description (Hartmann, 2001) and layout (Atkins & Rundell, 2008) of each entry will be based, with different types or 'templates' being also considered, e.g. a typical noun-entry type (Abel, 2012).

Opacity, frequency and predictability of form and meaning in words were aspects that had to be considered when deciding whether a complex lexeme or compound word should be contained in a dictionary or not (De Caluwe & Taeldeman, 2003), but

dictionaries and lexicographic traditions, in general, vary substantially. For example, derivational affixes have often received main entry status, with differences from dictionary to dictionary in their description: from dictionaries that identify them just as suffixes, to dictionaries that also point to their derivational or inflectional use (Alsina & DeCesaris, 1998).

Different approaches to lexicography also play a role in these various representations of morphological data. Linguistics-oriented dictionaries, guided by a linguistic theory for morphology and its terms, contrast with function-theoretic based (or communicative) works which are focused mainly on the morphological information needs of users in specific situations (Swanepoel, 2015; Bergenholtz & Tarp, 2005).

This context leads to a heterogeneous landscape when it comes to analysing the morphological description provided in dictionaries. Most traditional dictionaries do not cover morphological information extensively: usually, the morphological description of the lexical entry is limited to the list of the word forms that allow users to identify the morphological pattern to which the entry adheres, and hence generate the paradigm by themselves. Following this, word-forms that can be formed regularly are not listed. Moreover, the description of these 'reduced' inflection lists is often minimal on the assumption of users being familiar with the lexicographic tradition of the object language. For example, users of a German dictionary familiar with the German language easily interpret the description ***Na · me*** *der; -ns, -n* to refer to the gender of the entry, and its genitive singular and nominative plural endings. Other dictionaries, such as The K Dictionaries Multilingual Global Series[1], provide groups of word-forms inflected for case and number, along with the ending that is displayed in the user interface, as illustrated in Example 1.1.

This is similarly the case for Ancient Greek dictionaries, where noun entries will typically list the nominative singular form, the genitive singular ending, and the article (indicating the gender). This assumes the reader is able to work out the stem by comparing the nominative form with the abbreviated genitive ending. This, in combination with the gender, is then generally enough to produce other forms of the nominal paradigm. Additional forms of the noun are generally not given in the entry unless deemed impossible or non-obvious to produce from the standard information given.

For verbs it also very common to find verbal paradigms as a reference in the appendix of dictionaries. For example, Figure 1 shows the paradigm of the verb *amar* 'to love' as an example of a verb that inflects according to the 1st conjugation pattern in Spanish[2]. Even though such tables contain all forms of a lemma, the underlying morphological

[1] https://www.lexicala.com/resources#dictionaries

[2] http://www.rae.es/diccionario-panhispanico-de-dudas/apendices/modelos-de-conjugacion-verbal#advertencias, last accessed on 05.06.2019.

structure separating the stems from the regular and productive inflectional suffixes remains again implicit.

```
<HeadwordBlock>
  <HeadwordCtn>
    <Headword>Stipendiat</Headword> [...]
    <GrammaticalGender value="masculine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiaten</Inflection>
        <Display>-en</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <HeadwordCtn>
    <Headword>Stipendiatin</Headword> [...]
    <GrammaticalGender value="feminine" />
    <InflectionBlock>
      <InflectionCtn>
        <Inflection>Stipendiatin</Inflection>
        <Display>-</Display>
      </InflectionCtn>
      <InflectionCtn>
        <Inflection>Stipendiatinnen</Inflection
        >
        <Display>-nen</Display>
      </InflectionCtn>
    </InflectionBlock>
  </HeadwordCtn>
  <PartOfSpeech value="noun" />
</HeadwordBlock>
```

Example 1.1: An extract of the entry *Stipendiat* ‘scholarship holder’ from the K Dictionaries Global Series German Dictionary.

1. AMAR		Verbo modelo de la 1.ª conjugación		
INDICATIVO				
TIEMPOS SIMPLES				
presente	**pret. imperfecto / copretérito**	**pret. perfecto simple / pretérito**	**futuro simple / futuro**	**condicional simple / pospretérito**
amo	amaba	amé	amaré	amaría
amas (amás)	amabas	amaste	amarás	amarías
ama	amaba	amó	amará	amaría
amamos	amábamos	amamos	amaremos	amaríamos
amáis	amabais	amasteis	amaréis	amaríais
aman	amaban	amaron	amarán	amarían
TIEMPOS COMPUESTOS				
pret. perfecto compuesto / antepresente	**pret. pluscuamperfecto / antecopretérito**	**pret. anterior / antepretérito**	**futuro compuesto / antefuturo**	**condicional compuesto / antepospretérito**
he amado	había amado	hube amado	habré amado	habría amado
has amado	habías amado	hubiste amado	habrás amado	habrías amado
ha amado	había amado	hubo amado	habrá amado	habría amado
hemos amado	habíamos amado	hubimos amado	habremos amado	habríamos amado
habéis amado	habíais amado	hubisteis amado	habréis amado	habríais amado
han amado	habían amado	hubieron amado	habrán amado	habrían amado

Figure 1: Table of the inflectional paradigm of the verb *amar* 'to love' from the *Diccionario Panhispánico de Dudas* (Real Academia Española and Asociación de Academias de la Lengua Española, 2005).

From the examples just illustrated, it becomes clear that all the common approaches regarding the representation of morphological data rely highly on the implicit knowledge of the dictionary user about the language. As a consequence, morphological data varies greatly concerning their amount, their way of representation and interconnection to the relevant element they are contained in, i.e. the lemma or a form in a paradigm.

3. Overview of OntoLex-*lemon*

The OntoLex-*lemon* model[3] has been under development for several years and was originally based on the combination of the three pre-existing models (LingInfo (Buitelaar et al., 2006), LexOnto (Cimiano et al., 2007), LIR (Montiel-Ponsoda et al., 2011)) that were combined into a single model (lemon) by the EU project Monnet and later extended into the OntoLex-*lemon* model by the Ontology Lexicon Community

[3] The full specification can be consulted here: https://www.w3.org/2016/05/ontolex/.

