How to Solve Issues of Representing Morphological Data with MMoOn Core

Modelling Suggestions for a Morphology Module for OntoLex-Lemon

Bettina Klimek
Institute for Applied Informatics, Leipzig
KILT Research Group

OntoLex Face2Face Meeting
5th November 2018, Institute for the Dutch Language in Leiden

Content

1. Aim of the Talk
2. The MMoOn Core Ontology (the short story)
3. How to model morphological issues with MMoOn
    - 3.1. Modelling the position of morphs
        - 3.1.1. Infixation
        - 3.1.2. Internal modification
    - 3.2. Modelling stem allomorphy
    - 3.3. Modelling derivation
        - 3.3.1. Derivation 1
        - 3.3.2. Conversion
        - 3.3.3. Derivation 2
    - 3.4. Modelling complex word forms
4. Challenges for an OntoLex Morphology Module

1. Aim of the Talk

1. Initiate community work on morphology module

2. Show how morphological data can be represented with MMoOn Core

3. Collect feedback for the OntoLex morphology module

2. The MMoOn Core Ontology *(the short story)*

What is MMoOn Core?

MMoOn → **M**ultilingual **Mo**rpheme **On**tology

- language-independent vocabulary to represent morphological language data
- upper model that unifies language-specific datasets, so called MMoOn morpheme inventories
- first and only existing comprehensive domain model for morphology

Why did I create it?

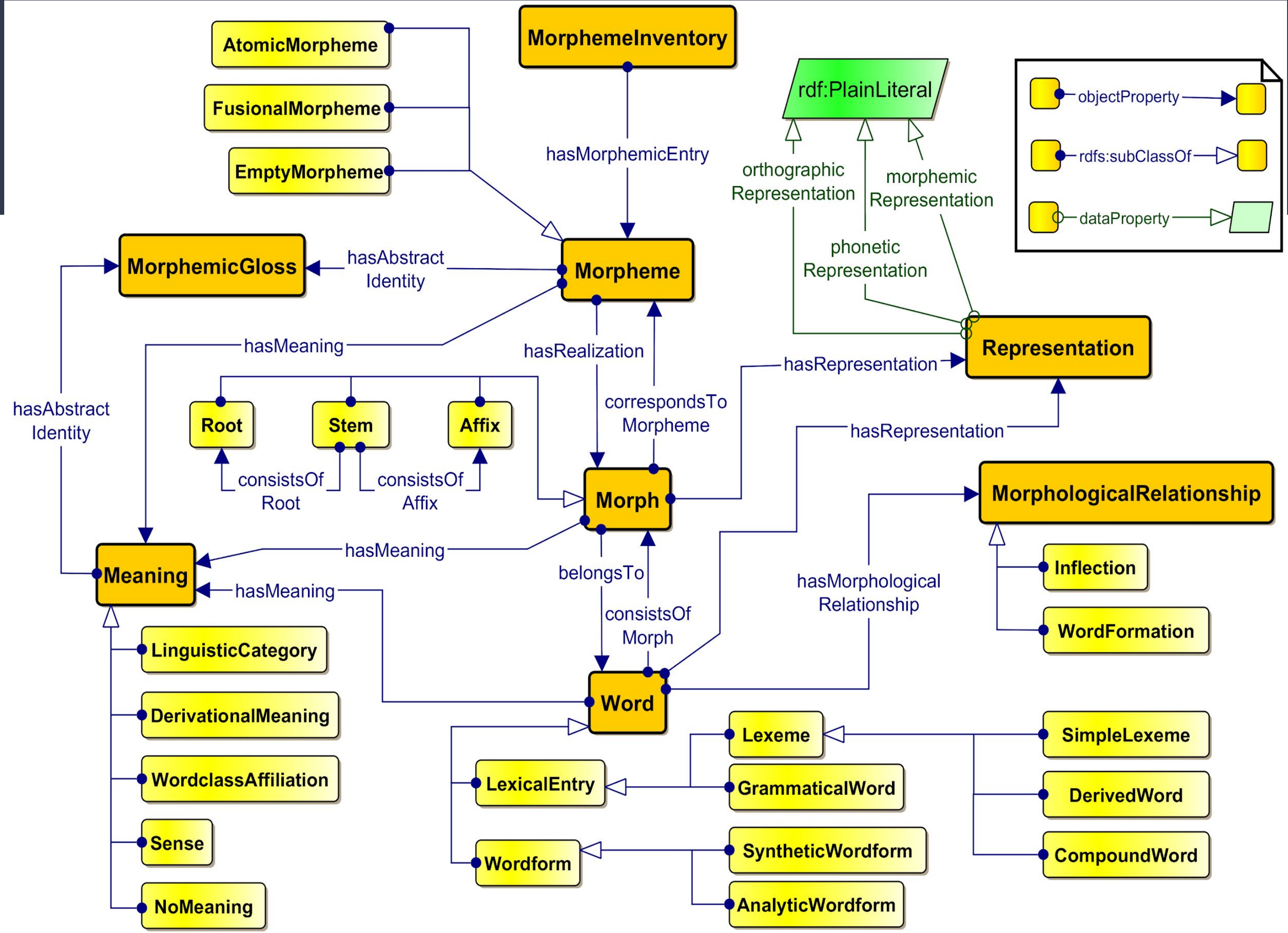
To enable the representation of (not possible with OntoLex-Lemon*):

- morphemes and morphs
- derivational and inflectional morphology
- relation between lexemes and their wordforms
- morphological segmentation

WARNING!

MMoOn Core is quite complex and fine-grained because it aims at linguists(!) as users (they have/produce the data 🤷‍♀️), which tend to define and specify every linguistic element they describe while avoiding to reach any general agreement that could be used as a shared basis for modelling language data.

*Klimek, Bettina (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*. 2017.



431 classes

37 object properties

5 datatype properties

600 individuals

MMoOn Core main classes definitions:

Primary language data: Language data which originates from a certain text compilation or could be applied to any text or token in order to identify the word-forms, morphs and morphemes of the morpheme inventory.

Morph: A morph is a **concrete realization of a single morpheme** which usually results from segmentation. A morph resource **describes only the perceptible side of a morpheme,** i.e. the significans. As such it is **not directly associated with a meaning** but with a corresponding morpheme resource. The mmoon core vocabulary, however, allows statements such as :Morph :hasMeaning :Meaning in case Morpheme resources are not yet documented.

