CEUR-WS.org/Vol-1899/OntoLex_2017_paper_6.pdf

# Proposing an OntoLex - MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models

Bettina Klimek[1]

AKSW/KILT, University of Leipzig,
klimek@informatik.uni-leipzig.de,
WWW home page: http://aksw.org/Groups/KILT.html

**Abstract.** This paper motivates and proposes to align the OntoLex and MMoOn Core models. It deals in particular with the ontolex and decomp modules and their potential as ontological foundation to represent the domain of morphological language data (MLD). It will be argued that ontolex and decomp provide only a basic modelling of the domain, which is not sufficient for representing fine-grained MLD, but suitable for interconnecting OntoLex with the Multilingual Morpheme Core Ontology (MMoOn Core). Both models each offer a modelling of a linguistic domain - OntoLex for lexical language data and MMoOn Core for morphological language data - that exhibits a notable amount of conceptual overlap. Thus, this paper investigates the potential of exploiting the overlap of both models for initiating an ontology-based interconnection of lexical and morphological datasets.

**Keywords:** MMoOn, OntoLex, morphology, ontology alignment

## 1 Introduction

The development of OntoLex as a standardized model for the ontological representation of lexical language data has gained high acknowledgement within the Linguistic Linked Open Data (LLOD) community. A reason for that lies in the far reaching modelling of lexical language data (LLD) that goes beyond the domain of lexicography. By providing five modules, the OntoLex model can be used according to the needs of a dataset creator to also represent morphological, syntactical, semantic and translational information about a lexical entry as well. I.e. the OntoLex model encompasses the representation of other linguistic domains as well.

This paper deals in particular with the ontolex and decomp modules and their potential as ontological foundation to represent the domain of morphological language data (MLD). It will be argued that ontolex and decomp provide only a basic modelling of the domain, which is not sufficient for representing fine-grained MLD, but suitable for interconnecting OntoLex with the Multilingual Morpheme Core Ontology (MMoOn Core)[1]. Both models each offer a modelling

[1] http://mmoon.org

of a linguistic domain – OntoLex for LLD and MMoOn Core for MLD – that exhibits a notable amount of conceptual overlap. The aim of this paper is, thus, to investigate the potential of exploiting this overlap of both models for initiating an alignment of both ontologies. Dataset creators of either OntoLex or MMoOn datasets would benefit from such a unification in that it enables the seamless extension of lexical OntoLex data with morphological MMoOn data or of morphological MMoOn data with lexical OntoLex data.

The remainder of the paper is structured as follows. Section 2 gives a brief overview of the domain of MLD and its overlap to the domain of LLD. In Section 3 the MMoOn Core model is summarized and presented as a suitable model for the domain of MLD. The main part of the paper constitutes Section 4 which investigates the representation of MLD in both models in a comparative way and which points to the overlapping aspects. It further serves to not only prove that the MMoOn Core model is qualified to be interconnected with OntoLex but also to show that both models would benefit from an alignment with regard to the representation of language data in both linguistic domains. Thereafter, in Section 5, specific interconnection points between both models are proposed together with practical issues that need to be considered for implementing an alignment of both ontologies. The paper closes with a summary in Section 6.

## 2 Scope and Delimitation of the Domain of Morphology

In traditional linguistics research fields such as phonology/phonetics, morphology, lexicology, syntax, semantics and pragmatics are distinguished. However, the study of one field reveals considerate inter-dependencies to other fields. E.g. the field of morphophonology investigates the interface of phonology and morphology. Similarly, there is an overlap of morphology and lexicology which in the view of linguistic data representation makes it hard to state where the domain of morphology ends and the domain of lexicography begins. Since "lexical items are the fundamental building blocks of morphological structure" [4] it is not satisfactory to represent lexical data only in lexicons and morphological data only in morphemicons. Even though such lexicons and morphemicons constitute valuable data resources, it is desirable to interconnect both. E.g. a lexicon entry might be the English adjective *unreal* and a morphemicon entry might be the negation prefix *un-*. In both separate dataset types the information that the adjective consists of this very prefix and the information which other lexical entries also contain this prefix is missing.

In the scope of Linked Data such information can be modelled in an ontology, which provides the necessary relations that interconnect lexical items and morphological items. But then the question arises: what kind of data should be represented in the linguistic domain of morphology? Figure 1 illustrates a data-driven view of the domain with the English example lexeme *(to) play*. The box in the middle indicates the narrow scope of the domain, i.e. which elements and their relations need to be modelled in order to represent MLD. The central entries of MLD are thus morphs, morphemes and meanings. But also information

