Representing Compounding with OntoLex. An Evaluation of Vocabularies for Word Formation Resources

Elena Benzoni[1], **Matteo Pellegrini**[2], **Francesco Dedè**[1], **Marco Passarotti**[2]
[1]Università degli Studi di Milano Statale, [2]Università Cattolica del Sacro Cuore
[1]Via Festa del Perdono, 7, 20122, Milano, Italy, [2]Largo Agostino Gemelli, 1, 20123, Milano, Italy
{elena.benzoni2@studenti.,francesco.dede@}unimi.it, {matteo.pellegrini,marco.passarotti}@unicatt.it

Abstract

This paper explores how compounds are represented in resources documenting word formation, and proposes ways to convert them into Linked Open Data using the OntoLex model. The ultimate purpose is to offer a broad empirical evaluation of which of the two OntoLex modules allowing for the representation of compounds – Decomp and Morph – fits best the different formats and theoretical approaches of the resources we examine. We show that the vocabulary of Decomp alone is rarely sufficient to account for all relevant facts; in almost all cases, it is necessary to resort to the vocabulary of Morph, either to reify the relation between compounds and their constituents or to represent specifically morphological information or other aspects. Special attention is devoted to the format of the Universal Derivations project: the modelling strategy that we propose can be applied to all resources harmonized in that format, potentially allowing for the conversion into Linked Open Data of a large amount of structured data.

Keywords: Compounding, Linked Open Data, Lexical Resources

1. Introduction and Motivation

After years spent developing digital resources for languages, the idea of not only collecting data published independently and in different formats in infrastructures (like CLARIN,[1] DARIAH,[2] and META-SHARE[3]), but also of making them structurally and semantically interoperable, has begun to gain ground. Principles and technologies of the Linked Open Data paradigm have been increasingly applied to achieve such interoperability. In this context, relations between data are represented explicitly by means of the RDF (Resource Description Framework) data model (Lassila and Swick, 1998), where information is coded in triples of the form subject-predicate-object, all the members of the triple being "resources" with their own URI (Uniform Resource Identifier), except for the object, that can also be a literal. Community efforts – such as the ones of the NexusLinguarum Cost Action[4] and the W3C consortium[5] – led to the development of ontologies and models for the modelling of both general purpose and specifically linguistic knowledge.

Among those models, OntoLex (McCrae et al., 2017), established itself as a *de facto* standard for the modelling of lexical resources in the framework of Linguistic Linked Open Data (Cimiano et al., 2020). One of the aspects of words on which lexical resources often provide information is their morphology. Starting from an already existing vocabulary designed to represent morphological information – namely, the Multilingual Morpheme Ontology (Klimek et al., 2021) – work has been conducted to release Morph (Klimek et al., 2019; Chiarcos et al., 2022b), a module designed to be fully integrated into the OntoLex core vocabulary and more specifically devised for the modelling of the kind of morphological information that is typically found in dictionaries, as well as in other kinds of lexical resources.

Originally, the morphological process of compounding was considered to be out of the scope of Morph: in the diagram sketched in Klimek et al. (2019), there is no dedicated class or property. This was motivated by the fact that another, already existing OntoLex module was considered to be capable to model the internal structure of compounds – namely, Decomp, which allows for the decomposition of complex lexical entries (both Multi-Word Expressions and compounds) in their constituents or components (see Subsection 2.2 below). At a later stage, however, it became clear that the Decomp vocabulary was not always sufficient for this purpose. Therefore, classes for compounding relations and rules were introduced in Morph alongside derivational ones (see Subsection 2.3 below).[6]

As a consequence, there are now two modules of OntoLex – Morph and Decomp – that can be used to represent compounding. This raises the question of which one is more appropriate in specific cases, and for what reasons. Such a question is made es-

[1] https://www.clarin.eu/.
[2] https://www.dariah.eu/.
[3] http://www.meta-share.org/.
[4] https://nexuslinguarum.eu.
[5] https://www.w3.org/.

[6] This history is for the most part not documented in published work, but it can be retraced in the minutes of the internal discussion of the group working on Morph within the NexusLinguarum Cost Action. These are all available at https://github.com/ontolex/morph/tree/master/minutes.

LREC-COLING 2024, pages 13958–13969
20-25 May, 2024.

pecially interesting by the fact that there is not only a plurality of lexical resources that provide information on compounds, using different formats and approaches, but also a plurality of possible theoretical views on how the process of compounding should be conceived in the architecture of grammar and the lexicon. The very concepts of "compound" and "compounding" are at the centre of a long-standing linguistic debate (Lieber and Štekauer, 2011), and many issues that are problematic for their theoretical definition also raise significant problems concerning their formal representation. Among these issues, there are two in particular that cannot be overlooked at all, namely the question of whether compounds are the outcome of a morphological or syntactic process (Gaeta and Ricca, 2009), and the question of whether compounds are formed starting from words or rather lexical morphemes (Scalise and Vogel, 2010). Each of these questions brings with it several other problems (just to mention one, the fuzzy boundary between proper compounds and other types of Multi-Word Expressions), which must be dealt with when devising resources for the formal (and computational) representation of complex words. Dealing with the more theoretical and general aspects of the compounding process, investigated with reference to as many languages as possible, is of the utmost importance even if one adopts a purely computational perspective, since it helps creating resources that, while devoted to a single language, may talk to each other and be linked together.