Group[4]. This model was developed around five basic principles: 1) it would be an RDF model that used the Web Ontology Language (OWL) (McGuinness, Van Harmelen, et al., 2004) for its semantics; 2) it would support multilinguality and avoid language-specific assumptions that might affect the applicability of the model to other languages; 3) it would use the principle of 'semantics by reference' as a basic semantic model (Cimiano et al., 2013); 4) it would embrace openness in being free of any financial costs or licensing as well as allowing contributions from any interested party, and 5) relevant standards and models would be reused wherever appropriate. This led to the core model that is depicted in Figure 2, which is based around a lexical entry, composed of a number of forms and a number of senses, which can then be linked to either lexical concepts or entities in an ontology.

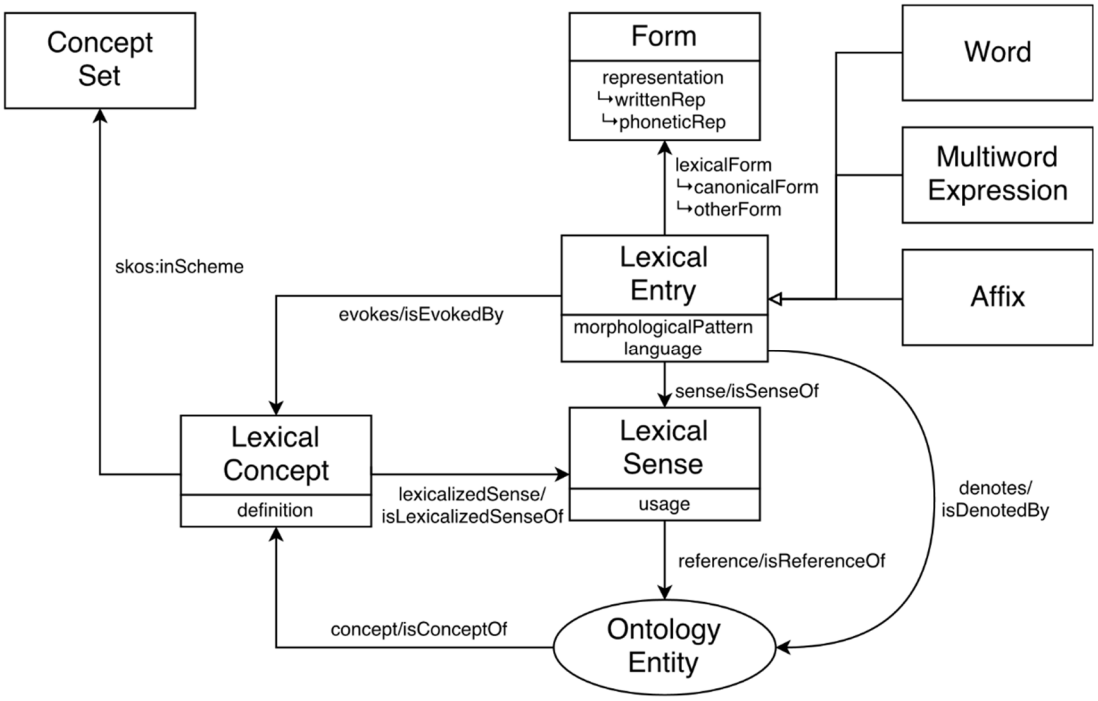

Figure 2: The core model of OntoLex-*lemon*.

In addition to this core, that is often also called "ontolex", there were four further modules developed in the initial release of the model:

Syntax and Semantics (synsem) This module describes how syntactic frames may be modelled and how they can be mapped to ontology structures,

Decomposition (decomp) The decomposition of multiword expressions and compound terms is described by this module,

Variation and Translation (vartrans) Modelling of translations and other kinds of

[4] https://www.w3.org/community/ontolex/

relations are provided by this module,

Linguistic Metadata (lime) This module provides metadata about the lexicon and the ontology and how this may be used to encourage interoperability between resources.

In addition, since then the group has continued to develop modules to extend the usefulness and applications of the model. One such extension, the recently released Lexicography Module (Bosque-Gil et al., 2017), has provided features for representing dictionaries in ways that are more compatible with traditional print dictionary forms. Other modules are in development, in particular this one along with a module for representing frequencies, attestations and corpus information[5], and a module for etymological and diachronic information (Khan, 2018).

Since its development, the OntoLex-*lemon* model has been extensively used for representing a vast amount of different lexical data: In addition to traditional dictionary data mentioned in Section 1, it has been applied to lexical databases like WordNet (McCrae et al., 2014), etymological resources (Chiarcos et al., 2016; Khan, 2018), and domain-specific lexicons (Bellandi et al., 2018).

4. Related work

The emerging OntoLex-*lemon* Morphology Module described in this paper aims to enable the representation of the morphological elements and processes that are involved in the decomposition and generation of lexical data (of both lexemes and their word-forms) by overcoming the representational limitations of traditional dictionaries as outlined in Section 2 and within the technical realm and the design principles of the overall OntoLex*lemon* model introduced in the previous section. Since the emergence of the (multilingual) Semantic Web in the early 2000s, several ontologies emerged from the lexicography, language resource and language documentation communities that already contain the modelling of morphological language data to some extent. Here we briefly describe some of these ontologies that are considered the most relevant with regard to the morphological data they allow to represent, together with an explanation to what extent they could or why they could not be reused within the OntoLex-*lemon* Morphology Module.

In the early development of the OntoLex-*lemon* model, its priorities have been on lexicalizing ontologies and knowledge bases. This was accompanied by a natural focus on lexical semantics, i.e., multilingual labels for the same concept, and, here, the original contribution of Monnet-Lemon, the predecessor of OntoLex-*lemon* has been to complement such labels with morphosyntactic information in order to facilitate context-adequate lexicalization. Morphology was only considered in the form of morphosyntax, i.e. inflectional features as well as the possibility to provide the adequate form for these.