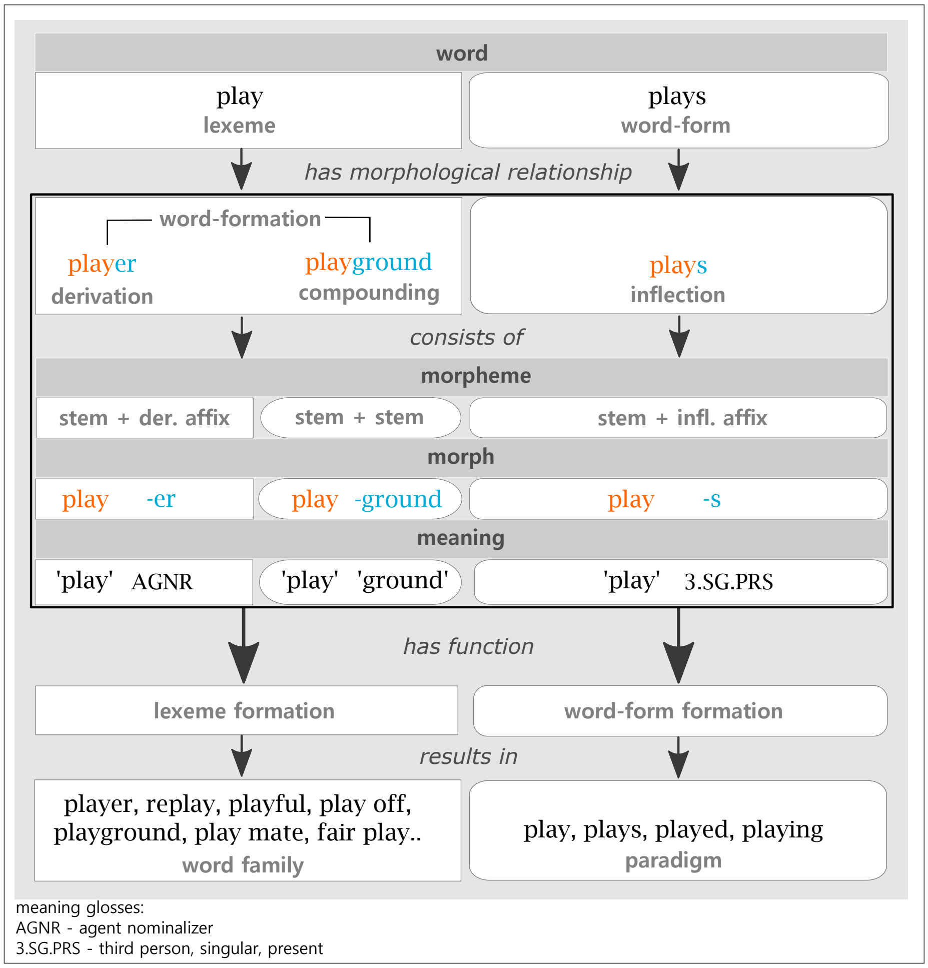


**Fig. 1.** Overview of the linguistic domain of morphology with the English example lexeme "play" (verb).

on the function as a derivational, compounding or inflectional morph/morpheme within a given word needs to be provided. The rest of the Figure shows how the the narrow MLD is interrelated to corresponding LLD. As a result, Figure 1 as a whole shows the wide scope of the MLD domain, which then also includes lexemes and word-forms. Consequently, only in the wide scope of MLD interrelating information between lexical and morphological items can be obtained, i.e. the identification of word-families and word-forms. This, however, means that there is no clear cut delimitation between the two domains of LLD and MLD and especially word-families and word-forms could be regarded as elements of both domains, since their representation requires lexical as well as morphological knowledge and data. Nonetheless, in the research field of LLOD this situation is not problematic. To the contrary, the Linked Data format is open for extension, so that existing lexical or morphological datasets in RDF can be interconnected across various vocabularies, e.g. by aligning the two domain models of OntoLex and MMoOn.

## 3 Why MMoOn?

To my knowledge, the *lemon* model [10, 9] and the resulting OntoLex model were the first models to provide an ontological representation of the domain MLD. Given that the central domain of these vocabularies is LLD, it is not surprising that the domain of MLD as shown in Figure 1 is only partially covered and questions arose on the applicability of both models for representing more fine-grained MLD [2].

In order to fill the gaps in these vocabularies (which will be discussed in Section 4) and to obtain a more extensive model that covers the full domain of MLD, the MMoOn Core ontology has been developed. The MMoOn model has proven to be applicable for the representation of inflectional languages, even for those exhibiting non-concatenative morphology, such as Hebrew [6]. Due to the shortage of space in this paper an image giving an overview of the MMoOn Core ontology can be found here: `http://mmoon.org/mmoon-core-model`, the full vocabulary here: `http://mmoon.org/core.owl` and the documentation of anything related to the ontology and emerging datasets here: `https://github.com/MMoOn-Project`. The MMoOn Core ontology is designed as a language-independent and theory-neutral model to create language-specific morpheme inventories. It consists of eight main classes: `MorphemeInventory`, `MorphologicalRelationship`, `MorphemicGloss` and `Meaning` which enable the representation of secondary language data and `Word`, `Morph`, `Morpheme` and `Representation` which are used to describe primary language data[2]. With regard to the modelling of secondary data, the OntoLex developers declare that the model "does not prescribe any vocabulary for doing so [i.e. recording linguistic properties] , but leaves it at the discretion of the user of the model to select an appropriate vocabulary [...]"[3]. As this complies to the common best practice for Linked Data to reuse existing vocabularies, such descriptive secondary language data will remain undiscussed within the modelling of MLD in OntoLex in this paper. It shall be noted, that MMoOn Core comes with nearly 300 meanings to which morphemic glosses are already assigned. Even though there is an overlap to vocabularies such as LexInfo [3], meaning resources are included in MMoOn because it includes also derivational meanings and facilitates the creation of a