The goal of this paper is to offer a broad empirical evaluation of the relevance of classes and properties of the OntoLex vocabulary for the representation of compounding in relation to the plurality of lexical resources capturing derivational morphology and word formation processes. As mentioned above, within the OntoLex framework, both Decomp and Morph allow for the modelling of compounding. Hence, we first describe in Section 2 the OntoLex core model (Subsection 2.1) and the Decomp (Subsection 2.2) and Morph (Subsection 2.3) modules. We then analyze in Section 3 some resources where compounds are included, to understand which module best suits each resource and to test the general guidelines against real-world data. We focus on resources that are not included in the UDer (Universal Derivations) project (Kyjánek et al., 2020) in Subsection 3.1 and move to ones that are included therein in Subsection 3.2. We deal with the specific data formats of standalone resources in Subsection 3.2.1, while in Subsection 3.2.2 we propose a more general modelling strategy that can be applied to any resource harmonized in the UDer format. In Section 4, we conclude that Decomp alone is rarely sufficient for the representation of compounding, and it is almost always necessary to resort to Morph too, and we sketch possibilities for future work.

2. Reference Vocabularies

2.1. The OntoLex Core Model

The general aim of the OntoLex vocabulary is to make it possible to provide rich linguistic grounding for ontologies, allowing for the representation of morphological, syntactic and semantic properties of lexical entries. Besides the core model, additional modules have been released to account for more specific information: Synsem for syntax and semantics,[7] Decomp (already mentioned in Section 1) for the decomposition of lexical entries,[8] Vartrans for variation and translation relations,[9] Lime for metadata,[10] Lexicog for lexicographic information;[11] other modules are currently being developed, such as the one for frequency and attestation in corpora, FrAC,[12] and Morph (already mentioned in Section 1).

As shown in Figure 1,[13] the OntoLex core model revolves around the central class `ontolex:LexicalEntry`, for which sub-classes[14] are introduced to cover not only for usual lexemes (`ontolex:Word`), but also for Multi-Word Expressions (`ontolex:MultiwordExpression`) and affixes (`ontolex:Affix`) . For each lexical entry, information can be provided on both formal and semantic aspects. Regarding formal aspects, there is a class `ontolex:Form` that can be used for the different concrete realizations of lexical entries – e.g., inflected wordforms. These can be mapped to the corresponding entry by means of either the property `ontolex:canonicalForm`, when dealing with the citation form, or `ontolex:otherForm` otherwise. Regarding semantic aspects, dedicated classes and properties are defined for senses (cf. the class `ontolex:LexicalSense`, the inverse properties `ontolex:sense` and `ontolex:isSenseOf` to map entries to

[7] https://www.w3.org/2016/05/ontolex/#syntax-and-semantics-synsem.

[8] https://www.w3.org/2016/05/ontolex/#decomposition-decomp.

[9] https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans.

[10] https://www.w3.org/2016/05/ontolex/#metadata-lime.

[11] https://www.w3.org/2019/09/lexicog/.

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

[13] See the W3C community report published at https://www.w3.org/2016/05/ontolex/ for additional details.

[14] These are indicated by arrows with a white head in this diagram and the following ones.

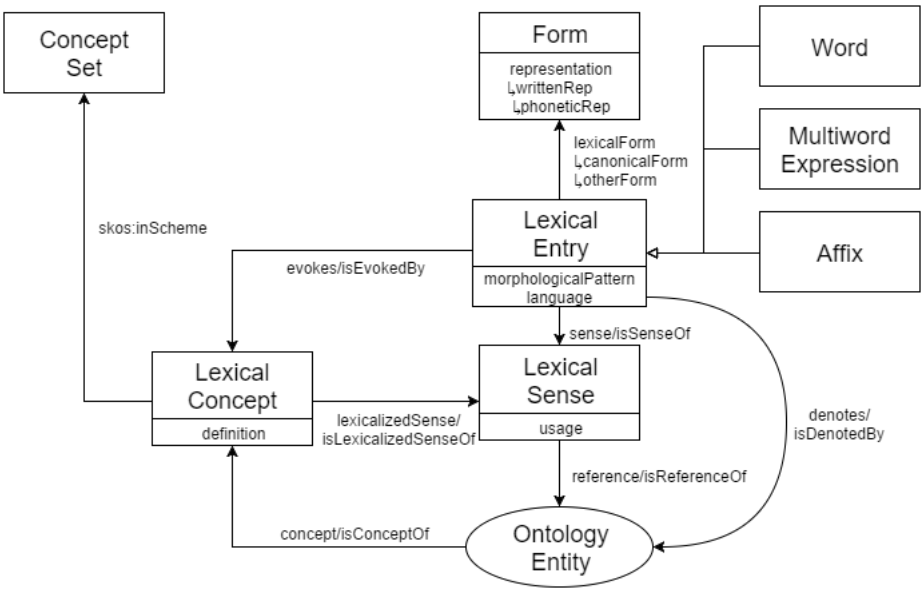

Figure 1: The OntoLex core model.

senses) and concepts (cf. the class `ontolex:LexicalConcept`, the inverse properties `ontolex:evokes` and `ontolex:isEvokedBy` to map entries to concepts, the inverse properties `ontolex:lexicalizedSense` and `ontolex:isLexicalizedSenseOf` to map concepts to senses, the inverse properties `ontolex:concept` and `ontolex:isConceptOf` to map concepts to ontology entities).

2.2. The OntoLex-Decomp Module

The OntoLex-Decomposition module (Decomp) aims to account for the structure of complex lexical entries, both Multi-Word Expressions and compounds. This can be achieved through two different properties: `decomp:subterm` or `decomp:constituent`. More technically, the range of the former property is the class `ontolex:LexicalEntry` (i.e., only individuals of that class can be used as its object), while the range of the latter property is the class `decomp:Component`, whose individuals may correspond to (`decomp:correspondsTo`) a lexical entry (`ontolex:LexicalEntry`), a semantic role (`synsem:Frame`) or a grammatical one (`synsem:Argument`), as shown in Figure 2.[15]

The choice between these two properties depends on the characteristics of the data under consideration: for instance, to represent segmentations that refer to the corresponding canonical forms of the lexical entries involved, `decomp:subterm` can be used, while to represent segmentations that identify constituents as they appear in the complex form, `decomp:constituent` seems to be the most suitable option.

