Linguistic Linked Open Data for Humanists

Anas Fahad Khan, Giulia Pedonese, Michele Mallia

Lisbon Summer School in Linguistics, July 1-5, 2024

Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca

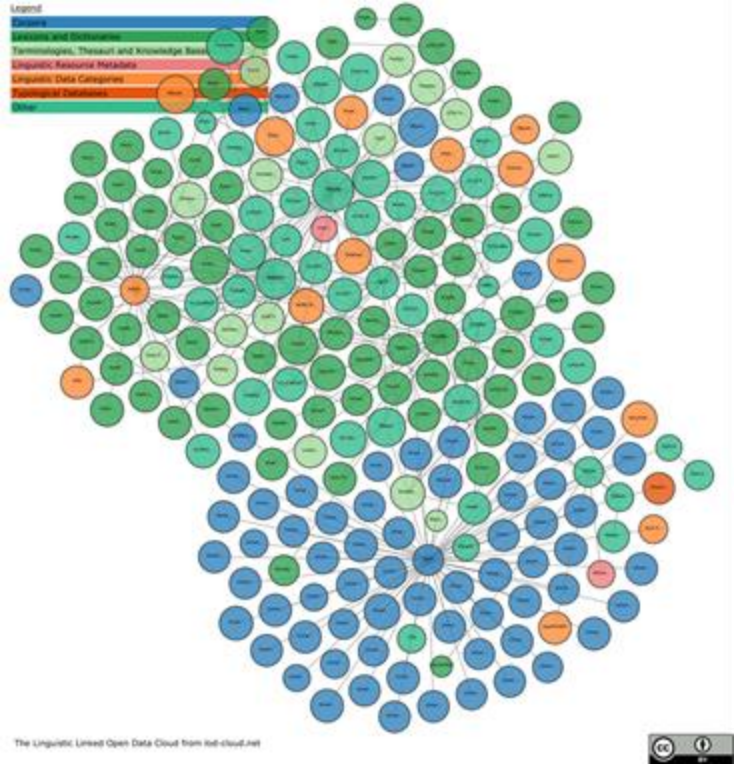
Italiadomani

Consiglio Nazionale delle Ricerche

Exploring the Linguistic Linked Open Data Cloud

Tour du Linguistic Linked Open Data Cloud

- What are some of the **vocabularies** (semantic artifacts) we could use for our building **own resources**? What are some **important resources** on the cloud of each type?
- We will do a brief **tour** of the LLOD cloud in the next few slides
- We will follow this with a deep dive into the category of **Lexicons and Dictionaries**

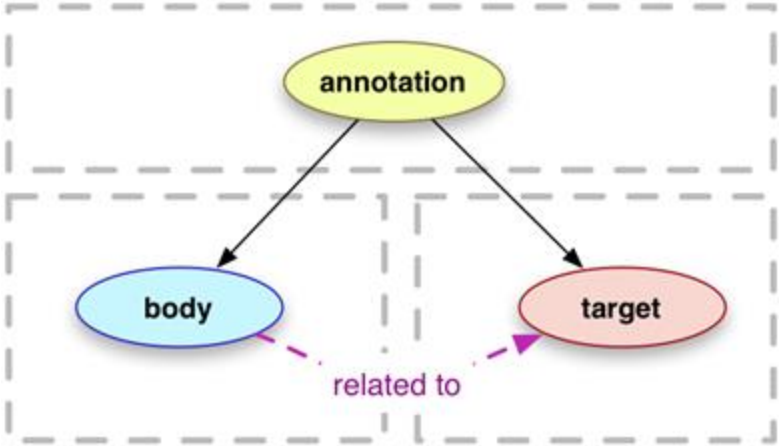


Annotations and Corpora

- **NLP Interchange Format** (NIF)
  - RDF vocabulary for **strings and their annotation**. Designed specifically for **NLP pipelines and web services**. Used in academic and industrial applications, esp. the **DBpedia** community
  - Namespace: http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
  - Prefix: nif:
  - Key Terms:
    - `nif:Context`: Represents the context of a text fragment.
    - `nif:String`: Represents a string within the text.
    - `nif:beginIndex` and `nif:endIndex`: Indicate the start and end positions of a text span.
    - `nif:annotation`: Links to annotations of the text span.

Annotations and Corpora

- **Web Annotation Format:** Offers a framework for **creating and sharing annotations** of web resources.
  - Namespace: http://www.w3.org/ns/oa#
  - Prefix: oa:
  - Key Terms:
    - `oa:Annotation`: Represents an annotation.
    - `oa:hasBody`: Links to the body of the annotation.
    - `oa:hasTarget`: Links to the target resource being annotated.
    - `oa:Motivation`: Represents the purpose or intention of the annotation.