[2] The former includes descriptive data which enables the assignment of linguistic features (or properties), e.g. grammatical categories or part of speech, and the latter contains all elements and their relations within a given language that are part of the domain, e.g. morphs, morphemes, word-forms. (For more detail see [7].)

[3] Every reference to the OntoLex model or any of its modules is made with regard to the model specification here: `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Linguistic_Description`.

MMoOn dataset (especially for linguists, who then do not have to deal with various vocabularies). The class hierarchies in MMoOn Core are fine-grained and interrelated with various object properties. This allows for explicitly stating which parts of the words are morphologically formed as well as to which words morphs and morphemes belong. As a result, a MMoOn morpheme inventory is more than a mere morphemicon: it is a semantically structured data graph that can be traced in both directions from words to morphemes in a semasiological and an onomasiological way. In particular the modelling encompasses the elements and their relations of the domain of MLD as shown in Figure 1. A dataset created with MMoOn is called a MMoOn morpheme inventory. Every morpheme inventory consists of three files: 1) The Core model, which functions as a cross-linguistic template for the domain of MLD, 2) a schema file, which is language-specific and describes the secondary language data and 3) an inventory file that contains only primary language data, i.e. only instance data. This schema file – or language-specific morpheme ontology – is derived from and imports the Core ontology. Hence, it contains all elements that are already provided in MMoOn Core and can be easily further adjusted and extended according to the morphological phenomena that shall be represented in a given language. Thus, the MMoOn Core model is suitable for the semantic modelling of MLD of any inflectional language and, therefore, an appropriate candidate for an alignment with the ontolex and decomp module.

## 4 Representing Morphological Data

In the following sections it will be shown how MLD is representable with MMoOn on the one side and with the ontolex and decomp modules on the other side. This direct comparison takes up Figure 1 as running example and aims at stressing why an interconnection of MMoOn and the two modules can be regarded as a valuable contribution to the ontological modelling of LLD and MLD in general.

### 4.1 Morphology on the Lexeme Level

A fundamental distinction in the domain of morphology is inflection and word-formation. The former involves word-form formation and the latter lexeme formation. Inflectional information on the lexeme level contains information on the building pattern of the word-forms of a lexeme.

As Example 1 shows, the ontolex object property `morphologicalPattern` can be used to express the inflectional class of a lexeme. The "?" in the subject slot indicates that the OntoLex model specification states that "the implementation of these patterns is not specified [...] but should be provided by some suitable vocabulary such as [the Lemon Inflectional and Agglutinative Morphology Module for OntoLex] LIAM[4]". What is more, the object property provided, does not differentiate inflectional and derivational relations of lexemes. The MMoOn Core

[4] `http://lemon-model.net/liam`

| ontolex/decomp | MMoOn |
|---|---|
| **Example 1:** Representing the inflectional morphological relationship of a lexeme. | |
| `ontolex:lex_play a ontolex:Word ;`<br>`ontolex:morphologicalPattern ? .` | `eng_inv:SimpleLexeme_play_v`<br>`a eng_schema:SimpleLexeme ;`<br>`mmoon:inflectionalRelation`<br>`eng_schema:RegularConjugation .` |
| **Example 2:** Representing the derivational morphological relationship of a lexeme. | |
| `ontolex:lex_play a ontolex:Word ;`<br>`ontolex:morphologicalPattern ? .` | `eng_inv:DerivedWord_player_n`<br>`a eng_schema:DerivedWord ;`<br>`mmoon:derivationalRelation`<br>`eng_schema:AgentNoun .` |
| **Example 3:** Representing that one lexeme is derived from another lexeme. | |
| `ontolex:lex_player a ontolex:Word ;`<br>`?decomp:subterm ontolex:lex_play .` | `eng_inv:DerivedWord_player_n`<br>`a eng_schema:DerivedWord ;`<br>`mmoon:isDerivedFrom`<br>`eng_schema:SimpleLexeme_play_v .` |
| **Example 4:** Representing that a lexeme is a compound word that is composed of two other lexemes. | |
| `ontolex:lex_playground`<br>`a ontolex:Word ;`<br>`decomp:subterm ontolex:lex_play ,`<br>`ontolex:lex_ground .` | `eng_inv:CompoundWord_playground_n`<br>`a eng_schema:CompoundWord ;`<br>`mmoon:isComposedOf`<br>`eng_schema:SimpleLexeme_play_v ,`<br>`eng_schema:SimpleLexeme_ground_n .` |

model, however, already contains a basic modelling of classes for inflection and word-formation within the `MorphologicalRelationship` main class, which are automatically reused and provided in every language specific MMoOn schema ontology, e.g. `eng_schema` in the provided examples.