[5] https://acoli-repo.github.io/ontolex-frac/

The current OntoLex-*lemon* representation of morphological information can complement ontology concepts with morphosyntactic categories (part of speech, a property of a lexical entry), and provide different forms with different morphosyntactic features (e.g., gender, case, number, etc.) Neither derivational morphology nor morphological information beyond the specification of grammatical features was expressible with this model, and lexicalizations of the same concept with different parts of speech required independent lexical entries, without being able to represent the systematic relations on the level of form and meaning that hold between them.

OntoLex-*lemon* does not provide any vocabulary of grammatical features, instead, it endorses the reuse of the existing ontologies and vocabularies for linguistic annotations, most notably, ISOcat, GOLD, OLiA, and LexInfo. ISOcat, a shared repository for linguistic concepts, features and data structures, was developed as a successor of the ISO Data Category Registry (DCR), originally designed as an RDF-based knowledge graph (Ide & Romary, 2004) and is built on XML technologies and resolvable URIs (Kemps-Snijders et al., 2009). ISOcat was a semistructured resource populated in a bottom-up process, so that it did not provide formal and consistent vocabulary, but its subsets became an important source of knowledge that more consolidated domain vocabularies described here drew from. GOLD, one of the first attempts in creating a linguistic ontology (Farrar & Langendoen, 2003), and OLiA (Chiarcos & Sukhareva, 2015) were designed primarily as solutions to harmonize linguistic categories and make markup schemes interoperable. In OLiA this is achieved by linking the hierarchy of abstract grammatical categories which constitutes the reference model with specific markup schemas that can vary for resources and languages.

Despite their interoperability and applicability to a vast amount of linguistic data, these ontologies are primarily focused on providing labels for the categories and lack the expressibility to represent morphosyntactic information.

LexInfo is an inventory containing various types, values and properties to describe linguistic categories (Cimiano et al., 2011). It is partially derived from ISOcat and is often used to represent linguistic annotations in Ontolex-*lemon* (however, this is not a requirement). Even though it covers certain aspects of morphology, it has a focus on inflectional morphology whereas it lacks expressiveness in describing derivational morphology.

Finally, the last relevant model is the MMoOn Core ontology[6] (Klimek et al., 2016). It is currently the only existing comprehensive domain ontology for the linguistic area of morphological language data. As such it is highly specialized and far more-fine grained than the desired modelling of the OntoLex-*lemon* Morphology Module requires. It contains, among other aspects, an extensive modelling of linguistic meanings, including derivational meanings in addition to grammatical categories. It also differentiates

[6] https://mmoon.org/core

between morph and morpheme resources and comes with a set of nearly 300 morphemic glosses to provide sufficient expressivity to represent morphological data contained in Flex or Toolbox datasets. At the same time, a specification of lexical data is not provided in MMoOn Core because this ontology was envisaged to be used complementary to OntoLex-*lemon*. Therefore, there is only one existing interconnection of the two domain ontologies so far, i.e. an established subclass relation between the two classes mmoon:LexicalEntry and ontolex:LexicalEntry. A more extensive ontology alignment has been thus far only proposed from the MMoOn Core perspective (Klimek, 2017) and might be considered for future implementation. Once the OntoLex-*lemon* Morphology Module will be officially released, further alignment options might be realized. Even though the MMoOn Core ontology exceeds by far the modelling needs of the Morphology Module, it served as a modelling template since the creation of MMoOn Core was initially motivated to fill the gap of representing morphological language data in OntoLex-*lemon* that still existed back then. So far, certain types of affix classes, e.g. mmoon:Simulfix) as well as the two object properties mmoon:consistsOf and mmoon:meaning have been reused in the OntoLex-*lemon* module, although only in an inspirational manner. These classes and properties are defined and integrated slightly differently within the morphology module and should not be confused as long as no explicit alignment has been implemented.

From this review of relevant existing ontologies it can be concluded that the emerging OntoLex-*lemon* morphology module adheres to the Semantic Web best practice of reusing existing vocabularies. Since none of the presented ontologies sufficiently satisfies the representation needs of morphological data in particular with regard to lexical data so far, the Morphology Module will adequately fill this gap. Furthermore, as a result of the outlined reuse choices, the Morphology Module could be kept user-friendly and manageable by replacing the usually necessary modelling of grammatical categories and morphological meanings of morph resources with the recommendation to use existing vocabularies instead, and also linguistically accurate because it is influenced by the more precise MMoOn Core domain ontology.

5. Challenges in developing a Morphology Module extension

Creating a descriptive modelling foundation for representing lexical data entails several design choices that directly affect the usability of the model. This does not only hold for ontology lexicons, but also for lexicon models in general. In what follows, challenges that arose during the development of the morphology module for OntoLex*lemon* will be outlined. With the ongoing development of modules, these issues gain increasing importance and can serve as orientation points of consideration for future module extension development efforts.

5.1 Scope and coverage

Description: The first question that arises when a new ontology is being created is who should use it for what purpose? As illustrated in Section 2, morphological information is highly implicit in the landscape of traditional dictionaries. However, along with the liberation from the limits of print dictionaries came almost unlimited possibilities of lexicographic data compilation in eLexicography, which are yet again broadened by the possibilities of the Linked Data paradigm. While some lexicographers only like to digitize a printed dictionary into Linked Data using RDF, others aim at transforming their already more fine-grained lexical databases and intend to use the resulting RDF dataset to generate more lexicographic content out of it, e.g. to generate inflectional paradigms including full word-forms together with the underlying morpho-phonological formation rules.