Attrib: Robert Sanderson, Paolo Ciccarese, Benjamin Young (eds.),

Annotations and Corpora

- **POWLA**
  - An **ontology for representing linguistic annotations**, particularly those derived from previous XML-based formats like PAULA.
  - Namespace: http://purl.org/powla/powla.owl#
  - Prefix: powla:
  - Key Terms:
    - `powla:Node`: Represents a node in a linguistic annotation graph.
    - `powla:Edge`: Represents an edge in a linguistic annotation graph.
    - `powla:hasParent`: Links a node to its parent node.
    - `powla:hasChild`: Links a node to its child node.
    - `powla:hasLayer` assigns a Relation or a Node a an annotation layer.
    - `powla:previous` a relation for connecting two powla:Nodes in a sequence.
- **CoNLL-RDF**
  - NIF subset, designed for simplicity and interoperability with (pre-LOD) standards in NLP, especially **CoNLL**-based ones

Annotations and Corpora

- Problems:
  - Although representing corpora in LOD makes them much more **interoperable**, the results can be very **verbose**, this can lead to a **large storage overhead and query complexity**
- Current challenges
  - Neccesity to **support more complex annotations**
  - **Harmonize LOD standards** for annotations (no one vocabulary used by everyone)
  - **Harmonize with pre-LOD standards**, e.g., ISO 24612:2012
- Current subject of discussions in several W3C Community Groups
  - **W3C CG Best Practices for Multilingual Linked Open Data**
  - **W3C CG Linked Data for Language Technology**
  - Open to everyone!

Annotation and Corpora

Corpora currently listed as being on the LLOD Cloud include:

- **Brown Corpus in RDF/NIF**
- **News-100 NIF NER Corpus**
  - 100 German news articles from the online news platform news.de
- **DBPedia abstract corpus NIF**
  - Contains manually disambiguated

Lexicons and Dictionaries

- Here we can cite one of the big success stories of linguistic linked data, **OntoLex-Lemon** and its various extensions
- Originally intended as a model for **enriching ontologies with linguistic information** but now seen as a de facto standard for creating and publishing lexicons on the Semantic Web (with or without ontologies)
- Used in many projects and in the conversion of many important datasets, including **Wordnets**, **Wiktionary**, the **Apertium** series of dictionaries

Lexicons and Dictionaries

- Popularity of OntoLex-Lemon may (partly) be due to how well **the content of lexicons (usually) lends itself to representation** in a graph based model
- The results are (usually) **not very verbose and the model is very intuitive**
- The **W3C Ontolex group** is a very active one and it is currently working on **two extensions**. It is open to everyone and we have regular meetings on a weekly basis. Demonstrates the importance of community groups
- In the next part of the presentation we will look **in depth** at OntoLex Lemon

Terminologies, Thesauri and Knowledge Bases

- We have already looked at SKOS which is intended for taxonomies and thesauri
- For **terminologies** there have been several proposals and there is some provsion available for encoding **a terminology as an RDF resource** but there is no general agreed upon approach to e.g., converting TBX into RDF
- …However There are plans to develop **a extension of OntoLex-lemon** for terminology bases too
- Anyone interested is advised to join the Ontology Lexicon group:
    - https://www.w3.org/community/ontolex/

Linguistic Resource Metadata/Linguistic Data Categories

- Models for creating linguistic metadata include **lime** (the metadata module of OntoLex-lemon) and the **wider coverage Metashare ontology**
- **Data Category Registries (DCRs)** are vital for ensuring **interoperability** across linguistic datasets
- Most well known Linguistic Linked Data DCR is **lexinfo** which is based on the now defunct **ISOCat registry**

Typological Databases

- Currently **not very much to report**…however see the following paper for a discussion of previous efforts in putting together such resources (and an overview of vocabularies and models for the LLOD cloud in general):
  - **Khan, Anas Fahad et al. ‘When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data’. 1 Jan. 2022 : 987 – 1050. 10.3233/SW-222859**

The State of the Cloud

- Last updated in **September 2023**.
- Missing quite a **few LLOD datasets, e.g., LiLa**
- Includes some broken links, and some **strange categorisations and colour scheme choices** (3 shades of green for 3 different categories)
- The LLOD cloud categories are arguably in need of revision to reflect e.g., **the CLARIN resource families**
- Still **a very useful index of LLOD resources**

Focus on Lexicons and Dictionaries in RDF

Introduction

- Before introducing OntoLex-Lemon model and its extensions and showing how it can be used in the **creation/conversion and publication of dictionaries and lexicons** (lexical resource) we will give some background on digital lexical resources
- We will look at related standards **TEI** and **LMF** studying the comparative advantages and disadvantages of each.
    - *There is a new format about to be published by OASIS which we won't look at here because it's still being drafted*
- Note that this is **a very popular topic** at the moment.An increasing number of lexicons are being published in as digital resources. Old legacy print dictionaries are also being converted into formats like TEI and are also being published as linked data (**retrodigitisation**).

Computational Lexicons

- Digital lexicons can be formatted in many different ways. One of the simplest ways is as **a raw text file** consisting of a list of lexical entries, each separated by **a space or a new line**.
- This can be hard to read and to navigate for human beings
- Searching/querying of my data will also be limited (e.g., **How do I find a list of all the lexical entries? All the verbs ending in '-er'?**).
- Requires the use of **special tools** in order to automatically extract the different kinds of information (**morphological, syntactic, semantic, pragmatic, encyclopedic?**) contained in the text file.