The case of Example 2, representing the derivational morphological relationship of a lexeme, is similar to Example 1. While the MMoOn vocabulary provides object properties that indicate an inflectional or derivational relation and also the kind of this relation in the subject slot of the triple, the ontolex object property remains ambiguous. Given that this property has no range declaration, it is, however, possible to use the MMoOn vocabulary to fill the subject slot. Further, it is important to note, that the LIAM vocabulary does not provide a general ontological modelling of morphological relationships such as MMoOn. Rather, it models the transformation rules that apply to a pattern underlying a specific morphological relation, which could then be applied for instance to `eng_schema:RegularInflection` or `eng_schema:AgentNoun`[5].

[5] `AgentNoun` is part of the class hierarchy: MorphologicalRelationship> WordFormation>Derivation>DerivedNoun>DeverbalNoun>AgentNoun

The Examples 3 and 4 show which other lexemes are involved in a word-formation process. The MMoOn vocabulary provides the two object properties `isDerivedFrom` and `isComposedOf` to state from which lexeme a derived word is derived and of which two lexemes a compound word is composed. The decomp object property `subterm` can be equivalently used for compound words in Example 4. The "?" in Example 3, however, indicates that this predicate is not appropriate for stating that the noun *player* is derived from the verb *play*, because `subterm` is defined as a property that "relates a compound lexical entry to one of the lexical entries it is composed of"[6].

As the examples show, the ontolex and decomp vocabulary is not accurate enough to represent the morphological relationship, either inflectional or derivational, of lexemes. In the cases of stating which lexemes are involved in the word-formation process, the model clearly favours compound words, while lacking an object property that interconnects a lexeme as the basis of a derived word. For such cases the MMoOn vocabulary would be a valuable addition to represent more fine-grained lexical data because it provides more specific object properties and also a more precise classification of lexical entries, i.e. it distinguishes simple lexemes, which are neither composed nor derived from other lexemes, derived words and compound words as subclasses of the MMoOn `LexicalEntry` class. What is more, an alignment of the ontolex `LexicalEntry` class with these classes would be cruicial in order to interconnect an OntoLex lexical dataset with a MMoOn morpheme inventory.

### 4.2 Morphology on the Word-form Level

In the domain of MLD word-forms play a central role, because these are the entities which contain the inflectional affixes that mark the grammatical variant of a lexical entry. Consequently, all word-forms of a lexeme need to be represented as separate resources in a dataset. As can be seen in Example 5, both the ontolex module and the MMoOn ontology provide properties and classes to do so[7]. While ontolex has one class, `Form`, in MMoOn the class `Wordform` is further specified for the two subclasses `SyntheticWordform` and `AnalyticWordform`[8]. In order to enable the extraction of inflectional paradigms of lexemes, word-form instances in MMoOn can be assigned to more specific morphological relationships. I.e. the synthetic word-forms *play* and *plays* belong to a regular present tense conjugation paradigm (which is not shown in the example but works similar to lexemes shown in Example 1). The analytic word-form *has played*, however, belongs to

[6] cf.:`https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`

[7] Note that the word-forms in Example 6 are not complete.

[8] The two concepts of 'synthetic' and 'analytic word-form' correspond to the definitions of Christian Lehmann: "A word form is synthetic [...] iff all its semantic and grammatical components are represented in one word form." and "A word form is analytic iff it consists of more than one word form such that the lexical meaning provides the root of one of them, while the grammatical meaning components are coded in the other word forms [...]. Cf. the entries "analytic structure" and "synthesis" at `http://linguistik.uni-regensburg.de:8080/lido/Lido`.

a regular past tense conjugation paradigm and consists of a word-form of *have* and the past participle of *play*. With the object property `consistsOfWord` every word or word-form which is contained in an analytic word-form can be explicitly stated and further represented as well. This is not possible in ontolex, since the appropriate object properties that are available do not take the class `Form` as range but only `LexicalEntry`[9].