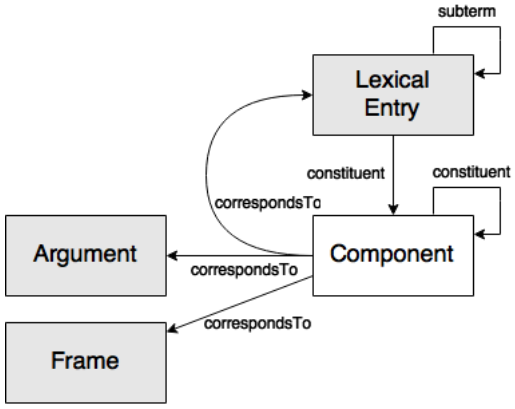

Figure 2: The OntoLex-Decomp module.

2.3. The OntoLex-Morph Module

The OntoLex-Morphology module (Morph), illustrated in Figure 3,[16] aims to model the morphological information expressed in dictionaries and other lexical resources. Morph is integrated into the OntoLex core model via the class `morph:Morph`, which is defined as a subclass of `ontolex:LexicalEntry`. Within its architecture, there are two main components, one for description (extensional morphology) and one for generation (intensional morphology) (Chiarcos et al., 2022b),

[15] See `https://www.w3.org/2016/05/ontolex/#decomposition-decomp` for additional documentation.

[16] We refer to version 4.13, as described in Chiarcos et al. (2022b); see `https://github.com/ontolex/morph/tree/master` for the current state of the model.

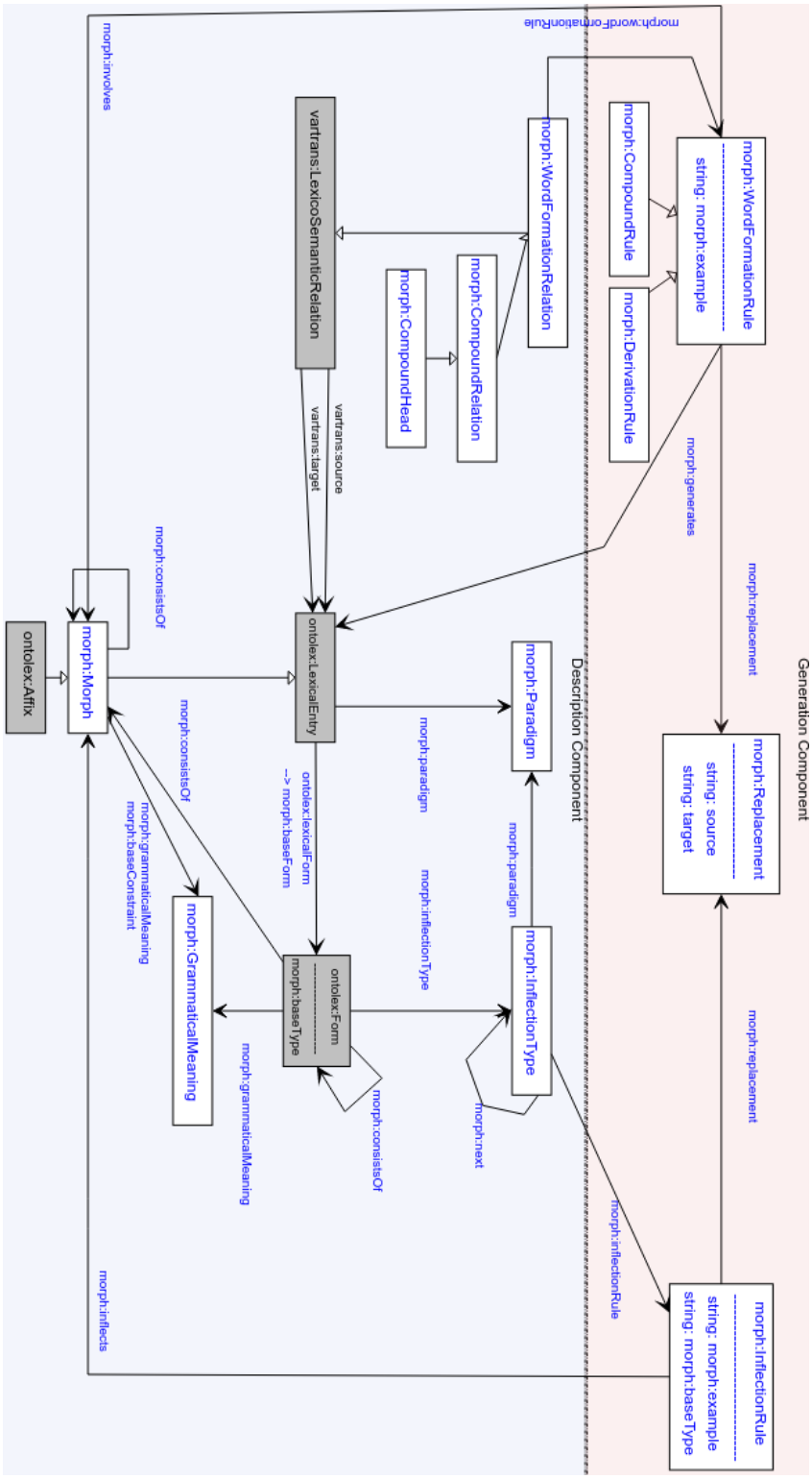

Figure 3: The OntoLex-Morph module, draft version 4.13 (2022).

both modelling derivational and inflectional morphology. As for the inflectional section of the module, it is important to note that it is possible to account for the segmentation of forms into morphs via the property `morph:consistsOf` (cf. the application to Morpholex in Subsection 3.1). Here, however, we will focus on the derivational section.