Computational Lexicons

Fè, _as_ Féde. _Also as_ Féce, _he did or made._

Feáce, _voluptuous, giuen to all sensuality._

Fébbre, _a feauer, an ague._

Fébbre c[o]ntínua, _a continuall ague._

Fébbre c[o]tidiána, _a quotidian ague._

Fébbre quartána, _a quartan ague._

Fébbre térzána, _a tertian ague._

Fébbreggiánte, _troubled with an ague._

Fébbreggiáre, _to haue an ague._

Fébbricélla, _a gentle or easie ague._

Fébbricitánte, _troubled with an ague._

Fébbricitáre, _to haue or be sicke of an ague._

Fébbricci[ó]s[o], _aguish, troubled with an ague._

Fébbric[ó]s[o], _sicke of, or full or a feauer._

Fébbr[ó]ne, _a violent ague, a burning feauer._

Fébbr[ó]s[o], _troubled with or hauing an ague._

Febéa, _vsed for Phebe or the Moone._

Febrái[o], Febrár[o], _the moneth Februarie._

Fecatélle, _as_ Fegatélle.

Fécat[o], _as_ Fégat[o].

Computational Lexicons

- Lexicons can also be **formatted with word processing packages** like Word or a markup language like HTML to make them more **humanly readable** and more like the print dictionaries we are used to browsing.
- But this seems like a missed opportunity: working with digital resources makes it much easier to enhance the information contained in lexicons/dictionaries, as well as **making the lexical knowledge contained in it more accessible both to humans and machines**.
- One of the best ways of doing the latter is by using a **specialised markup** that annotates the text for relevant morpho-syntactical/lexicographic information.

Computational Lexicons

- Rather than defining a new kind of markup per document/project/ institution it’s usually better to use a standard.
- That is, **the use of standards helps to promote interoperability and reusability of datasets**. Standards also (usually) represent the consensus of a community of experts.
- Standards for digital lexicons can be **viewed as an extended/modified version of the standards adopted by lexicographers when producing print dictionaries**, e.g., alphabetic ordering of words, the use of bold font for headwords.

Computational Lexicons - Two Typologies

- First dichotomy is between **native-born digital resources** and **legacy, retrodigitised, print resources**. The first type was created as a digital resource while the second is either a conversion of a previous print resource or is based on one.
- The second is the dichotomy between **NLP-dictionaries (NLP)** and **Machine Readable Dictionaries (MRD)**. The first category is specifically designed with Natural Language Processing applications in mind; the other is both for human and for machine consumption.
    - This distinction was important in the past but is perhaps less so now!

Lexical Markup Framework (LMF)

- The original LMF was a framework for producing computational lexicons and was an **ISO standard published in 2008**, developed under the aegis of the the ISO technical committee ISO-TC37/SC4.
- It was the result of five years of work and the input of around 60 different experts. Special care was taken to ensure that it could be used for non European languages.
- Was very influential on subsequent work (of great historical interest). But it is a closed standard which has effectively **limited its use**
- LMF was not **XML-native** but had a serialisation XML.
- Currently being revised and republished by ISO-TC37/SC4 (the new version has a TEI-XML serialisation).

LMF - an example

```
<LexicalEntry>
   <feat att="partOfSpeech" val="commonNoun"/>
   <Lemma>
      <feat att="writtenForm" val="clergyman"/>
   </Lemma>
   <WordForm>
      <feat att="writtenForm" val="clergyman"/>
      <feat att="grammaticalNumber" val="singular"/>
   </WordForm>
   <WordForm>
      <feat att="writtenForm" val="clergymen"/>
      <feat att="grammaticalNumber" val="plural"/>
   </WordForm>
</LexicalEntry>
```

LMF - an example

```
<LexicalEntry>
   <feat att="partOfSpeech" val="commonNoun"/>
   <Lemma>
      <feat att="writtenForm" val="clergyman"/>
   </Lemma>
   <WordForm>
      <feat att="writtenForm" val="clergyman"/>
      <feat att="grammaticalNumber" val="singular"/>
   </WordForm>
   <WordForm>
      <feat att="writtenForm" val="clergymen"/>
      <feat att="grammaticalNumber" val="plural"/>
   </WordForm>
</LexicalEntry>
```

The **Lexical Entry** element groups together form and sense information

LMF - an example

```
<LexicalEntry>
   <feat att="partOfSpeech" val="commonNoun"/>
   <Lemma>
      <feat att="writtenForm" val="clergyman"/>
   </Lemma>
   <WordForm>
      <feat att="writtenForm" val="clergyman"/>
      <feat att="grammaticalNumber" val="singular"/>
   </WordForm>
   <WordForm>
      <feat att="writtenForm" val="clergymen"/>
      <feat att="grammaticalNumber" val="plural"/>
   </WordForm>
</LexicalEntry>
```