**Example 5:** Representing the word-forms of a lexeme.

**ontolex/decomp**

```
ontolex:lex_play a ontolex:LexicalEntry ;
     ontolex:canonicalForm ontolex:form_play ;
     ontolex:otherForm ontolex:form_plays ,
                       ontolex:form_played ,
                       ontolex:form_playing ,
                       ontolex:form_has_played.
```

**MMoOn**

```
eng_inv:SimpleLexeme_play_v a eng_schema:SimpleLexeme ;
      mmoon: hasWordform  eng_inv:SyntheticWordform_play_1_v  ,
                          eng_inv:SyntheticWordform_plays_v ,
                          eng_inv:SyntheticWordform_played_v ,
                          eng_inv:SyntheticWordform_playing_V ,
                          eng_inv:AnalyticWordform_has_played_v .
```

**Example 6**: Representing the inflectional features that are encoded in a word-form.

**ontolex/decomp**

```
ontolex:form_plays a ontolex:Form ;
    lexinfo:number lexinfo:singular ;
    lexinfo:person lexinfo:thirdperson ;
    lexinfo:tense lexinfo:present .
```

**MMoOn**

```
eng_inv: SyntheticWordform_plays_v a eng_schema:SyntheticWordform ;
      mmoon:inherentInflectionalMeaning eng_schema:ThirdPerson ,
                                        eng_schema:Singular ,
                                        eng_schema:Present .
```

**Example 7**: Representing the morphs of which a word-form consists.

**MMoOn**

```
eng_inv: SyntheticWordform_plays_v a eng_schema:SyntheticWordform ;
      mmoon:consistsOfStem eng_inv:Stem_play_v ;
      mmoon:consistsOfAffix eng_inv:Suffix_s1 .
```

Example 5 further shows that MMoOn provides the property `hasWordform`, which is inverse of `belongsToLexeme`, for interrelating word-forms and lexemes. In ontolex, given that it is primarily concerned with lexical data, two properties

[9] Otherwise, analytic word-forms could be similarly representable in ontolex/decomp as constituents of multiword expressions.

are provided, i.e. canonicalForm and otherForm. This specification is clearly useful for compiling dictionaries. For stating all word-forms of a lexeme it might, however, not always be appropriate. At a first glance it seems as if the classes ontolex:Form and mmoon:Wordform could be used equivalently. That would be true, if all Form instances which are connected to a lexical entry via the two mentioned object properties could be regarded as word-forms of a lexical entry. In languages like German for instance, the canonical lexical entry of verbs is the infinitive, which is not an inflected word-form of the lexical verb entry. Querying all Form instances as 'word-forms' of a LexicalEntry in ontolex might thus return incorrect results. It needs to be mentioned here, that it is not clear from the OntoLex model specification if the representation of word-forms by using the Form class is considered or even intended. From the examples given in the specification one can conclude that different forms ("non-lemma") of lexical entries should be describable, but for a specific representation of the word-forms of a lexical entry the vocabulary seems not explicit enough with regard to the provided object properties and the rather general Form class[10].

Next to representing word-forms as separate resources, stating information about the grammatical features for which a word-form inflects is also part of the MLD domain. Example 6 shows that ontolex proposes here the use of the LexInfo vocabulary. Since one of the purposes of the MMoOn model is to enable a language-specific description of linguistic categories, a wide range of grammatical meanings is provided in the MMoOn Core vocabulary which are reused in every language-specific MMoOn schema ontology, e.g. eng_schema:Singular rdf:type mmoon:Singular. In addition, various differentiating object properties, such as inherentInflectionalMeaning in Example 6 or contextualInflectionalMeaning which are based on [1], are also established.

This kind of "annotating" word-forms or lexemes for their grammatical features is quite common, but of more significance in the domain of MLD is the identification of those meaningful parts within a word-form that encode the grammatical features and which are identifiable by segmentation, i.e. the morph entities. Consequently, it is necessary to state of which morphs a word-form (or word in general) consists. At this point the ontolex/decomp modules delimit the ontological representation to lexical data. Although, the ontolex class Affix is part of the vocabulary, the usage of this class remains quite limited. Because word-forms are not considered as ontolex LexicalEntry, but only as ontolex:Form instances, none of the ontolex/decomp object properties can be used for making more statements about the components of word-forms. Example 7 illustrates how morphs are explicated as segments of word-forms in MMoOn. A word-form always consists of a stem, which is the semantic core shared with the corresponding lexeme, and some inflectional affix(es). With the dedicated property consistsOfMorph[11], which is inverse of belongsTo, morph resources can

[10] Also in the model specification a property ontolex:form is used multiple times, even though not specified in the vocabulary.

[11] The two MMoOn Core object properties used in Example 7 are subproperties of mmoon:consistsOfMorph.

be assigned to the word-forms in which they occur. In this regard, a connection between both models would be very helpful in order to specify more information about word-forms in an OntoLex dataset.