Classes and properties for the representation of compounding were introduced into Morph when it became clear that the Decomp vocabulary was not always sufficient for this purpose. For instance, a proper modelling of compounding in the Word Formation Latin lexicon (Litta and Passarotti, 2019) requires a reification – i.e., a dedicated class –

of the relation between compounds as wholes and each of the members they are composed of, reification that cannot be expressed with Decomp (Pellegrini et al., 2022). Hence, the class `morph:CompoundRelation` was introduced as a subclass of `morph:WordFormationRelation`, which is, in turn, a subclass of a class from the Vartrans module – `vartrans:LexicalRelation`. These classes share the same structure: their individuals – i.e., relations – use properties from Vartrans to connect a source element (`vartrans:source`) and a target one (`vartrans:target`) – i.e., the two lexical entries involved. Thus, rather than resorting to Morph in some cases and Decomp in others, derivation and compounding can be treated with the same vocabulary, as it seems more adequate for resources like Word Formation Latin, where the two processes are distinct yet treated similarly.

An analogous bipartite structure is found in the generation component: word formation relations are connected to the class `morph:WordFormationRule` through a homonymous property `morph:wordFormationRule`, and there are dedicated subclasses for compounding and derivation rules (`morph:CompoundRule` and `morph:DerivationRule`), respectively.

3. Applications to Resources

Out of the 90 resources for word formation listed by Kyjánek (2020), we identified 8 that explicitly document compounding. Among those, here we will not consider the following ones: Word Formation Latin (Litta and Passarotti, 2018), because it is extensively treated in previous work (Pellegrini et al., 2022; Chiarcos et al., 2022b); WiktiWF,[17] because it is currently under development and there is neither a stable release nor any publication; the Russian Morphological Database,[18] because modelling options similar to the ones devised for other resources can be applied in that case too; and the Czech DeriNet (Vidra et al., 2021), because its format corresponds to the target format of the UDer project, that will be considered separately in Subection 3.2.2.

We will thus describe the resources that are not included in UDer in Subsection 3.1 – namely, Morpholex and Morphonette – and the ones that are included therein in Subsection 3.2 – namely, CroDeriV and Golden Compound Analyses. For each of these resources, we will evaluate which module between Decomp and Morph is best suited to model the data concerning compounding, considering the architecture of the two modules and the technical and linguistic characteristics of the resources – more precisely, the data format and the type of morphological analysis.[19]

3.1. Resources not in UDer

MorphoLex (Sánchez-Gutiérrez et al., 2018; Mailhot et al., 2020) consists of two datasets, one documenting English (Sánchez-Gutiérrez et al., 2017) and the other French (Mailhot et al., 2019). Both datasets can be considered as morpheme-oriented (Kyjánek, 2020), in that they present a segmentation of their entries which reduces the elements composing them to their citation form. Therefore, compounds can be considered as lexical entries composed of other lexical entries, and hence represented with Decomp, through the property `decomp:subterm`, as shown in the listing below.[20]

```
:ice a ontolex:LexicalEntry.
:pick a ontolex:LexicalEntry.
:icepick a ontolex:LexicalEntry;
    decomp:subterm :ice, :pick.
```

However, also wordforms other than the citation form are found – e.g., plural forms such as *icepicks*. These items fit the definition of the `ontolex:Form`, rather than `ontolex:LexicalEntry`, since they are grammatical realizations of lexical entries. Hence, as shown in the listing below, to represent their segmentation it is necessary to resort to Morph, as Decomp can only be used to specify the structure of lexical entries.

```
:icepicks a ontolex:Form;
    morph:consistsOf :ice, :pick.
```

The MorphoLex case is relevant because it shows that even if a resource is characterised by a morphological analysis based on segmentation, it is not necessarily the case that Decomp is the most suitable module for modelling its data.

Morphonette (Hathout, 2017) is a paradigm-oriented resource documenting French (Hathout, 2011). Morphonette groups morphological data into a morpho-lexical network originated by the crossing of two kinds of morphological relations, namely family relations and series relations, based on the sharing of segmental material between pairs of words. A family relation is the relation between words that share the same base – e.g., between

[17] https://github.com/lukyjanek/wiktionary-wf.

[18] http://courses.washington.edu/unimorph/userInterface/rvnkur.php.

[19] An overview of the modules used for each resource and the motivations behind the choice is given in the Appendix, also for the resources not discussed in this work, on which see Benzoni (2023).

[20] In our examples, we do not give a namespace for resources that are not in any existing vocabulary or dataset, but need to be introduced for the data at hand (e.g., the lexical entry `:icepick`).

Fr. *modifier*$_V$ and *modification*$_N$ 'change'. A series relation is the one between words that are formed by means of the same morphological process – e.g. between Fr. *modification*$_N$ and *rectification*$_N$ 'rectification'.