Modelling Choice: In line with OntoLex-*lemon* model, the Morphology Module also aims at being applicable for everyone working with lexicographic content who either focuses on the transformation of traditional dictionary data into RDF or on the conversion of more structured computational lexical data. Accordingly, the scope of the module is divided into two main parts: 1) enabling the representation of elements that are involved in the decomposition of lexical entries and word-forms, and 2) enabling the representation of building patterns that are involved in the formation of lexical entries and word-forms. A fine-grained description of phonological processes that are involved in any kind of stem or word formation on the phoneme level is, however, excluded and not representable with this Morphology Module. Only the elements between the lexical entry and the morph levels will be covered.

5.2 Consistency

Description: The ontolex and decomp modules of OntoLex-*lemon* already contain various classes and properties that can be used to describe morphological data. The ontolex:Affix and decomp:Component classes for instance already exist to represent sub-word units and can be put into relation to the lexical entries in which they are contained via properties like decomp:correspondsTo or decomp:subterm. Due to the widespread usage of OntoLex-*lemon*, the development of the Morphology Module is challenged with creating the necessary missing vocabulary by taking the existing classes and properties into account, while ensuring backwards compatibility at the same time.

Modelling Choice: Due to the incremental approach of developing the module for morphology and also future OntoLex-*lemon* extensions, it is inevitable to deal with overlapping existent vocabulary. Therefore, the OntoLex Community Group agreed to aim for the goal of reaching consistency by reusing as much of the existent vocabulary as possible and minimize duplication that results from creating similar classes and properties. Specifically, this entails that suitable existent vocabulary can be adapted as

long as the changes made are a) only additions to domain and range restrictions of properties or b) adaptions in the rdfs:comment description to broaden the applicability of classes. In this way, existing vocabulary can be coherently integrated into later developed modules while simultaneously preserving already established functionalities.

5.3 Terminological ambiguity

Description: During the module development process it turned out that one of the greatest challenges is to unambiguously define the terminology that is used to label the classes and properties of the new vocabulary. As intended, the widely set scope of the Morphology Module presented in Section 5.1 attracts the use of the module for various user groups which are, however, also coming from different terminological backgrounds. The understanding and usage of linguistic concepts like *morph* or *root* diverge considerably depending on whether the user of the module is, for example, a traditional linguist, a computer linguist or a lexicographer managing data for specific languages. This entails a high risk of an inappropriate usage of the ontological vocabulary that might result in an unintentional wrong data representation the user is generally not even aware of.

Modelling Choice: While the human-readable definition of ontology elements is defined within the rdfs:comment, the underlying machine-processable semantics are determined by implications and restrictions for an element and its relation to other elements of the ontology. For the computational processing of the data the former is not relevant, whereas the latter is formally fixed and unambiguous. What matters is the consistent usage of the vocabulary according to the ontologically defined semantics, notwithstanding that a user would have chosen a different label for an element. Moreover, providing a definition that is interpreted in the same way by all users is almost impossible. Therefore, the rdfs:comment descriptions of classes and properties are discussed and refined until the highest possible consensus is reached. In addition to that, the Morphology Module specification that will be published together with the release of the module contains usage examples and recommendations that support a shared understanding to ensure the consistent application of the module vocabulary.

6. Current state of the Morphology Module

6.1 Summary of the current state

The development of the Morphology Module is an ongoing joint effort by members of the OntoLex Community Group that started in November 2018. This paper presents the intermediate results which have been reached and the state of the module as of May 2019. The documentation creation process reflecting the discussions of the scope, identified representation needs and modelling steps can be consulted on the respective

OntoLex Wiki page[7]. It contains the outcomes as well as the links to the minutes of the regular calls that have been held.

So far, half of the defined scope for the Morphology Module (cf. Section 5.1) could be modelled. In particular this includes the fist scope, i.e. the representation of the decomposition of ontolex:LexicalEntry and ontolex:Form resources. An overview illustrating the resulting model structure is shown in Figure 3. The second scope of representing the automatic generation of entries and forms from morph resources is still in an early development stage and, hence, will not be addressed in detail in this paper. The model in Figure 3 displays how the Morphology Module is embedded within the existing OntoLex-*lemon* vocabulary it relates to. Classes and properties written in blue indicate the new vocabulary that is specified with the prefix morph with the class morph:Morph building the centre of the module. The two object properties decomp:subterm and decomp:correspondsTo are also represented in blue, thus, highlighting that these are vocabulary elements that will have to be adjusted by extending their ranges (as explained in Section 5.2) to arrive at an overall OntoLex-*lemon* model consistency. It has to be noted that the presented Morphology Module is not officially published yet and, therefore, not usable at this current stage. However, it can be assumed that the vocabulary elements that are described in the next Section will remain very close to their final published module specification.

6.2 New classes and properties

In order to solve the presented challenges outlined in Section 5, new classes and properties had to be developed for the Morphology Module. Altogether eleven new classes and seven object properties have been implemented into the modelling so far. In doing so, central concepts of the domain of morphological data could be reused from the OntoLex-*lemon* vocabulary, and a considerable reduction of overlap between the new and the existing vocabulary could be reached. The ontolex:Form class, for instance, was already appropriate to represent all forms of a lexical entry, which are crucial elements for the description of the segmentation of words. Table 1 and Table 2 present an overview of the module vocabulary with the definitions and restrictions that have been defined for all new classes and properties.

The morph:Morph class builds the centre of the module and is divided into six subclasses. As a result it will be possible to specify root, stem and certain affix types. The prominent affixes, i.e. prefix, suffix, infix and circumfix, are, however not part of the vocabulary because these can be reused from other ontologies such as LexInfo. The treatment and function of the ontolex:Affix class was highly debated for its potential re-usability. Since this class is a subclass of ontolex:LexicalEntry it cannot be used to represent bound morphs that are inflectional, because those are usually not described

[7] https://www.w3.org/community/ontolex/wiki/Morphology

as headwords in lexical databases or dictionaries. In order to avoid uncertainty within the classification of inflectional and derivational affixes, the morph:AffixMorph class has been created. Affixes that should be represented as lexical entries can be described with ontolex:Affix, whereas those that cannot should be described with the morph:AffixMorph class, regardless of their derivational or inflectional nature. Moreover, an explicit declaration for these two morphological functions has been enabled by providing the object property morph:hasMorphStatus and the class morph:MorphValue that already contains the two individuals morph:inflectional and morph:derivational ready for use.