### 4.3 Morphology on the Morph Level

| **Example 8:** Representing (bound) morphs. |
|---|
| **MMoOn** |

```
eng_inv:Stem_play_v a eng_schema:Stem ;
     mmoon:belongsTo eng_inv:SimpleLexeme_play_v ,
                     eng_inv:DerivedWord_player_n.
eng_inv:Suffix_s1 a eng_schema: Suffix ;
     mmoon:attachedToStem eng_inv:Stem_play_v ,
                          eng_inv:Stem_call_v ;
     mmoon:hasRepresentation eng_inv:Representation_s1 .
eng_inv:Representation_s1 mmoon:morphemicRepresentation "-s" .
```

| **Example 9**: Representing the meaning of morphs. |
|---|
| **MMoOn** |

```
eng_inv:Stem_play_v a eng_schema:Stem ;
     mmoon:hasSense mmoon:Sense_play_v_s1 ;
     mmoon:senseLink <http://lexvo.org/id/wordnet/30/verb/play_2_33_00> .
eng_inv:Suffix_s1 a eng_schema: Suffix ;
     mmoon:inherentInflectionalMeaning eng_schema:ThirdPerson ,
                                       eng_schema:Singular ,
                                       eng_schema:Present .
eng_inv:Suffix_er1 a eng_schema:Suffix ;
     mmoon:derivationalMeaning eng_schema:AgentNominalizer.
```

| **Example 10**: Representing morphemic homonymy and allomorphy. |
|---|
| **MMoOn** |

```
eng_inv:Suffix_s1 a eng_schema: Suffix ;
     mmoon:isHomonymTo eng_inv:Suffix_s2, eng_inv:Suffix_s3 .
eng_inv:Suffix_er1 a eng_schema:Suffix ;
     mmoon:belongsTo eng_inv:DerivedWord_player_n ,
     mmoon:isAllomorphTo eng_inv:Suffix_or .
```

Morph (and morpheme) resources constitute the morphemic entries of each MMoOn morpheme inventory and are in the center of the MLD domain. In general they correspond to the segmented line within an interlinear morphemic glossed text [8]. A morph is the perceivable side of a morpheme, i.e. it is orthographically and phonemically representable. For representing bound morphs, the ontolex module provides only the `Affix` class which is not further specified. The only possible statement which can be made, is to make an `Affix` class assignment of some suffix, prefix, infix or circumfix resource. Because of this limitation, the examples 8 to 10 only show MMoOn examples. Nothing that is illustrated can be expressed with the ontolex/decomp modules. The MMoOn vocabulary provides a `Morph` class which contains the following subclasses; `Affix`, `Stem` and `Root`. The `Affix` class is further broken up into the `Prefix`, `Suffix`, `Infix`, `Circumfix`, `Simulfix`, `Transfix`, `EmptyMorph` and `ZeroMorph` subclasses. By that, a precise representation of all morph elements which can be segmented from lexical entries or word-forms shall be enabled. Example 8 shows the representation of the verbal stem *play* and the inflectional suffix *-s*. For stem resources it can be further stated to which word resource they belong. In the example the stem *play* belongs to the simple lexeme *play* and the derived word *player*, but additionally it belongs to all word-forms of the simple lexeme *play*. For affix resources it can be stated to which root or stem resource an affix is attached to. In the example the suffix belongs to two stem resources, indicating that affixes are not only listed but also semantically interconnected to other morphemic or lexical entries in a MMoOn morpheme inventory. Further, the datatype property `morphemicRepresentation` is additionally provided to enable the representation of the morpheme boundary or position of the morph within a word.

By having separate morph resources one can additionally specify which parts of a word encode which meaning. Within word-forms the lexical meaning is usually encoded by the stem resource and the grammatical meaning by the affix(es). This is shown in Example 9. With MMoOn new senses can be defined for stem (and word) resources or one can link already existing senses via the `senseLink` property. Since the sense of a stem is the same as the sense of its corresponding lexeme, lexical sense resources provided in already existing LLD datasets could be used to assign sense information to MMoOn `Stem` instances. Lexical senses are not regarded as part of the MLD domain within the MMoOn Core model, but extensively modelled within OntoLex, which presents a potential interconnection point between both models. The grammatical meanings of inflectional affixes like `eng_inv:Suffix_s1` is stated with the same property as in Example 6. In contrast, however, this assignment to the suffix resource is more precise in terms of morphological segmentation. Moreover, MMoOn provides the property `derivationalMeaning` and a set of derivational meanings which can be used to specify resources such as `eng_inv:Suffix_er1`.

Finally, the MMoOn Core vocabulary contains two properties to state homonymous and allomorph relations between morphs, as is illustrated in Example 10. There are several *-s* suffixes in English which share the same surface form but encode different meanings. The `eng_inv:Suffix_s2` encodes plural in nouns and the `eng_inv:Suffix_s3` marks the genitive case, hence, they are represented as being homonym to `eng_inv:Suffix_s1`. For morphs which have different surface forms but share the same meaning the `isAllomorphTo` property is established. E.g. the two instances `eng_inv:Suffix_er1` and `eng_inv:Suffix_or` both encode the derivational meaning of agent nominalizer but occur in complementary distribution, i.e. they attach to distinct verb stems. As the examples show, the

MMoOn vocabulary enables a fine-grained representation of morphemic language data that is semantically relatable to lexical language data. A connection of the ontolex/decomp modules with MMoOn Core would facilitate a morphological description of lexical data with MMoOn on the one side and a lexical description of morphemic data with ontolex/decomp on the other side.