Given the central role played by relations in this resource, it is reasonable to reify them using the vocabulary of Morph. It is important to note that in Morphonette compounding and derivation are not explicitly distinguished: hence, the more general class `morph:WordFormationRelation` can be used for both. Since family and series relations are specific kinds of word formation relations, it seems reasonable to establish two subclasses of morph:WordFormationRelations to represent them, as shown in the listing below.

```
:FamilyRelation rdfs:subClassOf
    morph:WordFormationRelation.
:SeriesRelation rdfs:subClassOf
    morph:WordFormationRelation.

:aberroscope a ontolex:LexicalEntry.
:microscope a ontolex:LexicalEntry.

:rel a :SeriesRelation;
    vartrans:source :aberroscope;
    vartrans:target :microscope.
```

This case is notable for two reasons: first, Morphonette is a resource that nicely exemplifies that a strict correlation might occur between theoretical approaches and the technical aspects of a resource; second, precisely for the aforementioned correlation, this example proves the flexibility of Morph in front of the plurality of possible theoretical views, concerning not only compounds but morphology in general.

3.2. Resources in UDer

3.2.1. Standalone Resources

Golden Compound Analyses (GCA) (Vodolazsky and Petrov, 2021) is a resource documenting Russian containing compounding rules for the training, test and validation of a compound splitter (Vodolazsky and Petrov, 2021). Hence, GCA is a resource based on word formation rules, involving the input and output PoS of each part of the compound, or (for non-lexical items like interfixes) their morphological status – e.g., rule754([noun + ITFX] + adj → adj) for the example we provide below in this subsection. Given the presence of rules that relate the members of compounds, a reification of compounding relations, to which those rules can be connected, seems more appropriate for the modelling of GCA, thus requiring to resort to the vocabulary of Morph. The data are structured in this way: firstly, there is a rule, followed by a compound that matches the given rule and then its output PoS is provided. Finally, the compound elements are indicated through their citation form and respectively associated with their input PoS. For instance, *zasukhoustoychivyy* 'drought-resistant' is coded to be an adjective, formed according to rule 754 ([noun + ITFX] + adj → adj), and its members are the noun *zasukha* 'drought' and the adjective *ustoychivyy* 'resistant'. Based on this structure, the compound as a whole and each of its members can be considered as instances of `ontolex:LexicalEntry` and rules as instances of `morph:CompoundRule`, as shown in the listing below.

```
:zasukhoustoychivyy a ontolex:LexicalEntry.
:zasukha a ontolex:LexicalEntry.
:ustoychivyy a ontolex:LexicalEntry.

:rule754 a morph:CompoundRule;
    morph:generates :zasukhoustoychivyy.

:rel1 a morph:CompoundRelation;
    vartrans:source :zasukha;
    vartrans:target :zasukhoustoychivyy;
    morph:wordFormationRule :rule754.

:rel2 a morph:CompoundRelation;
    vartrans:source :ustoychivyy;
    vartrans:target :zasukhoustoychivyy;
    morph:wordFormationRule :rule754.
```

The source of interest in GCA is plainly the presence of word formation rules. Besides Word Formation Latin, among the resources considered, this is the only one that offers the opportunity to provide a practical example of modelling that requires a reification due to the presence of rules.

CroDeriV (Šojat et al., 2023) is a lexeme-oriented resource for Croatian, that also provides a morpheme-oriented double segmentation of its entries: there is a "deep" segmentation – where grammatical morphemes are reduced to their underlying form and lexical morphemes to their citation form – and a "surface" segmentation – where the components are given in the segmental shape in which they are found in the entry (Filko et al., 2020). While Decomp is a suitable option for the former segmentation, it is not sufficient for the latter, where we find also morphological information – i.e., linking elements – that requires the use of Morph. Furthermore, it seems advisable to introduce a reification of the relation between a compound and its members. This is motivated on the one hand by the fact that compounding and derivation are treated in a uniform fashion (compounding is simply seen as a derivational relation with more than one parent), thus making it undesirable to treat the former with Decomp and the latter with Morph. On the other hand, for compounds, one of the two relations is identified as primary: a compound belongs to the same family of its primary constituent, and not to the same family of the other constituent. To rep-

resent this information, we propose to introduce a specific subclass of `morph:CompoundRelation`, namely `:MainRelation`, as shown in the listing below.

```
:dvolik a ontolex:LexicalEntry;
    decomp:subterm :dva;
    decomp:subterm :lik;
    ontolex:canonicalForm :dvolik_form.

:dvolik_form a ontolex:Form;
    morph:consistsOf :dv, :o, :lik.

:rel1 a morph:CompoundRelation,
        :MainRelation;
    vartrans:source :lik;
    vartrans:target :dvolik.

:rel2 a morph:CompoundRelation;
    vartrans:source :dva;
    vartrans:target :dvolik.
```

The modelling of CroDeriV data gave us a chance to show a case where the vocabulary of Morph is needed to convey specifically morphological information about compounds – namely, the linking element. Moreover, even if there might be some redundancy in the coexistence of Decomp and Morph, the joint use of the two modules shows that they can be used to model different pieces of information regarding compounds and their representation.

3.2.2. The UDer Format

The Universal Derivations (UDer) project (Kyjánek et al., 2021) aims to harmonise word formation data originating from different resources to standardise their annotation based on a single format, so that data published independently, documenting different languages and using diverse formats, can be more easily compared (Kyjánek et al., 2020). More precisely, the UDer project shares the same goal of Universal Dependencies (UD) (de Marneffe et al., 2021) and Universal Morphology (UniMorph) (Batsuren et al., 2022), even if they establish annotation standards for different linguistic aspects, i.e., respectively, derivational morphology, syntax and inflectional morphology.


```
167945.1    lichopřeslen#NNI??-----A---?  Lichopřeslen  NOUN
Animacy=Inan&Gender=Masc&Loanword=False
End=8&Morph=lichopře&Start=0&Type=Root|End=12&Morph=slen&Star
t=8&Type=Suffix    167945.0
Sources=149597.0,167945.0&Type=Compounding
{"corpus_stats":   {"absolute_count":   72,   "percentile":
81.64787843087431, "relative_frequency": 2.8724850505704583e-
08,  "sparsity":  7.541742222856092},   "is_compound":  true,
"techlemma": "lichopřeslen"}
```


Figure 4: An example of data in UDer format.