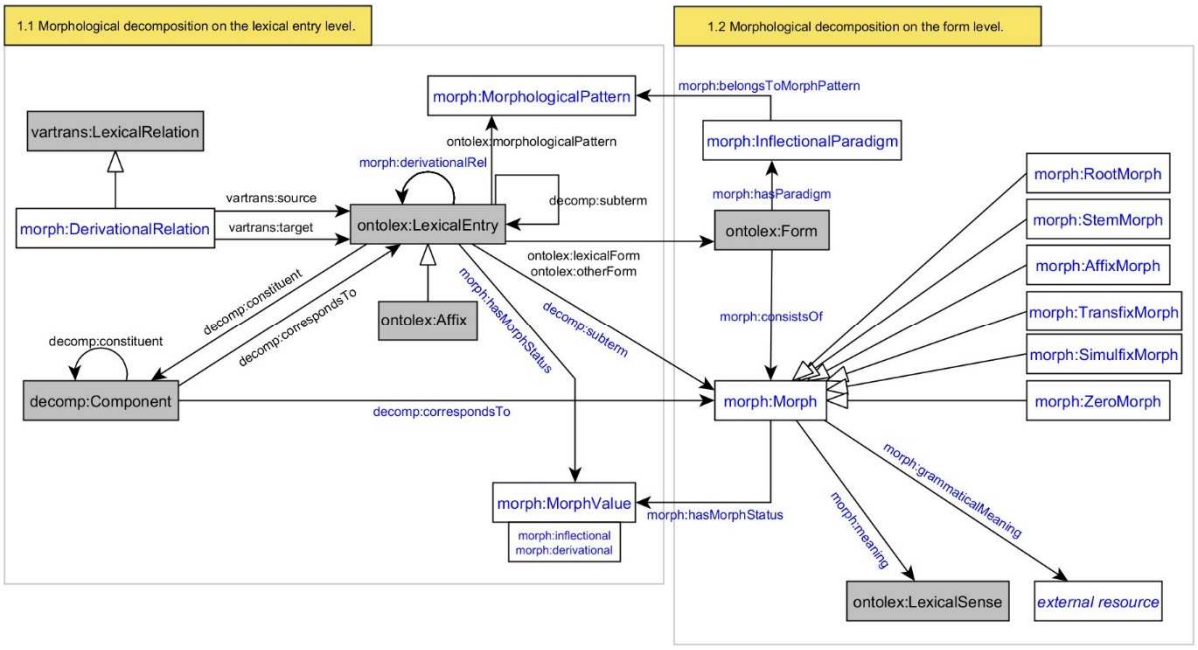

Figure 3: Current proposal of the Ontolex-*lemon* morphology module.

Since the derivational morphs of a derived lexical entry are now explicitly representable within the Morphology Module, a possibility to state that one derived lexical entry is derived from another lexical entry should be provided. This has been achieved by creating the class morph:DerivationalRelation that is defined as a subclass of vartrans:LexicalRelation. Therefore, it inherits the same domain and range restrictions which mean it can represent the direction of the derivational relation between two lexical entries, i.e. one can explicate that one derived lexical entry is derived by a specific derivational relation from another lexical entry. Furthermore, more generically all lexical entries that can be created through a derivational relation from another lexical entry can be expressed by using the object property morph:derivationalRel. Examples illustrating the use of this class and this property will be provided in Section 6.3.1.

Class Name	**Definition**	**Class Relation**
Morph	A morph is a concrete primitive element of morphological analysis.	owl:disjointWith ontolex:LexicalEntry
RootMorph	A morph that constitutes the semantic nucleus of a stem. It cannot be further segmented and is often not specified for a part of speech.	rdfs:subclassOf morph:Morph
StemMorph	The stem is the morph to which inflectional marking applies.	rdfs:subclassOf morph:Morph
AffixMorph	An affix is a bound segmental morph.	rdfs:subclassOf morph:Morph
TransfixMorph	A transfix is a discontinuous affix.	rdfs:subclassOf morph:Morph
SimulfixMorph	A simulfix is a bound morph that entails a change or replacement of vowels or consonants (usually vowels) which changes the meaning of a word, e.g. *eat* in past tense becomes *ate*.	rdfs:subclassOf morph:Morph
ZeroMorph	A morph that that corresponds to no overt form, i.e. orthographic or phonetic representation.	rdfs:subclassOf morph:Morph
MorphValue	The value of a morph states the relationship that holds between the morph and the forms or lexical entries in which it can occur.	class instances: morph:inflectional morph:derivational
DerivationalRelation	A 'derivational relation' is a lexical relation that relates two lexical entries by means of a derivational affix.	rdfs:subclassOf vartrans:LexicalRelation
MorphologicalPattern	The morphological pattern states the inflectional, derivational or compositional building pattern that applies to a lexical entry.	none
InflectionalParadigm	A structured set of inflected forms according to specific grammatical parameters.	none

Table 1: Overview of new classes of the Morphology Module.

With the foresight to enable also the automatic generation of ontolex:LexicalEntry resources from given morph:Morph and ontolex:Affix resources, the necessary conceptual frame has been modelled already. Figure 3 shows that the existing ontolex:morphologicalPattern object property was an initial proposal but remained under specified due to the non-existent Morphology Module at the point of its creation. This lack of expressivity has been now resolved by creating the two classes morph:MorphologicalPattern and morph:InflectionalParadigm which interrelate

ontolex:LexicalEntry and ontolex:Form within the graph structure of the module via the two established object properties morph:hasParadigm and morph:belongsToMorphPattern. Even though the specific usage of this part of the module is not sufficiently attested yet, the example for it provided in Section 6.3 illustrates the intended utilization.