Morpheme: The morpheme class contains the smallest **meaning-bearing units** of a language. These comprise all semantically or grammatically distinct parts which are identifiable by segmentation of the morphs of which a word or a word-form consists.

LexicalEntry: A lexical entry is a word as it appears as an entry in a dictionary. It can be a lexeme or a grammatical word. All lexical entries that inflect have a **holistic abstract sense representing the core meaning shared by a set of closely-related word-forms**. The lexical entry can be one of the word-forms which is chosen as the representative of the inflectional paradigm of the lexeme.

Wordform: A word-form is an inflectional variant of a lexical entry.

Representation: A linguistic representation of a word or morph.

MMoOn Core main classes definitions:

Secondary language data: The kind of data which enables the description of the primary language data.

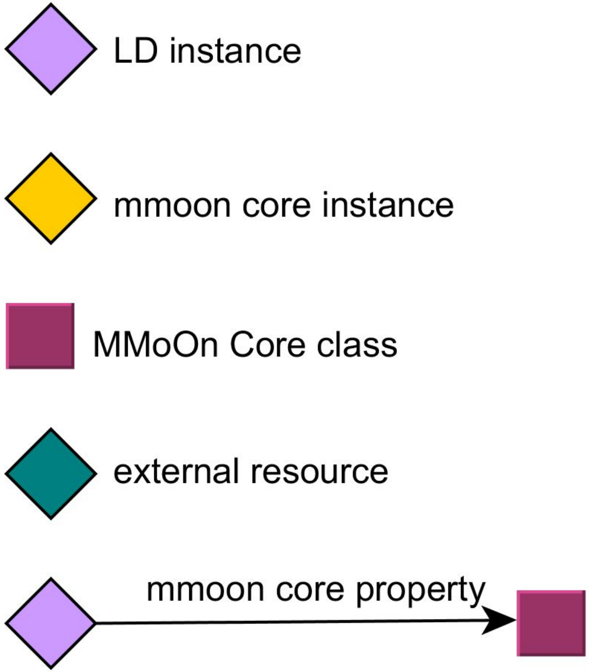
Meaning: This class comprises a wide range of meanings a word, morph or morpheme can be associated with, e.g. **linguistic categories, word-class affiliation, (lexical) senses, derivational meanings**.

MorphemicGloss: The gloss is the **abstract identity of a morpheme** and/or a meaning. It serves as a metalinguistic representation of (mostly morphological) meanings.

MorphologicalRelationship: Is the r**elationship between word-forms of a lexical entry (inflection) or the relationship between lexical entries of a word family (derivation and compounding)**. [Haspelmath and Sims: Understanding morphology. 2002:18]

MorphemeInventory: The morpheme inventory is the object that contains morphemic entries. It is specified for the natural language it describes.

3. How to model morphological issues with MMoOn



Primary Language Data

Secondary Language Data (instances)

3.1 Modelling the position of morphs

I1: The phones making up a morpheme don’t have to be contiguous
Inflection may cause a stem to break up or change. Morphology may occur at any point in the stem.

Lakhota verbs:

- lówan ‘he sings’ => **wa-**lówan ‘I sing’
- máni ‘he walks’ => ma-**wá-**ni ‘I walk’

Irish nouns: 'cat' => 'cat'
'a c**h**at' => 'his cat'
'a **g**cat' => 'their cat'

→ 3.1.1 Modelling infixation

I2: The form of a morpheme doesn’t have to consist of phones
Morpheme may alter the stem rather than adding to it.

German nouns:

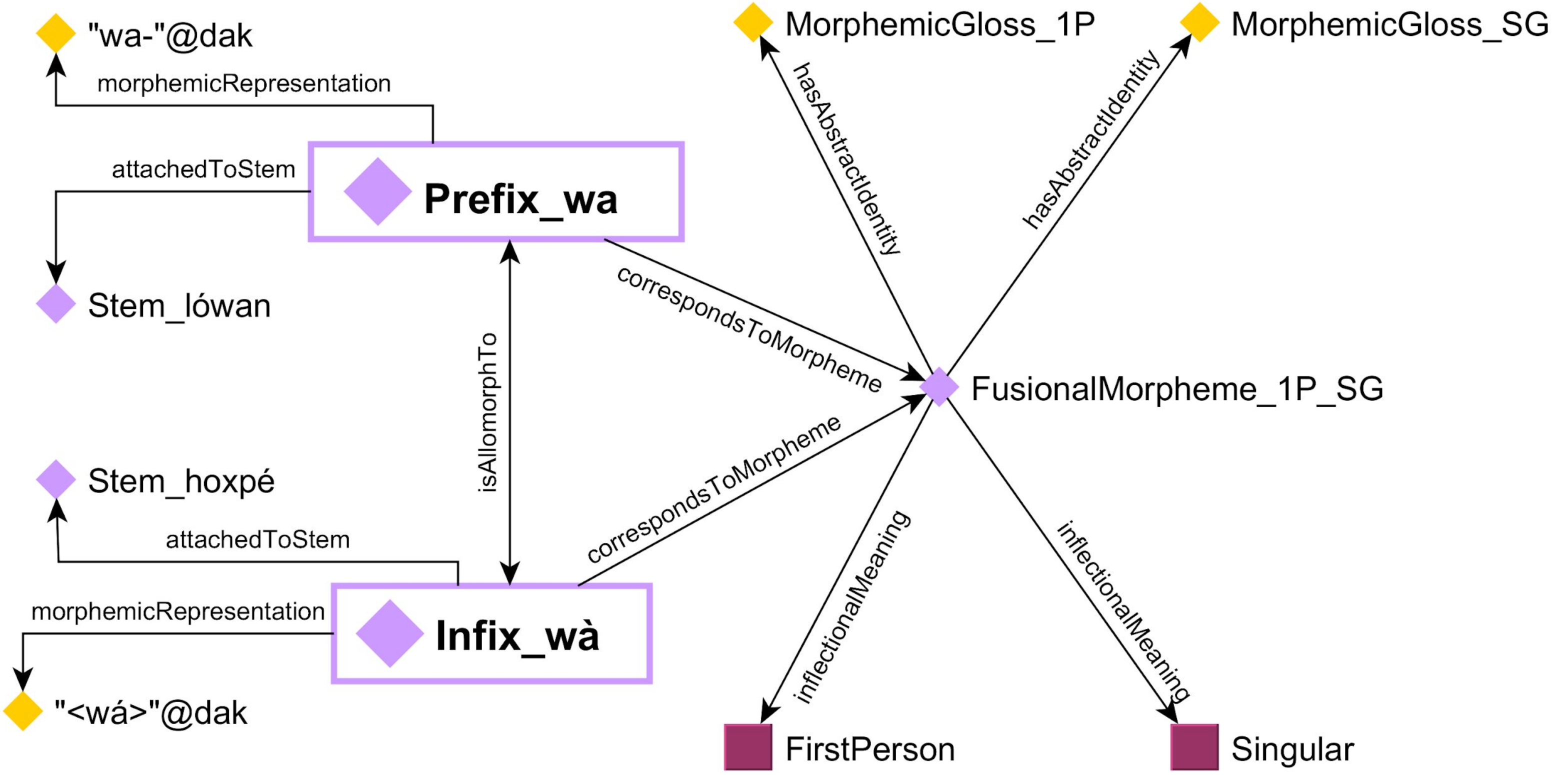
- 'M**u**tter' (mother) => 'M**ü**tter' (mothers)