The target format of the data harmonised in UDer is the one of DeriNet 2.0 (Vidra et al., 2019), a lexeme-oriented resource documenting Czech, and it is inspired by the CoNLL-U format (de Marneffe et al., 2021). As illustrated in Figure 4, for each lexical entry, relevant aspects of derivational morphology concerning individual lexemes are annotated via key-value pairs. Due to the nature of the project, the file format needs to be flexible enough to allow for the representation of resource-specific data and information. However, given the relevance of the UDer project, we here outline the modelling of those pieces of information that can be systematically found in the file containing harmonised data. Each aspect will be explained by taking the compound *lichopřeslen* 'verticillaster' from DeriNet as an example.

Out of the key-value pairs listed by Vidra et al. (2019), for the modelling of the UDer format, we will consider the PoS Tag, the Morphological Features, the Main Parent ID and the Parent Relation. Within the DeriNet file, there is also a key to represent morphological segmentation, but since it is not present in other files collecting harmonised data from resources documenting compounding, we will not consider it here – although it can be easily represented with Morph. We will also exclude the JSON column, because it is used for the coding of unpredictable, resource-specific data, so it is difficult to devise a modelling strategy that would be appropriate for all resources.[21] For the PoS Tag, the Universal Part-Of-Speech Tagset (Petrov et al., 2012) is used (as in UD), whereas, for features, there is only a partial overlap between the UDer Morphological Features set and the Universal Features set used in UD, since the former has some *ad hoc* additional features (e.g., to code whether an entry is a loanword). The *ad hoc* nature of those additional features hinders the usage of LexInfo, the vocabulary OntoLex relies on to represent morpho-lexical features (Cimiano et al., 2011). Hence, we propose to model the content of the columns for PoS Tags and Morphological Features by introducing two corresponding subclasses of `oa:Annotation`, from the Web Annotation data model, illustrated in Figure 5 .[22] In this model, annotations connect a target (the item to be annotated) and a body (the content of the annotation). This allows for the flexibility that we need: when a URI is available for the content of the annotation (as happens for UD features), we point to that URI; when this is not the case (e.g., for the coding of loanwords), we point to a literal (i.e., the string found in the column). The listing below provides examples that illustrate this difference.

[21] However, since the information is coded in such a common and structured format as JSON, it can easily be recovered for further modelling.

[22] https://www.w3.org/TR/annotation-model/.

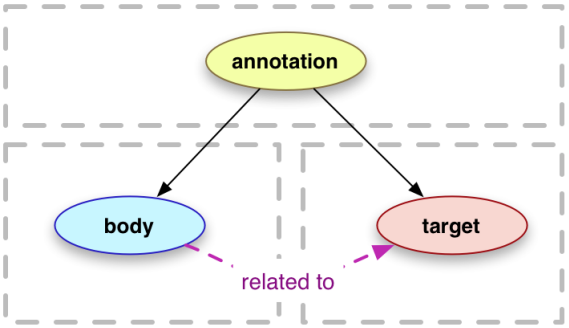

Figure 5: The Web Annotation data model

```
:lichopreslen a ontolex:LexicalEntry.

:MorphologicalFeatures rdfs:subClassOf
    oa:Annotation.

:lichopreslen_morph_feat
    a :MorphologicalFeatures;
    oa:hasBody
        <https://universaldependencies.
            org/cs/feat/Gender#masc-
            masculine-gender>,
        "Loanword=False";
    oa:hasTarget :lichopreslen.
```

As for the Parent Relation field, there are four types of relations, namely Variant (for allographs), Derivation, Conversion and Compounding. We propose to introduce *ad hoc* classes that can be integrated into OntoLex and its modules by establishing `rdfs:subClassOf` relations with the already existing classes. Since the `:TypeVariant` relation is not a morphological relation, it can be integrated as a subclass of `vartrans:LexicalRelation`. `:TypeDerivation` and `:TypeConversion` can be integrated as subclasses of `morph:WordFormationRelation`. On the other hand, for the UDer Compounding relation, there is no need to introduce a specific subclass, as a class specifically for compounds is already defined in Morph.

There is a last aspect that is worth mentioning, namely the identification of the Main Parent ID. The solution we propose is similar to the one described in Subsection 3.2 above for the same issue in CroDeriV. Every UDer relation has a main parent. Thus, we propose to introduce another subclass of `vartrans:LexicalRelation`, named `:MainParentRelation`. One of the two relations introduced for a compound will be at the same time an instance of `morph:CompoundRelation` and of `uder:MainParentRelation`, as shown in the listing below.

```
:rel1 a morph:CompoundRelation;
    vartrans:source :lichy;
    vartrans:target :lichopreslen.

:rel2 a morph:CompoundRelation,
        :MainParentRelation;
    vartrans:source :preslen;
    vartrans:target :lichopreslen.
```

Therefore, with only a few extensions, the vocabulary of Morph proves to be also capable of modelling the data format of UDer – an advantageous possibility, given the many resources that are being released in that format, that once again proves the flexibility of Morph.

Indeed, it is important to observe that a way to convert the data of UDer into RDF triples using the vocabulary of OntoLex and Morph has already been sketched in Chiarcos et al. (2022a) for German. Compared to their proposal, the one put forward in this work is on the one hand less precisely tailored on the details provided by a resource for a given language, but on the other hand it has a wider scope. Rather than outlining a way to convert the data of a single resource in UDer for a specific language, with all the details that it encodes, we propose a more general modelling solution that, while leaving aside some details, has the potential to be applied to any resource in UDer format in a semantically richer fashion, by virtue of the use of different (sub-)classes corresponding to different kinds of relations, and of the possibility to model information of different nature with `oa:Annotation`. In any event, our proposal is also broadly compatible with the one by Chiarcos et al. (2022a) in that both use (sub-classes of) the class `morph:WordFormationRelation` to express the relations between morphologically related lexical entries.