With the **Lemma** element we can mark out the headword of the entry

LMF - an example

```
<LexicalEntry>
   <feat att="partOfSpeech" val="commonNoun"/>
   <Lemma>
      <feat att="writtenForm" val="clergyman"/>
   </Lemma>
   <WordForm>
      <feat att="writtenForm" val="clergyman"/>
      <feat att="grammaticalNumber" val="singular"/>
   </WordForm>
   <WordForm>
      <feat att="writtenForm" val="clergymen"/>
      <feat att="grammaticalNumber" val="plural"/>
   </WordForm>
</LexicalEntry>
```

We can also specify different variants of the lexical entry using the element **WordForm**

LMF - an example

```
<LexicalResource dtdVersion="16">
 <GlobalInformation>
  <feat att="label" val="Simple English LMF test suites"/>
  <feat att="languageCoding" val="ISO 639-3"/>
 </GlobalInformation>
 <Lexicon>
  <feat att="language" val="eng"/>
  <LexicalEntry>
    <!--Inflected forms of clergyman are clergyman and clergymen-->
    <feat att="partOfSpeech" val="commonNoun"/>
    ....
  </LexicalEntry>
 </Lexicon>
</LexicalResource>
```

The **Lexical Resource** element can group together more than one lexicon

LMF - an example

```
<LexicalResource dtdVersion="16">
 <GlobalInformation>
  <feat att="label" val="Simple English LMF test suites"/>
  <feat att="languageCoding" val="ISO 639-3"/>
 </GlobalInformation>
 <Lexicon>
  <feat att="language" val="eng"/>
  <LexicalEntry>
    <!--Inflected forms of clergyman are clergyman and clergymen-->
    <feat att="partOfSpeech" val="commonNoun"/>
    ....
  </LexicalEntry>
 </Lexicon>
</LexicalResource>
```

The **Global Information** element includes information that holds throughout all the lexicons in the resource, e.g., language or script coding

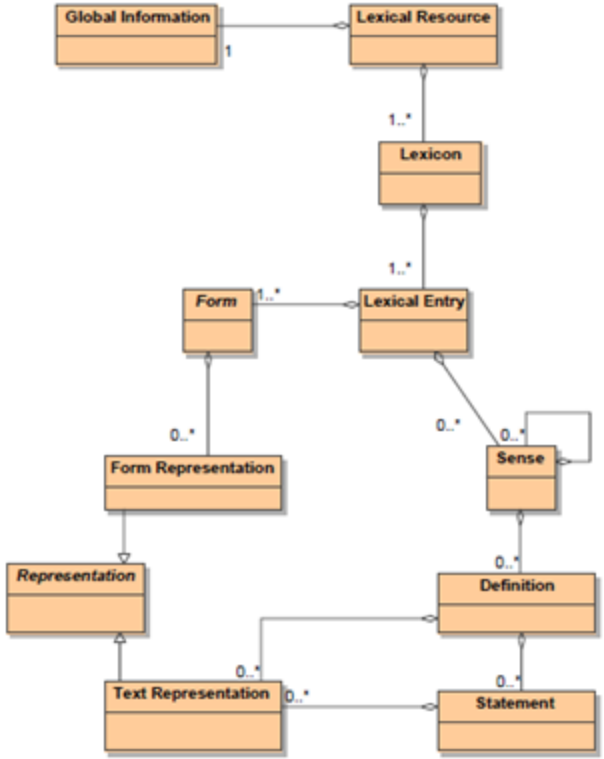
LMF - an example

```
<LexicalResource dtdVersion="16">
 <GlobalInformation>
  <feat att="label" val="Simple English LMF test suites"/>
  <feat att="languageCoding" val="ISO 639-3"/>
 </GlobalInformation>
 <Lexicon>
  <feat att="language" val="eng"/>
  <LexicalEntry>
    <!--Inflected forms of clergyman are clergyman and clergymen-->
    <feat att="partOfSpeech" val="commonNoun"/>
    ....
  </LexicalEntry>
 </Lexicon>
</LexicalResource>
```

The **Lexicon** element groups together all the lexical entries of a lexicon. It also allows us to specify information pertaining to the whole lexicon such as e.g., language

LMF - the Core Model

- The core model of LMF contained the classes we saw in the previous example. They include a **Sense** class too, representing each of the different **meanings** that a lexical entry can have.
- These classes were considered the most ‘essential’ for building lexicons.
- The additional modules contained classes that are specific to NLP dictionaries and MRDs, or for expressing morpho-syntactic/semantic properties



Text Encoding Initiative (TEI)

- The Text Encoding Initiative (TEI) refers both to a widely used standard for encoding digital texts as XML as well as to the consortium that maintains and develops them.
- The TEI standard is set down as a series of guidelines which taken together define an XML schema. The TEI guidelines are divided up into several parts and include a number of specialist modules each dealing with a different kind of text, including a module for dictionaries.
- The TEI guidelines also include a **special module for encoding dictionaries (TEI-DICT)**. Developed for MRDs rather than NLP dictionaries.
- TEI-DICT is a very popular standard for publishing lexical resources.