→ 3.1.2 Modelling internal modification

The position of prefix, suffix and circumfix in relation to the stem is clear per definitionem.

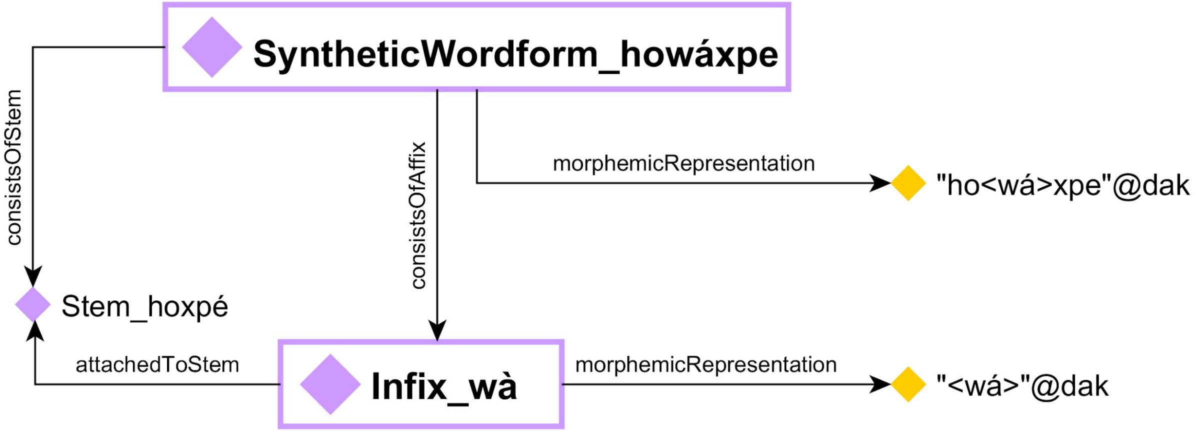
BUT :

How do we know at which position the infix is inserted into the stem?



3.1.1 Modelling infixation

MMoOn solution:
model infix position in morphemic representation of the wordform resource.



Other idea:
create vocabulary for inner word positions and list elements for every word resource with rdf list property. (danger of instance overload)

E.g.

SbInfix	Infix	SaInfix
ho	wá	xpe

3.1.1 Modelling infixation

Lakhota verbs:

- lówan ‘he sings’ => **wa-**lówan ‘I sing’
- máni ‘he walks’ => ma-**wá-**ni ‘I walk’

→ the same morpheme (1P.SG) but different forms (prefix and infix)

Irish nouns: 'cat' => 'cat'
'a c**h**at' => 'his cat'
'a **g**cat' => 'their cat'

→ different morphs (infix <h> and prefix g-) and different morphemes (3P.SG.OBJ and 3P.PL.OBJ)

3.1.1 Modelling infixation

Example:
German

- 'Mutter' (mother) => 'Mütter' (mothers)

Lango (a Nilo-Saharan language of Uganda). These examples are all different inflected forms of the verb ‘to stop’. They all agree with a first person singular subject (‘I’), but differ in their aspect. The only difference between the first two forms is in the tone associated with the final syllable.

- àgíkò ‘I stop (something), perfective’
- àgíkô ‘I stop (something), habitual’
- àgíkkò ‘I stop (something), progressive’

Problem:
How to represent a morph that that entails a process?!

a) u-->ü
b) ü

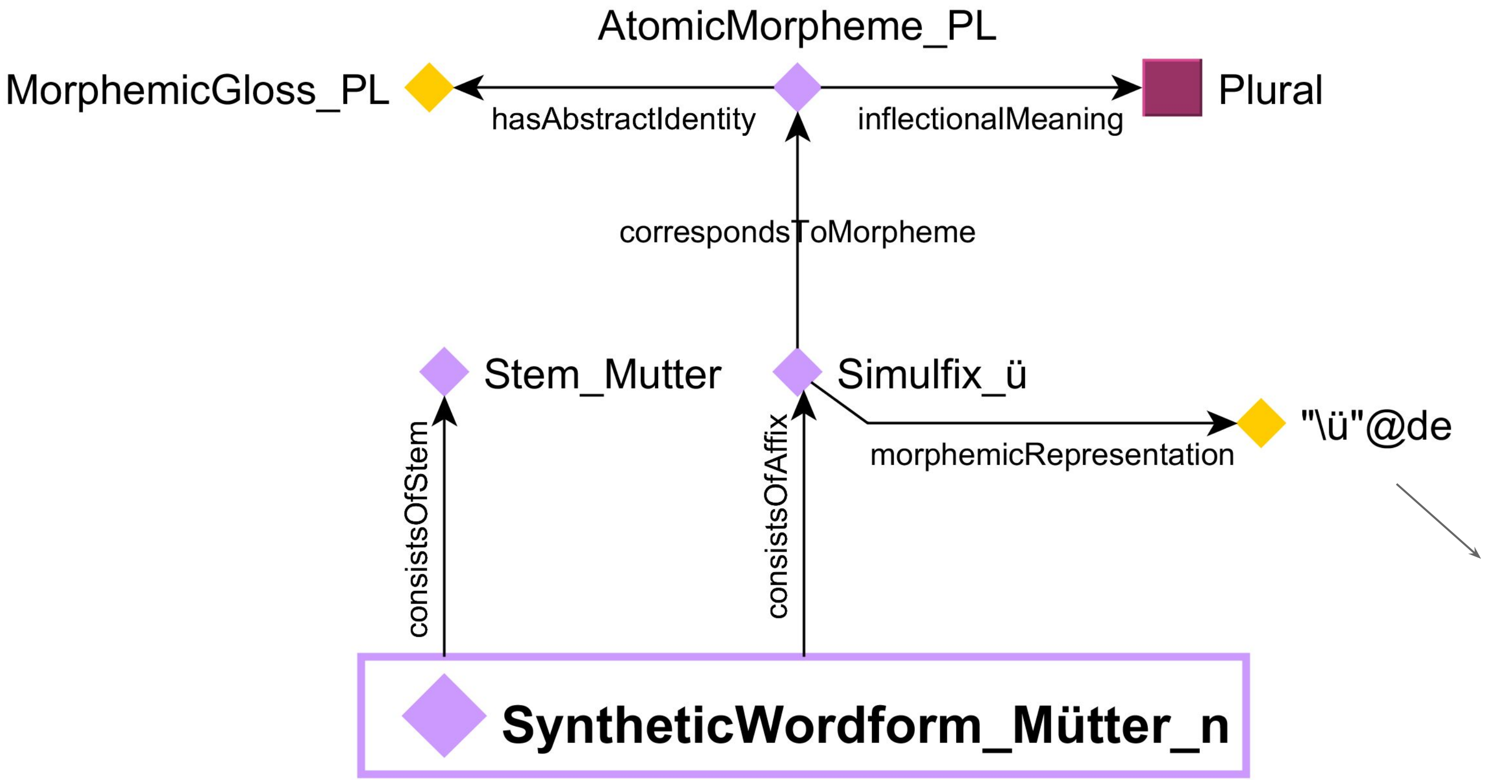
Representing elements and processes of internal modifications