4. Conclusions and Future Work

In this paper, we have reviewed several resources documenting compounding, focusing on how they can be modelled as Linked Open Data using the OntoLex vocabulary (cf. Subsection 2.1). For each case, we have highlighted the peculiar characteristics that require resource-specific modelling, regarding both technical (e.g., data format) and theoretical aspects. In doing so, we have provided a broad empirical evaluation of the two different modules that can be used for the representation of compounding – namely, Decomp (cf. Subsection 2.2) and Morph (cf. Subsection 2.3).

On the one hand, we have seen that Decomp alone is often not sufficient to model compounding. This confirms on a larger scale what has been suggested in previous work focusing on specific word formation resources for individual languages, namely Latin (Pellegrini et al., 2022) and German (Chiarcos et al., 2022a). On the other hand, Morph proves to be flexible enough to be capable to cover

for a remarkable variety of resources. This highlights the advantages of introducing an OntoLex module specifically devised for the treatment of morphological information, and shows that having the possibility of using also this module to model compounding yields benefits that overcome the redundancy implicit in having two different vocabularies (Morph and Decomp) for the same phenomenon (compounding).

Furthermore, the evaluation presented in this paper suggests best practices to decide which vocabulary is more appropriate to handle specific cases. To summarize, Decomp should be preferred whenever it is sufficient – that is to say, in cases where compounding can be simply modelled as a decomposition of lexical entries, with no loss of relevant morphological information. The use of Morph should be restricted to cases where such decomposition-based modelling is not sufficient. Existing resources for compounding have provided examples of aspects that require the use of Morph to be modelled efficiently, namely:

- segmentation of forms, rather than lexical entries (cf. Morpholex, Subsection 3.1);
- reification of word formation relations (cf. Morphonette, Subsection 3.1);
- representation of word formation rules (cf. GCA, Subsection 3.2.1);
- modelling of specifically morphological information, like linking elements (cf. CroDeriV, Subsection 3.2.1);
- unified treatment with derivation (cf. CroDeriV, Subsection 3.2.1).

Given this state of affairs, one may wonder – as two anonymous reviewers did – why not just resort to Morph, making it the only vocabulary whose usage is recommended for resources of this kind. The reasons why such an option is not adequate can be retraced in the history of the development of OntoLex itself. As was hinted in Section 1, the Decomp module was part of the Ontolex vocabulary from the very beginning, to allow for a modelling of complex lexical entries, and their decomposition into sub-parts, not only for compounds, but also for complex lexical entries of different kind (chiefly, Multi-Word Expressions). On the other hand, the Morph module has been developed at a later stage, with the modelling of morphology (chiefly, inflection and derivation) in mind. Compounding is obviously potentially relevant in that it is part of the morphology of languages. A key principle of Linked Open Data is to reuse whenever possible and introduce new vocabulary only when necessary. As a consequence, since Decomp classes and properties were already available and explicitly envisaged as to be used (also) for compounding, the initial idea was leaving compounding out of the scope of Morph, and just resorting to Decomp for that. The fact that Decomp proved to be in many cases insufficient for a proper modelling of compounding motivates the addition of classes and properties for that purpose in Morph. However, classes and properties of Decomp cannot be disregarded, as they are needed for other purposes that are not covered by Morph, being outside the scope of morphology. Therefore, having only one vocabulary for compounding is not an option: this is what motivates the need for guidelines and best practices for the usage of Morph or Decomp in that area. Our proposal to use Decomp whenever sufficient and Morph whenever necessary is consistent with the history of the development of Ontolex, and it has the additional benefit of backward-compatibility with previous resources that might have used Decomp for compounds. For instance, Decomp has been used for the modelling of GermaNet (Chiarcos et al., 2022a). Such a modelling would still be Ontolex-compliant despite the presence of classes and properties of Morph usable in the same context.

As a further step, by using the vocabulary of Morph and only a few, minor extensions, we have been able to propose a model for the data format of the UDer project (cf. 3.2.2). This model can be used not only for resources containing compounds, but also for all the other resources harmonised in UDer, paving the way for future work to convert and publish UDer data as Linked Data. This would be particularly useful considering the remarkable amount of data available in that format.[23] Furthermore, it would facilitate interoperability with other datasets in standard formats for which a widely applicable procedure for conversion to RDF has already been proposed – e.g., UD (Chiarcos et al., 2020; Mambrini et al., 2022) or Unimorph (Chiarcos et al., 2020, 2022a) – allowing for interesting crossed queries – e.g., identifying all compounds that are assigned a given dependency relation in one or more UD treebanks, or extracting all inflected forms listed in a Unimorph dataset for entries that display a specific word formation relation. From a more general perspective, having also the UDer format easily translatable to RDF, alongside UD and Unimorph, is advantageous both for the Linked Open Data community – that can enlarge its coverage of large-scale data regarding different levels of analysis for diverse languages – and for the communities behind these enterprises – that can have their data communicate with each other under the same framework.

[23] At the time of the writing of this paper, 31 resources were included (https://ufal.mff.cuni.cz/universal-derivations).

5. Bibliographical References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. Unimorph 4.0: Universal morphology. *arXiv preprint arXiv:2205.03608*.