TEI-DICT - an Example

pŏdĭum , ii, n., = πόδιον,

I.an elevated place, a height.

I. In gen. (post-class.): “podia ternis alta pedibus fabricantur,” Pall. 1, 38.—

II. In partic.

A. A projection in a building, a jutty, balcony, podium (postAug.), Plin. Ep. 5, 6, 22; Vitr. 3, 3; 5, 7; 7, 4, 4; Dig. 33, 7, 12, § 22.—

B. A projecting part in the circus or amphitheatre, a parapet or balcony next to the arena, where the emperor and other distinguished persons sat, Suet. Ner. 12; cf. Plin. 37, 3, 11, § 45: “omnes ad podium spectantes,” Juv. 2, 147.

(Source: A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary. revised, enlarged, and in great part rewritten by. Charlton T. Lewis, Ph.D. and. Charles Short, LL.D. Oxford. Clarendon Press. 1879.)

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> , =
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> , =
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

The **entryFree** element contains a single unstructured entry

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> , =
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

The **entryFree** element contains a single unstructured entry

In TEI there is more than one way of creating or encoding a single lexical entry:

- The element **entry** provides a way of creating structured lexical entries
- The element **entryFree** allows for the encoding of lexical entries without the constraints imposed on entry
- The element **superEntry** allows for the grouping together of entries that function as one (such as e.g., homographs)

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> , =
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

The **orth** element allows for the encoding of the orthographical form of the headword

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> ,=
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

The **itype** element encodes the inflectional type of an entry*

The **gen** element encodes morphological gender

*Represented in this case by the ending of the genitive form

TEI-DICT - an Example

```
<entryFree id="n36781" type="greek" key="podium" opt="n"><orth extent="full" lang="la" opt="n">pŏdĭum</orth>,
    <itype opt="n">ii</itype>,
    <gen opt="n">n.</gen> , =
    <foreign lang="greek">πόδιον</foreign>,
    <sense id="n36781.0" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.1" n="I" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.2" n="II" level="1" opt="n">
        …
    </sense>
    <sense id="n36781.3" n="A" level="2" opt="n">
        …
    </sense>
    <sense id="n36781.4" n="B" level="2" opt="n">
        …
    </sense>
</entryFree>
```

The **sense** element groups together information on a single sense of a lexical entry

TEI-DICT - an Example

Let's take a closer look at one of the individual senses...

pŏdĭum , ii, n., = πόδιον,
I.an elevated place, a height.
I. In gen. (post-class.): "podia ternis alta pedibus fabricantur," Pall. 1, 38.—
II. In partic.

A. A projection in a building, a jutty, balcony, podium (postAug.), Plin. Ep. 5, 6, 22; Vitr. 3, 3; 5, 7; 7, 4, 4; Dig. 33, 7, 12, § 22.—

B. A projecting part in the circus or amphitheatre, a parapet or balcony next to the arena, where the emperor and other distinguished persons sat, Suet. Ner. 12; cf. Plin. 37, 3, 11, § 45: "omnes ad podium spectantes," Juv. 2, 147.

TEI-DICT - an Example

Let’s take a closer look at one of the individual senses…

```
<sense id="n36781.4" n="B" level="2" opt="n">
        A projecting part in the circus or amphitheatre,
        <hi rend="ital">a parapet</hi> or <hi rend="ital">balcony next to the arena</hi>
        emperor and other distinguished persons sat,
        <bibl n="Suet. Nero 12" default="NO" valid="yes">
                <author>Suet.</author> Ner. 12
        </bibl>; cf.
        <bibl n="Plin. Nat. 37.45" default="NO" valid="yes"><author>Plin.</author> 37, 3, 11, §
45
        </bibl>:
        <cit><quote lang="la">omnes ad podium spectantes,</quote>
                <bibl n="Juv. 2.147" default="NO" valid="yes">
                        <author>Juv.</author> 2, 147
                </bibl>
        </cit>.
</sense>
```

We can specify the ‘co-ordinate’ for a sense (in a hierarchy) using **@level** and **@n**

TEI-DICT - an Example

Let’s take a closer look at one of the individual senses…

```
<sense id="n36781.4" n="B" level="2" opt="n">
        A projecting part in the circus or amphitheatre,
        <hi rend="ital">a parapet</hi> or <hi rend="ital">balcony next to the arena</hi>, where the
        emperor and other distinguished persons sat,
        <bibl n="Suet. Nero 12" default="NO" valid="yes">
                <author>Suet.</author> Ner. 12
        </bibl>; cf.
        <bibl n="Plin. Nat. 37.45" default="NO" valid="yes"><author>Plin.</author> 37, 3, 11, §
45
        </bibl>:
        <cit><quote lang="la">omnes ad podium spectantes,</quote>
                <bibl n="Juv. 2.147" default="NO" valid="yes">
                        <author>Juv.</author> 2, 147
                </bibl>
        </cit>.
</sense>
```