“Internal modification is a morphological process which produces an alteration in the root or stem itself to express inflectional or derivational categories.”

Elements that express internal modification are called replacive morph or simulfix. These are like infixation in not being peripheral to the base, but they differ from it in that the grammatical meaning in question is not associated with a single string of segments which, if subtracted, leaves the base.

German: *M**u**tter* ‘mother’ → *M**ü**tter* ‘mothers’: IMG (L1) Mutter\PL; IMG (L2) mother\PL
Lango: *àgík**ò*** ‘I stop (something), perfective’ → *àgík**ô*** ‘I stop (something), habitual’ → *àgíkk**ò*** ‘I stop (something), progressive’:
IMG (L1) àgíkò\1SG.PFV , àgíkô\1SG.HABIT , àgíkkò\1SG.PROG IMG (L1) stop\1SG.PFV , stop\1SG.HABIT . stop\1SG.PROG

3.1.2 Modelling internal modification



MMoOn does not represent processes, only data!

PHOIBLE?
LIAM?
Fahad (SWRL rules)?

German
Mütter\ü
‘mother\PL’

3.1.2 Modelling internal modification

3.2 Modelling stem allomorphy

I3: The form of a morpheme (root or affix) can be sensitive to its morphological context
The suffix may cause a change in the stem of the original word to phonetic/orthographic rules.

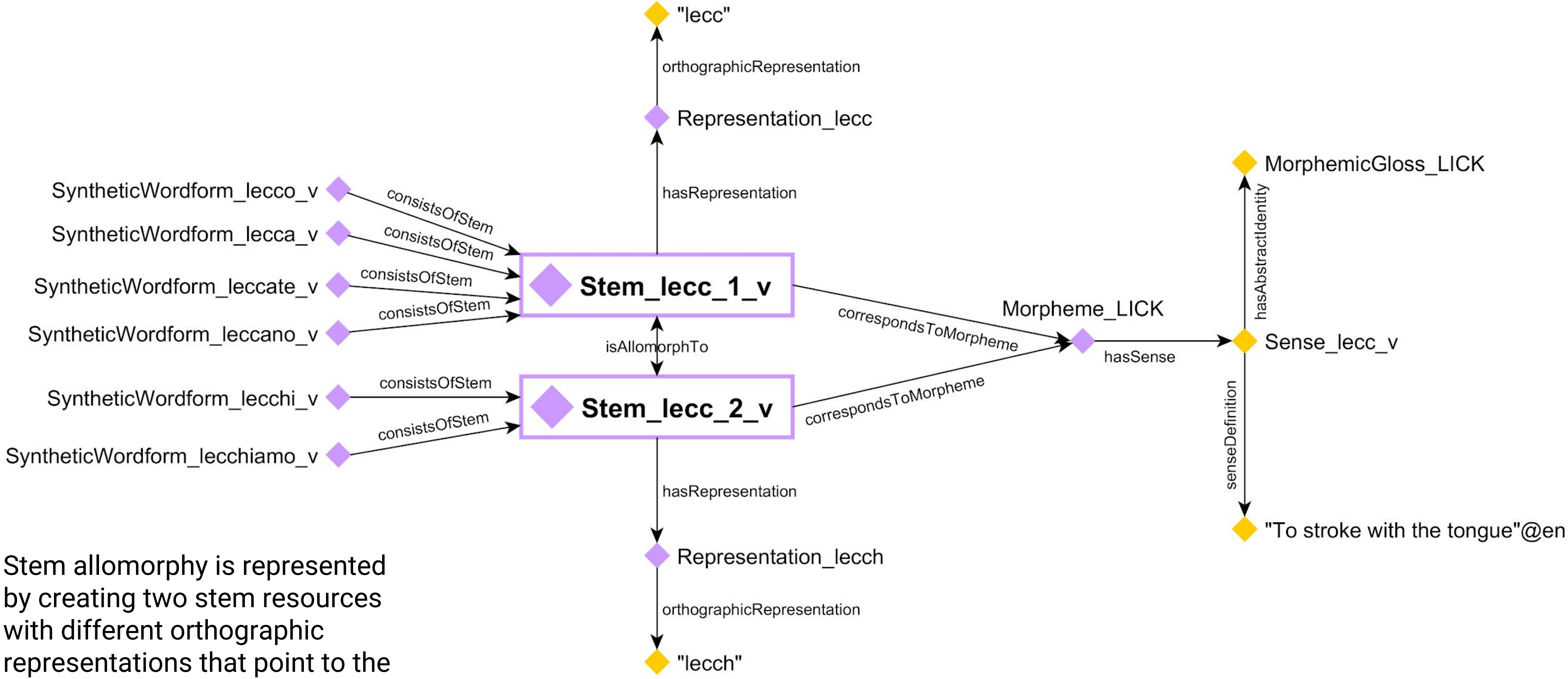
Italian verbs: stem inserts 'h' to preserve hard /k/ sound at the end of 'leccare', in two forms of the present tense:

- 'lecc + o' => '**lecc**o' (I lick)
- 'lecc + i' => '**lecch**i' (you lick)
- 'lecc + a' => '**lecc**a' (he licks)
- 'lecc + iamo' => '**lecch**iamo' (we lick)
- 'lecc + ate' => '**lecc**ate' (ye lick)
- 'lecc + ano' => '**lecc**ano' (they lick)

→ 3.2 Modelling stem allomorphy

“An allomorph of a morpheme is one of the morphs which instantiate the morpheme.”