As a central component of the morphological data domain the representation of the meaning of morph:Morph resources had to be modelled as well. Therefore, the two object properties morph:meaning and morph:grammaticalMeaning have been implemented in the module. The underlying concepts of morph:StemMorph and morph:RootMorph resources can be expressed by the former property by pointing to a ontolex:LexicalSense resource and the grammatical categories that are encoded in resources that represent grammatical morphs, usually bound affixes, can be expressed by pointing to an external resource. As already mentioned, the creation of an extensive modelling of possible linguistic categories has been considered to be out of scope for this module, and it is recommended to reuse existing vocabulary elements, e.g. from LexInfo, instead. The possible lack of a grammatical catogory in any existing ontology can be then compensated by using the morph:grammaticalMeaning property alternatively together with a newly created vocabulary.

Property Name	**Definition**	**Restrictions**
derivationalRel	The property relates two lexical entries that stand in some derivational relation.	domain: ontolex:LexicalEntry ontolex:LexicalEntry
consistsOf	This property states into which Morph resources a Form resource can be segmented.	domain: ontolex:Form morph:Morph
hasMorphStatus	The property states whether a morphological element functions as inflectional or derivational.	domain: morph:Morph, ontolex:Affix morph:MorphValue
hasParadigm	This property assigns a form to an inflectional paradigm.	domain: ontolex:Form morph:InflectionalParadigm
belongsToMorphPattern	This property assigns an inflectional pattern of a form as belonging to a morphological pattern of a lexical entry.	domain: morph:InflectionalParadigm morph:MorphologicalPattern
meaning	This property assigns a lexical sense to a morph resource.	domain: morph:Morph ontolex:LexicalSense
grammaticalMeaning	This property assigns a grammatical meaning to a morph resource.	domain: morph:Morph

Table 2: Overview of new object properties of the Morphology Module.

Finally, a relation was needed that states that an ontolex:Form resource consists of morph:Morph resources analogously to the ontolex:constituent object property that interrelates ontolex:LexicalEntry resources and decomp:Component resources. This relation manifests itself in the object property morph:consistsOf which is used to identify the segmentable morphs of inflected words, whereas ontolex:constitutent can identify the lexical parts of derived or compounded words. By further extending the range of ontolex:correspondsTo and ontolex:subterm for the class morph:Morph it is even possible to identify inflectional affixes within complex lexical entries. This is a particularly useful functionality of the morphology module for many languages that involve the expression of an inflectional morph in the process of word-formation. German nominal compounds, for example, can consist of some linking morph that can be identified as a case marking morph (or depending on the underlying linguistic theory as a zero morph), e.g. as in *Haushalt-s-kasse*, 'household-GEN-budget'.

6.3 Representing morphological decomposition

In what follows the usage of the introduced vocabulary of the Morphology Module will be illustrated by the example displayed in Figure 4. It shows the graph modelling evolving around the English noun *speaker*, including all the properties, classes and instances that are involved. For better understandability the graph is reduced to the representation of only one derived lexical entry, i.e. the adjective *speakerless* and only two word-forms of *speaker*, assuming that there are more. All boxes highlighted in yellow represent the new classes of the Morphology Module vocabulary.

6.3.1 On the lexical entry level

Looking at the resource :lex_speaker_n as the subject of this graph clarifies which morphological information can be explicated by creating the following statements:

1) It consists of two constituents which are decomp:Component resources which again can be said to correspond to another ontolex:LexicalEntry and a morph:AffixMorph resource, i.e. the verb :lex_speak_v and the derivational suffix :suffix_er. This suffix has been specified with the value morph:derivational and the ontolex:LexicalSense :agentNominalizer. This modelling indicates that in this example dataset this derivational suffix *-er* is explicitly not a lexical entry but could, however, be easily turned into one by changing its type assertion to ontolex:Affix.

2) It can be created with the morphological pattern :pattern_CommonNouns. As mentioned already, this is technically not implemented yet but it is intended to use the two decomp:Component resources :component_speak and :component_er for this purpose.

3) It can be linked to other lexical entries by using the morph:derivationalRel property in order to state which other derived words can be derived from :lex_speaker_n. This is, however, only a very generic statement but one that is often found in lexical or dictionary data.

Finally, the statement in 3) can be specified in a fourth statement by turning :lex_speaker_n into an object of a statement that describes it as the target of the derivational relation :derivRel_speaker_AgentNoun. While the property in statement 3) just states that there is some derivational relation between two ontolex:LexicalEntry resources, triples with a morph:DerivationalRelation instance in the subject position explicitly interlink the source lexical entry and the target lexical entry for which a unique derivational relation holds.

6.3.2 On the form level

The interconnection between lexical entries and the forms that can be built from them has been already established within OntoLex-*lemon* with the ontolex:otherForm property and has been, therefore, used in this example accordingly to relate the two forms :form_speakers1 and :form_speakers2 to the lexical entry :lex_speaker_n.

Considering these two instances as the subjects when consulting Figure 4 makes it possible to create the following statements about them:

1) They are both specified to belong to the inflectional paradigm :paradigm_NounInflecion. This paradigm defines the grammatical form variants of the ontolex:Form resources, i.e. case and number, and is itself assigned to the overall building pattern :pattern_CommonNouns for ontolex:LexicalEntry resources that are nouns like :lex_speaker_n.

2) They are both segmentable into morph:Morph resources that are stated with the morph:consistsOf property. As it is clear from Figure 4, they both share the same morph:StemMorph resource but consist of two different morph:SuffixMorph resources.