The element **cit** contains citational information

quote contains associates a quotation with the citation

The element **bibl** groups together bibliographic data

TEI-DICT

TEI allows us to take three different views of dictionary data:

- *(a) the **typographic** view—the two-dimensional printed page, including information about line and page breaks and other features of layout*
- *(b) the **editorial** view—the one-dimensional sequence of tokens which can be seen as the input to the typesetting process; the wording and punctuation of the text and the sequencing of items are visible in this view, but specifics of the typographic realization are not*
- *(c) the **lexical** view—this view includes the underlying information represented in a dictionary, without concern for its exact textual form*

(Taken from the TEI guidelines http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html)

TEI-DICT

- In other words TEI allows us to model both how lexical information is represented, the exact sequence of words/punctuation used, where the lines breaks are, etc, and the lexical information itself. **The *how* and the *what* at the same time**.
- TEI therefore has to be ***flexible enough to represent all the different ways in which dictionaries represent lexical (and other) information***. If we're only interested in the content of the information (and not the exact sequence of words and punctuation used to present it), then the flexibility of TEI DICT might be a hindrance.
- A model like ***LMF***, and as we will see also ***Ontolex-Lemon***, focuses on the lexical content itself and not how it is represented. Laurent Romary, Ana and others are currently working on a ***standardised and simplified version of TEI DICT called TEI-Lex0.***

lemon

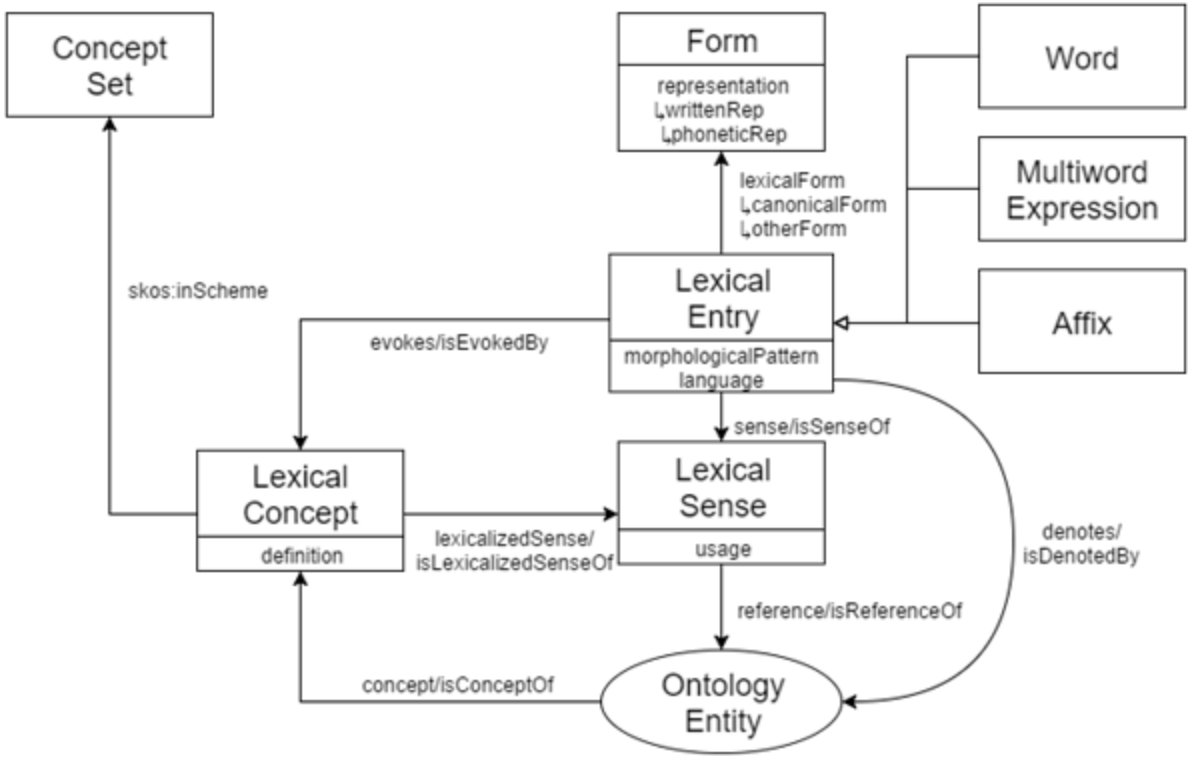
- Lemon stands for the **Lexicon Model for Ontologies**.
- This was an ontology developed as part of the **Multilingual Ontologies for Networked Knowledge (Monnet)** (2010-13) project as a collaboration between several European universities and academic institutes. It was **closely based on previous standards/models**, and in particular on LMF.
- Unlike those previous models **it was RDF-native**
- Lemon was originally intended as a model for enhancing knowledge bases and ontologies like DBpedia with linguistic knowledge: that is for grounding such resources with linguistic information
- You can browse the model here (for historical interest only):
  - https://lemon-model.net/

OntoLex-Lemon

Lemon soon became the most popular model for representing lexicons in RDF, taking on the status of a de facto standard. It was used to model the P**rinceton (and other) Wordnets, DBnary (the linked data version of Wiktionary), FrameNet and VerbNet**.