The two forms (i.e. the morphs) *lecc* and and *lecch* of the stem *lecc* occur in complementary distribution depending on the vowel of the suffix.



Stem allomorphy is represented by creating two stem resources with different orthographic representations that point to the same morpheme.

3.2 Modelling stem allomorphy

3.3 Modelling derivation

I6: Morphology crosses part-of-speech boundaries
Morphological processes can turn one part-of-speech into another, effectively creating a distinct LexicalEntry.

English:

- "to play" (verb) => "played" (adjective)
- "to play" (verb) => "the playing" (noun)

I7: Morphology affects the meaning of words
Morphological processes may cause the meaning of the word to change in a systematic manner.

Diminutives create a new noun with a meaning of being smaller, this could be modelled by means of adding a small classes to the meaning of a noun.

features/ derivation types	**word class change**	**affixal marking**	**additional derivational meaning**
Conversion Ex.: play (v) → play (n)	+	- (zero-morph)	-
Derivation 1 Ex.: play (v) → playing (n)	+	+	-
Derivation 2 Ex.: book (n) → booklet (n) play (v) → player (n)	+ -	+	+

The morphological processes that create new lexemes, e.g. derivation and compounding, are modelled as subclasses of the class `MorphologicalRelationship`. It is possible to model these classes as subclasses of the respective wordclass classes. I.e. `DeverbalNoun` as subclass of `Noun`.

Overview of derivation types that can be represented with MMoOn.

3.3 Modelling derivation

I6: Morphology crosses part-of-speech boundaries
Morphological processes can turn one part-of-speech into another, effectively creating a distinct LexicalEntry.

English:

- "to play" (verb) => "played" (adjective)
- "to play" (verb) => "the playing" (noun)

→ 3.3.1 Derivation 1

I7: Morphology affects the meaning of words
Morphological processes may cause the meaning of the word to change in a systematic manner.

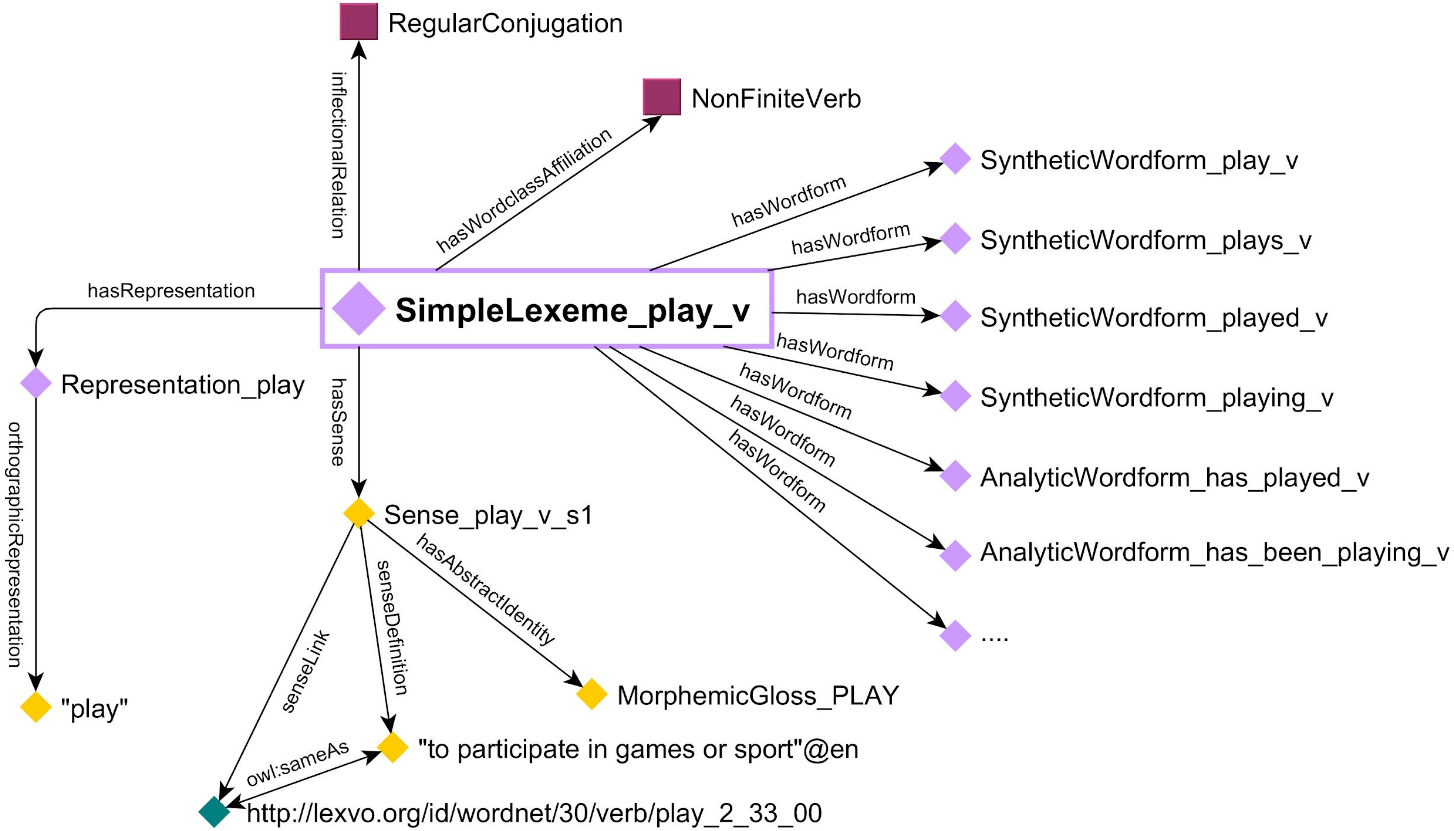
Diminutives create a new noun with a meaning of being smaller, this could be modelled by means of adding a small classes to the meaning of a noun.