In addition to that, the three morphs :stem_speaker_n, :suffix_s1 and :suffix_s2 can be further specified for their meanings by pointing to ontolex:LexicalSense instances and grammatical values for the linguistic category case reused from the LexInfo vocabulary. It is essentially due to this enabled decomposition chain that makes it possible to not only identify, specify and interrelate all meaningful sub-word units but also the lexical entries and forms contained in lexical data, that all these elements can be disambiguated and described within a dataset modelled with the Morphology Module and OntoLex-*lemon*.

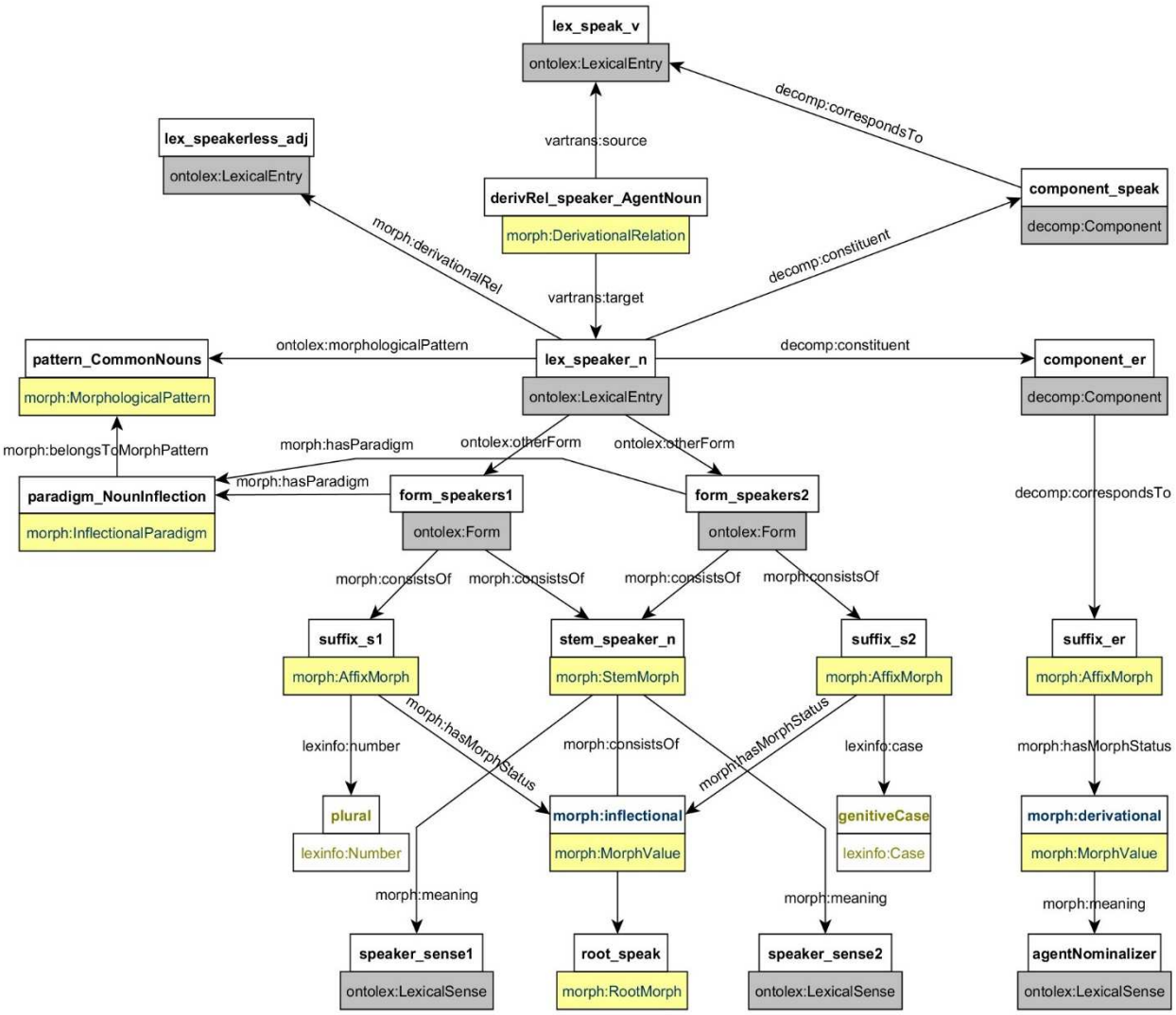

Figure 4: Graph representation for the example entry :lex_speaker_n.

7. Future work

Even though the modelling outcomes presented here have been largely agreed upon, several issues remain open for future work. Due to the various linguistic backgrounds of the OntoLex Community Group members some desired implementation options have been raised that might be still realized and included within the final Morphology Module specification. The following three features have been proposed for additional realization and are still under discussion:

1) **Morphemic glosses:** Since interlinear glossed text language data is an emerging source of lexical data that can be also represented in RDF, interest has been indicated to include the representation of morphemic glosses. So far it has been discussed if a modelling of glosses would exceed the scope of the Morphology Module, while the option to provide a shallow modelling with an

alignment to the MMoOn Core vocabulary that already provides a representation of glosses is also considered.

2) **Ordering:** For some highly polysynthetic and morphology-rich languages it is desirable to have a more precise representation of the internal morphological structure of lexical entries and forms. Therefore, it has been decided that a more expressive possibility for representing the position and ordering of morphs should be implemented to be available next to the currently used but very inexpressive rdfs:list object property. Proposals for that have been already made, but no agreement has been reached yet.

3) **Multiple segmentations:** Taking into account that a lexical dataset created based on the Morphology Module could be also applied in the context of computational linguistics, the processability of this data for machines might require the representation of more than one possible segmentation strategy. Allowing for the explication of that would be also interesting for linguists who want to document and analyse competing segmentations of words in their research.

In addition to these yet unrealized features it is necessary to focus on the refinement of the definitions of the newly created vocabulary elements. The exchanges within the community group have revealed that some of the presented rdfs:comment information is not precise enough and might lead to misunderstandings. In order to avoid misunderstandings in the usage of the vocabulary, time and attention will be invested again to resolve currently ambiguous or unclear definitions.