### 4.4 Morphology on the Morpheme Level

**Example 11:** Representing morphemes.

**MMoOn**

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme .
```

**Example 12**: Representing the meaning a morpheme resource represents.

**MMoOn**

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme ;
    mmoon:derivationalMeaning eng_schema:AgentNominalizer ;
    mmoon:hasAbstractIdentity mmoon:MorphemicGloss_AGNR .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme ;
    mmoon:inflectionalMeaning eng_schema:ThirdPerson ,
                              eng_schema:Singular ,
                              eng_schema:Present ;
    mmoon:hasAbstractIdentity mmoon:MorphemicGloss_3P ,
                              mmoon:MorphemicGloss_SG ,
                              mmoon:MorphemicGloss_PRS .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme ;
    mmoon:inflectionalMeaning eng_schema:NoMeaning ;
    mmoon:hasAbstractIdentity mmoon:MorphemicGloss_E .
```

**Example 13**: Representing the relation between morphemes and morphs.

**MMoOn**

```
eng_inv:AtomicMorpheme_AGNR a eng_schema:AtomicMorpheme ;
    mmoon:hasRealization eng_inv:Suffix_er1 ,
                         eng_inv:Suffix_or .
eng_inv:FusionalMorpheme_3P_SG_PRS a eng_schema:FusionalMorpheme ;
    mmoon:hasRealization eng_inv:Suffix_s1 .
eng_inv:EmptyMorpheme_E a eng_schema:EmptyMorpheme ;
    mmoon:hasRealization eng_inv: EmptyMorph_u .
```

Next to morphs morphemes are the central resources within the domain of MLD. Morphemes are the smallest meaningful units of language and represent the conceptual side, i.e. the meaning, of morphs. Such data is not part of a lexical dataset and, thus, not modelled in the ontolex/decomp modules. The MMoOn Core vocabulary provides three `Morpheme` subclasses, i.e. `AtomicMorpheme`, `FusionalMorpheme` and `EmptyMorpheme`, which can be used to represent morpheme resources. For illustration serves Example 11. If exactly one meaning is represented in a language, the morpheme instance is of the type `AtomicMorpheme` If, however, more than one meaning is represented and fused into one morph within a language, the morpheme instance is of the type `FusionalMorpheme`. Some theories of morphology assume morphemes that have no meaning. For representing such elements, e.g. the `EmptyMorph` instance *-u-* in the English adjective *factual*, an `EmptyMorpheme` instance can be created to account for the empty conceptual side of *-u-*[12].

Since morphemes are only meanings, i.e. mental representation of concepts, they are represented by abstract identities in order to be referable. This is done by `MorphemicGloss` instances which are provided for each of the 299 `Meaning` classes in the MMoOn Core model and which also apply to every language-specific schema instance derived from the Core model, e.g. `eng_schema:Singular mmoon:hasAbstractIdentity mmoon:MorphemicGloss_SG`. It has to be noted here, that the `hasMeaning` object subproperties can be used to describe `Word`, `Morph` and `Morpheme` resources in MMoOn Core, as has been shown in the Examples 6, 9 and 12 and seems to over-model the data. While these are just possibilities of describing the meaning of different linguistic elements, within a consistent MMoOn morpheme inventory it is sufficient to model the meaning on the `Morpheme` resources, because these are traceable through the data graph via the corresponding morphs to the word-forms and lexemes in which they occur.

Finally, a morph and its corresponding morpheme must be interrelated because they constitute a unity of a linguistic expression and its conceptualization. Example 13 illustrates the association between morphemes and morphs. The object property `hasRealization` which is inverse of `correspondsToMorpheme` is provided and links a morpheme to all morphs by which it is realized in a given language.

It has to be noted that so far – to my knowledge – no RDF dataset exists which contains morpheme resources as proposed in the MMoOn Core model. However, in linguistic field research and in the general practice of documenting the morphological level of languages, it is common to create interlinear glossed texts, which distinguish morph and morpheme resources in a similar way. While it might be effortful (but not impossible) to create morpheme resources as proposed in MMoOn from scratch or manually, the vocabulary could be useful for representing existing interlinear glossed text resources in MMoOn RDF.

## 5 Intersections and Issues of an OntoLex - MMoOn Alignment

As the previous sections illustrated, the conceptual overlap of the ontolex/decomp modules and the MMoOn Core model provides an auspicious basis for interconnecting both domain models. In order to align both vocabularies, several

[12]It depends on the choice of the dataset creator if empty morphemes are assumed. One could also assume a suffix *-ual* as being an allomorph to *-al*.

intersections of elements could be used to bring them into mutual agreement. Since ontology alignment and merging might cause "unforeseen implications" [5], this task should be solved together by the OntoLex and MMoOn community groups. Nonetheless, in what follows, elements are proposed which are assumed to be necessary for mapping in order to enable a consistent extension of OntoLex datasets with a MMoOn morpheme inventory and conversely.