→ 3.3.3 Derivation 2

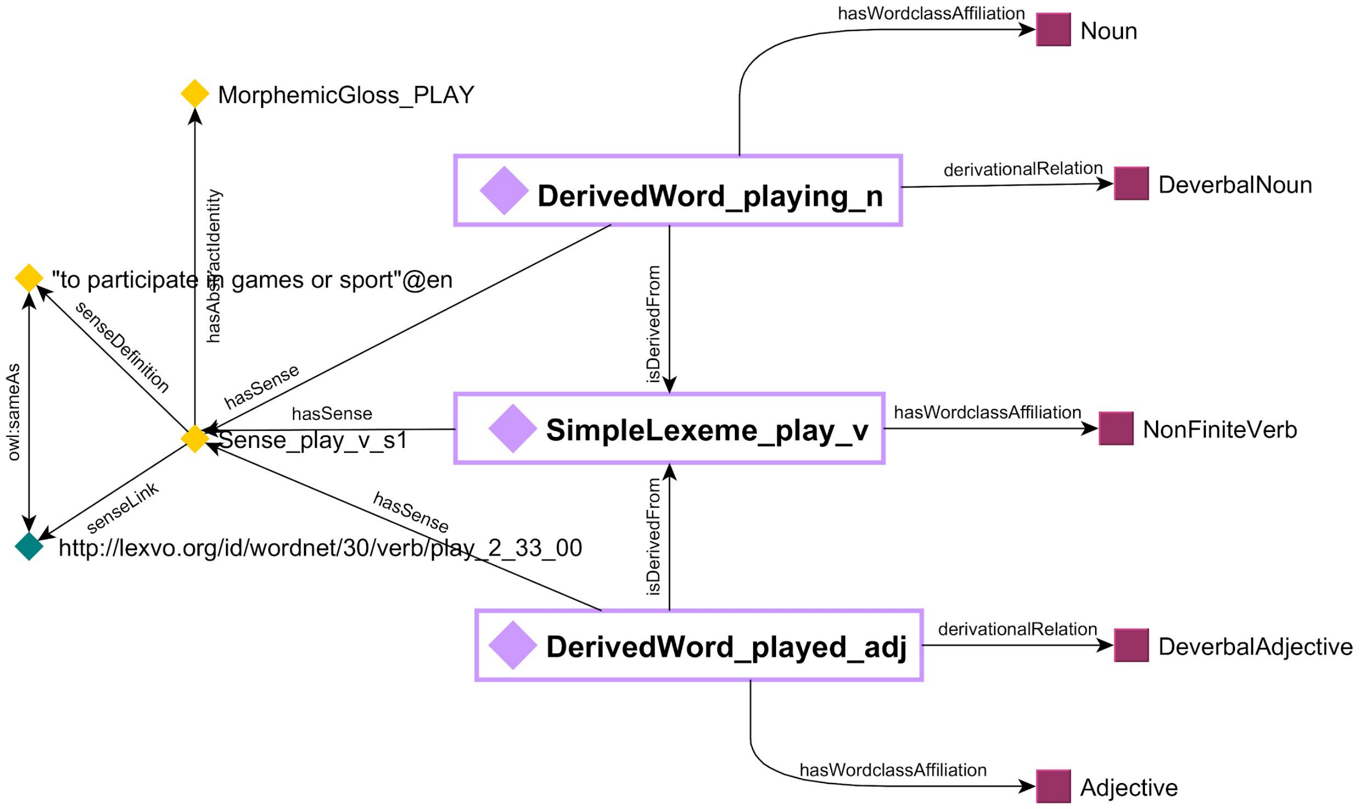
English:

- "to play" (verb) => "the play" (noun)

→ 3.3.2 Conversion



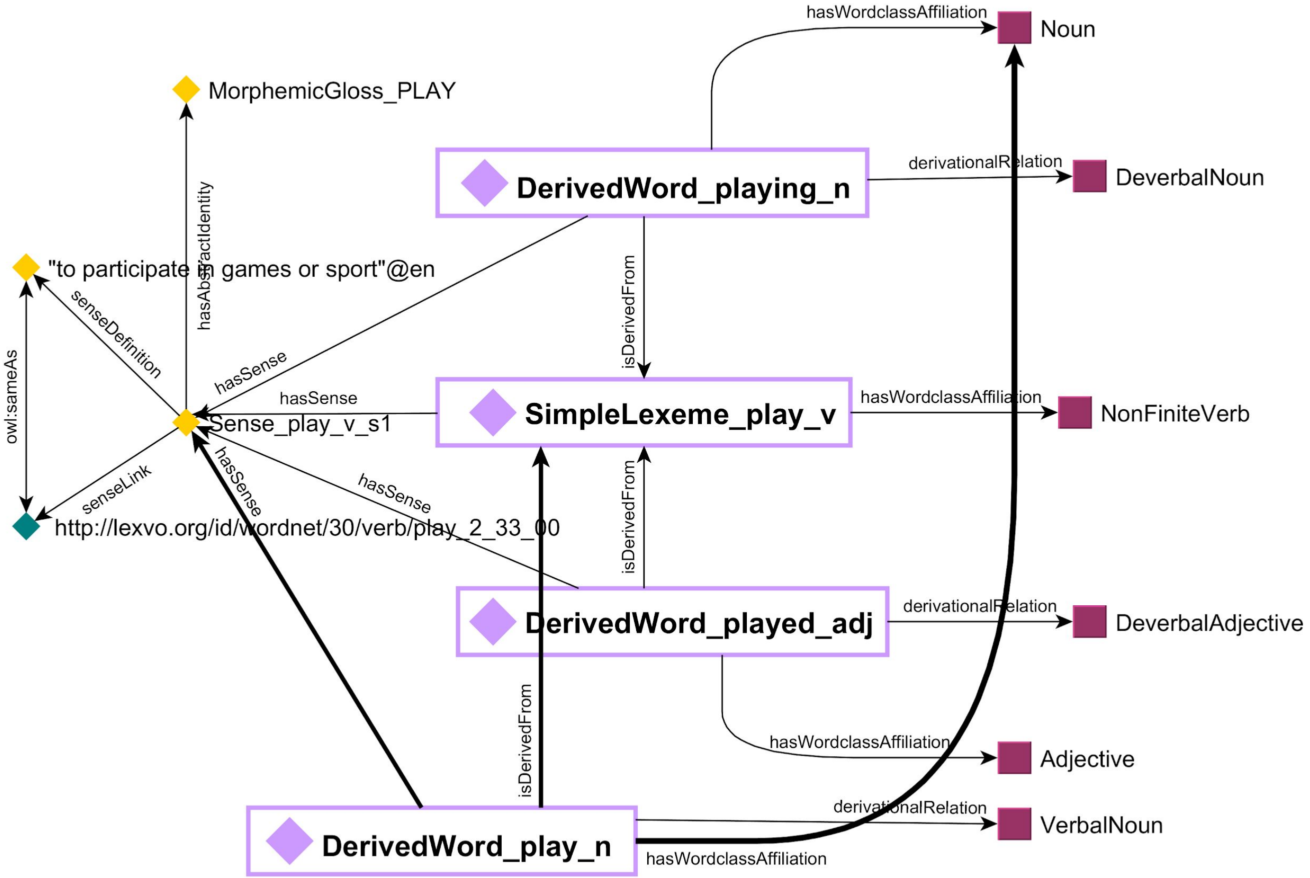
Representation of inflectional information for the simple lexeme verb *play*.