Elena Benzoni. 2023. La rappresentazione delle parole composte in Linguistic Linked Open Data tra teoria e pratica. Master's thesis, Università degli Studi di Milano Statale.

Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2020. Annotation interoperability for the post-ISOCat era. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5668–5677, Marseille, France. European Language Resources Association.

Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2022a. Unifying Morphology Resources with OntoLex-Morph. A Case Study in German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4842–4850, Marseille, France. European Language Resources Association.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022b. Computational Morphology with OntoLex-Morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Matea Filko, Kresimir Sojat, and Vanja Stefanec. 2020. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistic*, 115:83–104.

Livio Gaeta and Davide Ricca. 2009. Composita solvantur: Compounds as lexical units or morphological objects? *Rivista di Linguistica*, 21(1):35–70.

Nabil Hathout. 2011. Morphonette: a paradigm-based morphological network. *Lingue e linguaggio*, 2/2011:245–264.

Bettina Klimek, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann. 2021. MMoOn Core–the Multilingual Morpheme Ontology. *Semantic Web*, 12(5):813–841.

Bettina Klimek, John P McCrae, Julia Bosque-Gil, Maxim Ionov, James K Tauber, and Christian Chiarcos. 2019. Challenges for the representation of morphology in ontology lexicons. *Proceedings of eLex 2019*, pages 570–591.

Lukáš Kyjánek. 2020. Harmonisation of language resources for word-formation of multiple languages. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Lukás Kyjánek, Zdenek Zabokrtskỳ, Magda Sevcikova, and Jonás Vidra. 2020. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, 115:5–30.

Ora Lassila and Ralph R. Swick. 1998. Resource Description Framework (RDF) Model and Syntax Specification.

Rochelle Lieber and Pavol Štekauer. 2011. Introduction: Status and Definition of Compounding. In *The Oxford Handbook of Compounding*. Oxford University Press.

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, pages 224–239. De Gruyter, Berlin / Boston.

Hugo Mailhot, Maximiliano A Wilson, Joël Macoir, S Hélène Deacon, and Claudia Sánchez-Gutiérrez. 2020. MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior research methods*, 52:1008–1025.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. The index Thomisticus treebank as linked data in the LiLa knowledge base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4022–4029, Marseille, France. European Language Resources Association.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, and Martina Verdelli. 2022. Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension. *The Prague Bulletin of Mathematical Linguistics*, 119:67–92.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. 2018. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior research methods*, 50:1568–1580.

Sergio Scalise and Irene Vogel. 2010. Why compounding? In *Cross-disciplinary issues in compounding*, pages 1–18. John Benjamins.

Jonáš Vidra, Zdeněk Žabokrtskỳ, Magda Ševčíková, and Lukáš Kyjánek. 2019. DeriNet 2.0: towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89.

Daniil Vodolazsky and Hermann Petrov. 2021. Compound splitting and analysis for russian. *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology*, pages 149–157.

6. Language Resource References

Hathout, Nabil. 2017. *MORPHONETTE. French Morphological Network*. PID http://redac.univ-tlse2.fr/lexiques/morphonette_en.html.

Kyjánek, Lukáš and Žabokrtský, Zdeněk and Vidra, Jonáš and Ševčíková, Magda. 2021. *Universal Derivations v1.1*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL). PID http://hdl.handle.net/11234/1-3247. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Litta, Eleonora and Passarotti, Marco. 2018. *Word Formation Latin (WFL)*. CIRCSE Research Centre, Università Cattolica del Sacro Cuore. PID http://hdl.handle.net/20.500.11752/OPEN-531. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Mailhot, Hugo and Wilson, Maximiliano A. and Macoir, Joël and Deacon, S. Hélène and Sánchez-Gutiérrez, Claudia H. 2019. *MorphoLex-FR*. PID https://doi.org/10.3758/s13428-019-01297-z.

Sánchez-Gutiérrez, Claudia H. and Mailhot, Hugo and Deacon, S. Hélène and Wilson, Maximiliano A. 2017. *MorphLex-EN*. PID https://doi.org/10.3758/s13428-017-0981-8.

Vidra, Jonáš and Žabokrtský, Zdeněk and Kyjánek, Lukáš and Ševčíková, Magda and Dohnalová, Šárka and Svoboda, Emil and Bodnár, Jan. 2021. *DeriNet 2.1*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (UFAL). PID http://hdl.handle.net/11234/1-3765. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vodolazsky, Daniil and Petrov, Hermann. 2021. *Golden Compound Analyses*. PID https://github.com/s231644/rucompoundsplitter.

Šojat, Krešimir and Srebačić, Matea and Pavelić, Tin. 2023. *CroDeriV*. PID http://croderiv.ffzg.hr/.

Appendix. Summary table of resources documenting compounding and modules used for their modelling

Resource	*Module(s) used*	*Motivation*
CroDeriV	Decomp and Morph	"deep" segmentation→ Decomp "surface" segmentation → Morph Need for a reification of the compounding relation → Morph (more than one source → `morph:CompoundRelation`)
DeriNet / UDer	Morph	Need for a reification of the compounding relation (more than one source → `morph:CompoundRelation`)
Golden Compound Analyses	Morph	Need for a reification of the compounding relation
MorphoLex	Decomp and Morph	Segmentation of Lexical Entries → Decomp Segmentation of Forms → Morph
Morphonette	Morph	Need for a homogeneous representation of derivation and compounding
Russian Morphological Database	Decomp and Morph	Segmentation → Decomp Need for a homogeneous representation of derivation and compounding → Morph
WiktiWF	Morph	Need for a homogeneous representation of derivation and compounding
Word Formation Latin	Morph	Need for a reification of the compounding relation (more than one source → `morph:CompoundRelation`)