Furthermore, the second part of the Morphology Module that will enable the generation of forms with existing productive morphs in a dataset is also a part of the future work. However, the modelling is envisaged to produce lexical entries and forms based on patterns and paradigms, including also discontinuous morphs like transfixes and infixes. As it turned out in previous discussions such a formal representation is not trivial to model, especially with regard to the aim to be language-independently applicable.

8. Conclusion

To summarize, the current state of the Ontolex-*lemon* Morphology Module has been presented. The created vocabulary has been introduced and its usage illustrated. From that it becomes clear that the new module overcomes the limitations of the current representation of morphological data contained in traditional dictionaries by enabling the explication of formerly implicit information. With the Morphology Module modelled so far it is possible to represent the decomposition of lexical entries and forms with regard to both their derivational and inflectional morphs and underlying building patterns.

Furthermore, the challenges that arose from integrating the module into the existing

Ontolex-*lemon* model have been explained and design choices have been supported. It has been also shown that the module applies to existing Semantic Web standards by reusing relevant existing ontologies within its framework.

The remaining open issues have been presented and will be addressed in future work in order to arrive at the release of the final Morphology Module specification.

9. Acknowledgements

John McCrae is supported in part by a research grant from Science Foundation Ireland, cofunded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289, as well as by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure) and 825182 (Prêt-à-LLOD).

Julia Bosque-Gil is supported by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program.

Maxim Ionov and Christian Chiarcos are supported by the German Ministry for Education and Research (BMBF) through a project Linked Open Dictionaries (LiODi, 2015-2020) as a part of an Early Career Research Group on eHumanities.

10. References

Abel, A. (2012). Dictionary Writing Systems and Beyond. In S. Granger and M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. Chap. 5, pp. 86–106.

Alsina, V. & DeCesaris, J. (1998). *Morphological structure and lexicographic definitions: The case of -ful and -like*.

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Bellandi, A., Giovannetti, E. & Weingart, A. (2018). Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* 9.3, p. 52.

Bergenholtz, H. & Tarp, S. (2005). Dictionaries and inflectional morphology. In *Encyclopedia of Language and Linguistics*. Pergamon Press, pp. 577–580.

Borin, L. et al. (2014). Representing Swedish Lexical Resources in RDF with lemon. In: *International Semantic Web Conference (Posters & Demos)*. Citeseer, pp. 329– 332.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. (2017). Towards a module for lexicography in OntoLex. In: *DICTIONARY News* 7.

Bosque-Gil, J., Gracia, J., Aguado-de-Cea, G. et al. (2015). Applying the ontolex model to a multilingual terminological resource. In *European Semantic Web Conference*. Springer, pp. 283–294.

Buitelaar, P. et al. (2006). LingInfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the OntoLex Workshop at LREC*.

Chiarcos, C. & Sukhareva, M. (2015). Olia – ontologies of linguistic annotation. *Semantic Web* 6.4, pp. 379–386.

Chiarcos, C., Abromeit, F. et al. (2016). Etymology Meets Linked Data. A Case Study In Turkic. In *Digital Humanities 2016, DH 2016, Conference Abstracts*. Krakow, Poland: Alliance of Digital Humanities Organizations (ADHO), pp. 458– 460. ISBN: 978-83-942760-3-4.

Chiarcos, C., McCrae, J. et al. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer, pp. 7–25.

Cimiano, P., McCrae, J. et al. (2013). "On the role of senses in the ontology lexicon". In *New trends of research in ontologies and Lexical resources*. Springer, pp. 43–62.

Cimiano, P., McCrae, J. & Buitelaar, P. (2014). *Lexicon Model for Ontologies: Community Report*. W3C Community Group Final Report.

Cimiano, P., Buitelaar, P. et al. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* 9.1, pp. 29–51.

Cimiano, P., Haase, P. et al. (2007). "LexOnto: A model for ontology lexicons for ontology-based NLP". In *Proceedings of the OntoLex07 Workshop held in conjunction with ISWC'07*.

De Caluwe, J. & Taeldeman, J. (2003). 2.5 Morphology in dictionaries. In P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*. Vol. 6. John Benjamins Publishing, pp. 114–126.

Farrar, S. & Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT international* 7.3, pp. 97–100.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem et al. (eds.) *Proceedings of eLex 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.

Hartmann, R. R. K. (2001). *Teaching and researching lexicography*. Routledge.

Ide, N. & Romary, L. (2004). A registry of standard data categories for linguistic annotation. In *4th International Conference on Language Resources and Evaluation-LREC'04*, pp. 135–138.

Kemps-Snijders, M. et al. (2009). ISOcat: Remodeling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)* 4.4, pp. 261–276.

Khan, F. (2018). Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In J. P. McCrae et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). isbn: 979-10-95546-19-1.

Khan, F. et al. (2017). The Challenges of Converting Legacy Lexical Resources to Linked Open Data using Ontolex-Lemon: The Case of the Intermediate Liddell-Scott Lexicon. In: *LDK Workshops*, pp. 43–50.

Klimek, B. et al. (2016). Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Klimek, B. (2017). Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*.

McCrae, J., .Fellbaum, C. & Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

McCrae, J., Aguado-de-Cea, G. et al. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation* 46.6, pp. 701–709.

McGuinness, D. L., Van Harmelen, F. et al. (2004). *OWL: Web Ontology Language overview*. W3C recommendation.

Montiel-Ponsoda, E. et al. (2011). Enriching ontologies with multilingual information. *Natural language engineering* 17.3, pp. 283–309.

Real Academia Española and Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Santillana Ediciones Generales.

Schierholz, S. J. (2015). Methods in Lexicography and Dictionary Research. *Lexikos* 25, pp. 323–352.

Swanepoel, P. H. (2015). The design of morphological/linguistic data in L1 and L2 monolingual, explanatory dictionaries: a functional and/or linguistic approach? *Lexikos* 25, pp. 353–386.