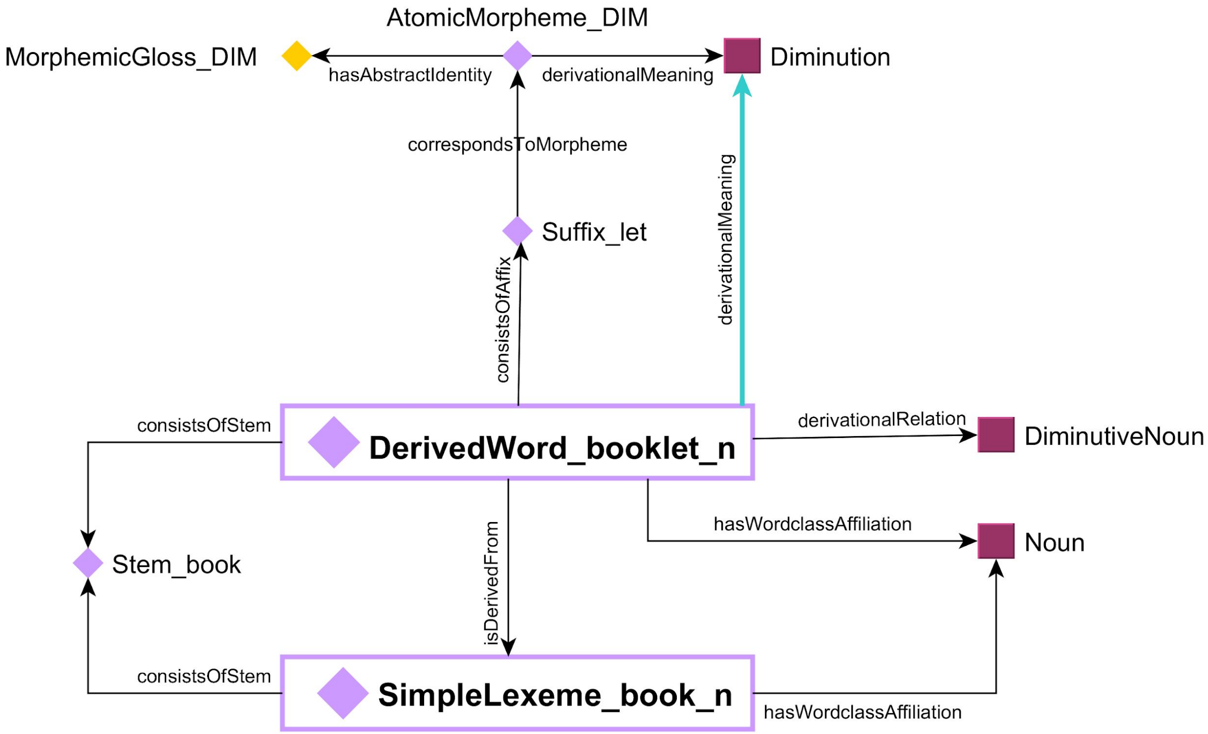
Remember:
`DerivedWord` and `SimpleLexeme` are subclasses of `LexicalEntry`

In MMoOn:
The meaning of the derived word is a combination of the sense defined for the simple lexemes they are derived from plus its word class affiliation.

3.3.1 Modelling Derivation 1



3.3.2 Modelling Conversion



The shortcut interrelating the lexical entry instance directly with the meaning class causes the loss of specifying the meaning at the actual element it encodes (i.e. only the suffix *-let*) and the loss of allomorph identification (e.g. *-ling* as in *duckling*).

`DiminutiveNoun` is a subclass of `DenominalNoun`.

3.3.3 Modelling Derivation 2

3.4 Modelling complex wordforms

I5: The morphosyntax of a language describes how the morphemes in a word affect its combinatoric potential
Words may combine in potentially unbounded manner, such that tables for morphological inflections are not alone sufficient.

Japanese: verbs may productively combine, e.g., to make other passive forms, or to include modifiers to the verb (e.g., tsukusu - to do something to exhaustion), these can be combined and have normal inflections, i.e., past tense or negative form:

- 食べる (taberu - (he) eats), 食べた (tabeta - (he) ate), 食べない (tabenai - (he) did not eat), 食べなかった (tabenakatta - (he) did not eat)
- 食べられる (taberareru - (he) is eaten), 食べられた (taberareta - (he) was eaten), 食べられない (taberarenai - (he) is not eaten), 食べられなかった (taberarenakatta - (he) was not eaten)
- 食べ尽くす (tabetsukusu - (he) eats completely), 食べ尽くした (tabetsukushita - (he) ate completely), 食べ尽くさわない (tabetsukusawanai - (he) did not eat completely), 食べ尽くさわなかった (tabetsukusawanakatta - (he) did not eat completely)
- 食べ尽くされる (tabetsukusareru - (he) is eaten completely), 食べ尽くされた (tabetsukushita - (he) was eaten completely), 食べ尽くされない (tabetsukusarenai - (he) is not eaten completely), 食べ尽くされなかった (tabetsukusarenakatta - (he) was not eaten completely)

3.4 Modelling complex wordforms

My segmentation guess:

食べる	*tabe-ru*	‘(he) eats’	eat-3P.SG
食べられる	*tabe-rare-ru*	‘(he) is eaten’	eat-PASS-3P.SG
食べ尽くす	*tabe-tsuku-su*	‘(he) eats completely’	eat-completely-3P.SG
食べ尽くされる	*tabe-tsuku-sare-ru*	‘(he) is eaten completely’	eat-completely-PASS-3P.SG