1) `ontolex:LexicalEntry` and `mmoon:LexicalEntry`: These two classes are central in both domain models and are regarded as the the most important intersection because they are crucial for the interconnection of lexical entries and morph resources. The OWL property `owl:equivalentClass` could be an appropriate mapping choice, since it would allow to infer that all more specific `mmoon:LexicalEntry` subclasses are also subclasses of `ontolex:LexicalEntry`. With consideration of the use of MMoOn properties which have some of these subclasses in their domain and range restrictions, however, it is debatable if a stated equivalency between these two classes will be sufficient or if a separate mapping of each `mmoon:LexicalEntry` subclass might be required.

2) `ontolex:Affix` and `mmoon:Affix`: These two classes can be also mapped via `owl:equivalentClass`. This would allow to later classify `ontolex:Affix` resources for the more specific `mmoon:Affix` subclass types by remaining of the `ontolex:Affix` type at the same time.

3) `decomp:subterm` and `mmoon:isDerivedFrom`; `mmoon:isComposedOf`: The decomp module clearly favours the representation of compound words. Therefore, an interconnection of `mmoon:isDerivedFrom` and `mmoon:isComposedOf` as being subproperties of `decomp:subterm` would enable more specific interrelations of lexical entries in OntoLex if desired.

4) `ontolex:LexicalSense` and `mmoon:Sense`: A reuse of `ontolex:LexicalSense` resources for `mmoon:Stem` resources would facilitate the assignment of senses to stems a lot. Although, the `owl:equivalentClass` property could be used here as well, a more elegant solution would be the implementation of an axiom that automatically creates a link between the `ontolex:LexicalSense` resource (of a given `ontolex:LexicalEntry` instance) and the `mmoon:Stem` instance of which the lexical entry consists.

Even though more elements could be considered for an alignment, the proposed mappings already bear a significant impact for the use of present OntoLex and MMoOn datasets and advantages for future datasets as well. E.g. the considerable amount of linguistic categories and derivational meanings provided with MMoOn Core could be directly used for OntoLex. Moreover, the morphological segmentation of `ontolex:LexicalEntry` instances is easily describable with an aligned MMoOn Core model. Finally, the `ontolex:Affix` resources as part of a lexicon, would be enriched with information on the specific kind of affix in question, its interrelation to the lexical entries in which it occurs and the inflectional or derivational meaning it carries.

## 6 Conclusion

At the moment, the OntoLex and the MMoOn model coexist as two separate ontologies, even though both models exhibit a conceptual overlap in the representation of the LLD and MLD domains. This paper motivated an alignment of both models, since it could be shown that the ontolex/decomp modules are not sufficient to describe fine-grained MLD in such an extensive way as the the MMoOn Core model does. Therefore, the undertaken comparison of the capabilities of both models to represent MLD revealed intersecting elements of both vocabularies and proved that the MMoOn model is a suitable candidate for achieving extensibility of OntoLex datasets with MLD. Further, the paper pointed out intersecting elements for which mapping possibilities have been suggested and discussed. The aim of this paper was to propose a unification of both models. Now, it is up to the LLOD and OntoLex community to discuss and to decide whether the proposed alignment of these two linguistic domain models is desired and to work together on the realization of an OntoLex-MMoOn alignment. The author is convinced, that it would indeed enhance the exploitation of linguistic Linked Data in the Semantic Web world and would moreover contribute to the development of more coherent linguistic Linked Data datasets in general.

## 7 Acknowledgements

This papers research activities were partly supported and funded by grants from the EU's H2020 Programme ALIGNED (GA 644055) and the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) for the SmartDataWeb Project (GA-01MD15010B).

## References

1. Booij, G.: The grammar of words: An introduction to linguistic morphology. Oxford University Press (2012)
2. Chavula, C. and Keet, C.M.: Is lemon Sufficient for Building Multilingual Ontologies for Bantu Languages? In: OWLED, pp. 61-72 (2014)
3. Cimiano, P., Buitelaar, P., McCrae, J. and Stintek, M.: LexInfo: A declarative model for the lexicon-ontology interface. In: Web Semantics: Science, Services and Agents on the World Wide Web. Elsevier, pp. 29–51 (2011)
4. Haspelmath, M. and Sims, A.: Uderstanding morphology. Routledge (2013)
5. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In IJCAI-2001 Workshop on ontologies and information sharing. pp. 53–62 (2001)
6. Klimek, B., Arndt, N., Krause, S. and Arndt, T.: Creating Linked Data Morphological Language Resources with MMoOn - The Hebrew Morpheme Inventory. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (2016)
7. Lehmann, C.: Data in linguistics. In: The Linguistic Review. Vol. 21, pp. 175–210 (2004)

8. Lehmann, C.: Interlinear morphemic glossing. In: Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung. 2, pp. 1834–1857 (2004)
9. McCrae, J. et al.: The lemon cookbook. (2010)
10. McCrae, J., Spohr, D. and Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In: The semantic web: research and applications, Springer, 2011, pp. 245–259 (2011)