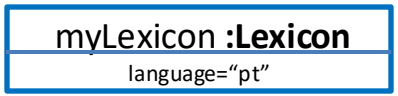
This success led to the development of a new version, **OntoLex-Lemon**, published in December 2016. It consists of a **core module** as well as a **metadata module (lime)**, a **syntax and semantics module (synsem)**, a **decomposition module (decomp)**, and a **variation and translation module (vartrans)**.

Ontolex-Lemon Core



LLD – Ontolex-Lemon: an example

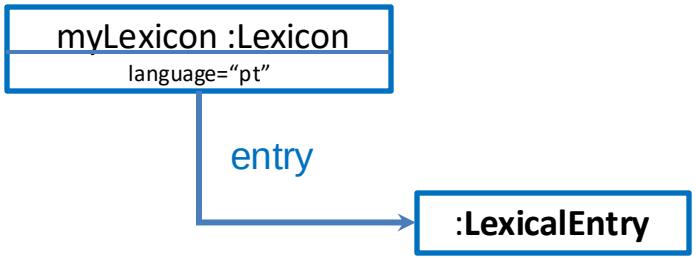
Lexicon: The object representing the lexicon as a whole.



[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: an example

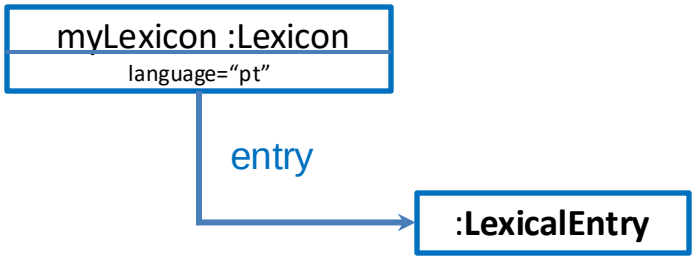
Lexical Entry: An entry in a lexicon is a container for one or several forms and
one or several meanings of a lexeme.



[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: an example

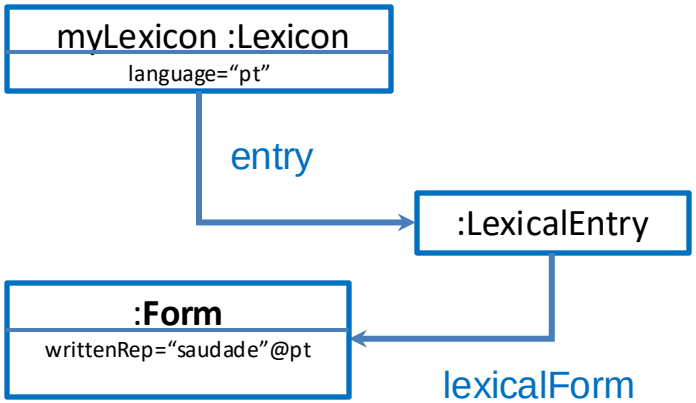
Lexical Entry: An entry in a lexicon is a container for one or several forms and
one or several meanings of a lexeme.



[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: an example

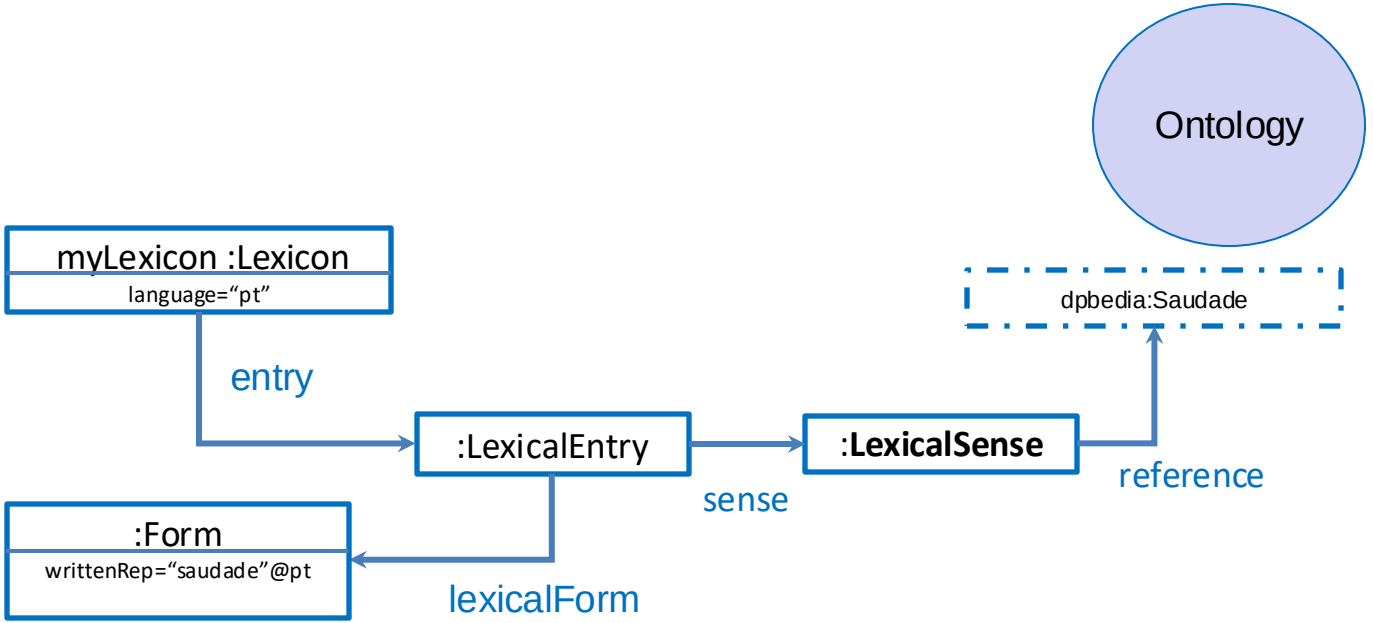
Lexical Form: An inflectional form of an entry. A given lexical form may have several representations in different orthographies.