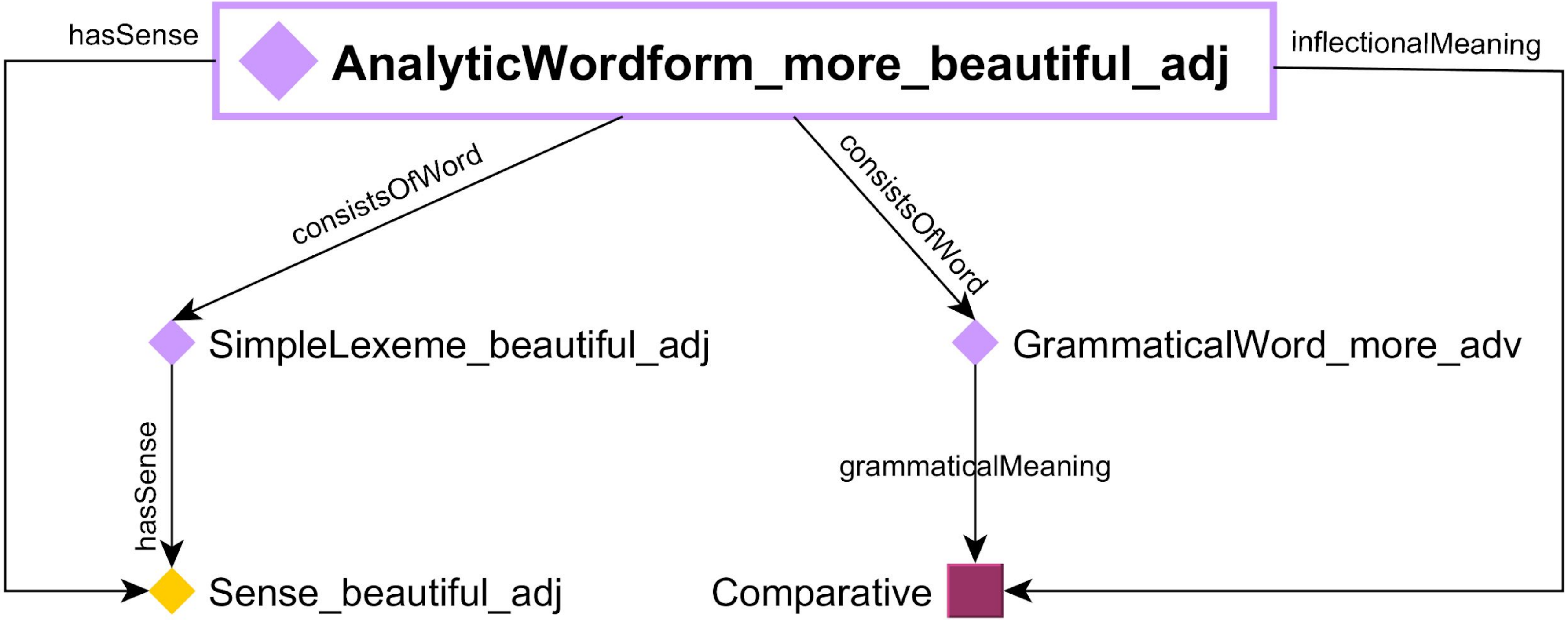
→ 3.4 Modelling complex wordforms

Treated as instances of the class `AnalyticWordform`

“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 (some of them possibly in the lexical word form).” (Lido, Christian Lehmann)

English:

- comparative adjective forms: *more beautiful*
- perfect tense verb forms: *has played*



3.4 Modelling complex wordforms

Example:
Irish (téigh - to go)
The Irish verb to go is a suppleted verb, consisting of three verbs that are used in different forms, with two of these forms having no lemma in the modern languages. This is similar to the suppletion of 'to go' with the verb 'to wend' in English, e.g., 'he goes', 'he went' (form of 'wended').

- Present: "téigh" => "Téann sé" (he goes), "Téimid" (we go)
- Past: †"cuaigh" => "Chuaigh sé" (he went), "Chuamar" (we went)
- Future: †"rachaigh" => "Rachaidh sé" (he will go), "Rachaimid" (we will go)

Compare regular verb (cuardaigh - to help)

- Present: "cuardaigh" => "Cuardaíonn sé" (he helps), "Cuardaímid" (we help)
- Past: "cuardaigh" => "Chuardaigh sé" (he helped), "Chuardamar" (we helped)
- Future: "cuardaigh" => "Cuardóidh sé" (he will help), "Cuardóimid" (we will help)

† Non-extant form

I4: Suppletive forms replace a stem+affix combination with a wholly different word

Words frequently use multiple stems and inflections in different tenses can be based on distinct stems.

4. Challenges for an OntoLex Morphology Module

Content:

What? morphs and morphemes or only morphs

Where? position/order of morphs in word form

How? addition, internal modification, infixation

→ What is the main purpose of the OntoLex morphology module (morph. data in dictionaries)?

→ What kind of morphological data exists and in what format?

Modelling:

Domain delimitation: how to avoid overlap to decomp and ontolex module?

How should paradigms be represented?

To what extent will MMoOn be reusable for the OntoLex morphology module?

→ MMoOn vocabulary aligned with MMoOn (only different namespace)
→ new vocabulary aligned with MMoOn
→ new vocabulary different from MMoOn (no reuse)

Further Reading

MMoOn website: http://mmoon.org/

MMoOn Core ontology: http://mmoon.org/core/

MMoOn projects, data and more: https://github.com/MMoOn-Project

MMoOn publications:

- Bosch, S.; Eckart, T.; Klimek, B.; Goldhahn, D. & Quasthoff, U. (2018) Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment. In: *The 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Japan, Miyazaki*.
- Eckart, T., Klimek, B., Goldhahn, D., & Bosch, S. (2018, October). Using Linked Data Techniques for Creating an IsiXhosa Lexical Resource-a Collaborative Approach. In *CLARIN Annual Conference 2018*.
- Klimek, Bettina (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*. 2017.
- Klimek, B.; Arndt, N.; Krause, S. & Arndt, T. (2016) Creating Linked Data Morphological Language Resources with MMoOn.The Hebrew Morpheme Inventory. In: *The 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Slovenia, Portorož*.

OntoLex Wiki: : https://www.w3.org/community/ontolex/wiki/Morphology