[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: an example

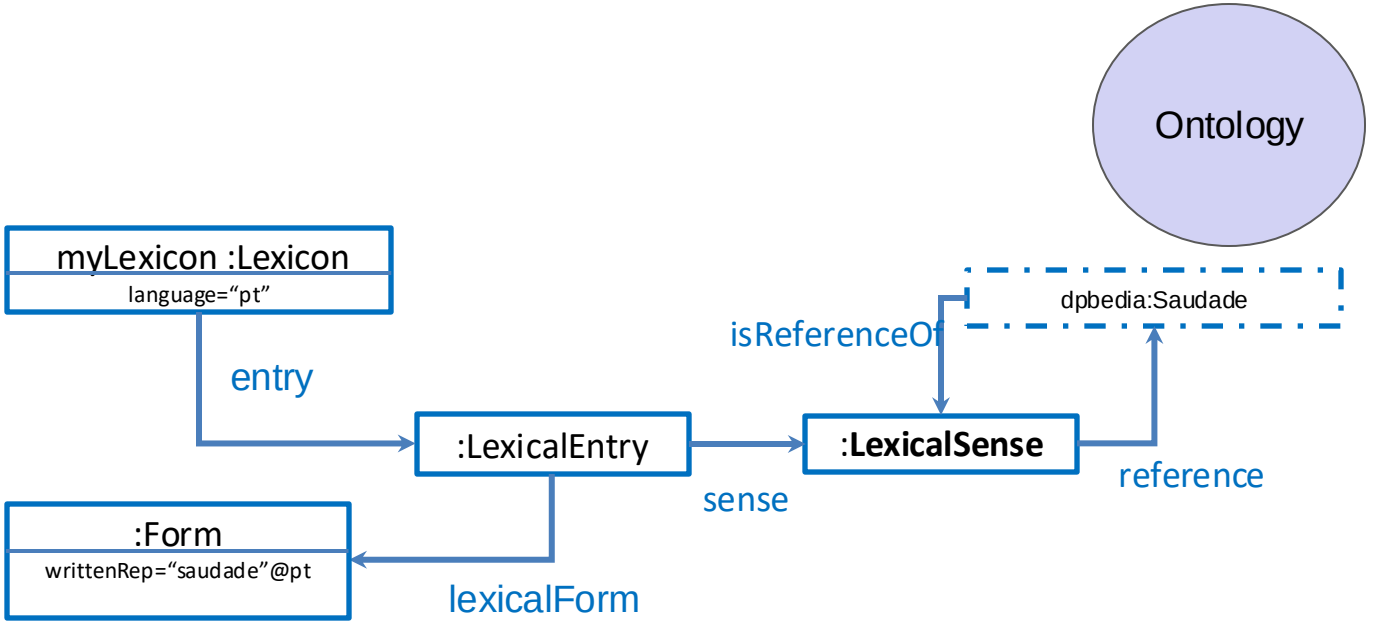
Lexical Sense: A sense links the lexical entry to the reference (ontology term) used to describe its meaning.



[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: an example

Lexical Sense: A sense links the lexical entry to the reference (ontology term) used to describe its meaning.



[Slide based on Gracia and Khan @ SDLLOD-22]

Entry in Turtle (using blank nodes)

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix dbpedia: <http://dbpedia.org/resource/>.


:myLexicon a lime:Lexicon ;
       lime:language “pt”;
       lime:entry :saudade_entry .

:saudade_entry a ontolex:LexicalEntry ;
  ontolex:canonicalForm [
     ontolex:writtenRep “saudade”@pt ] ;

  ontolex:sense [
    ontolex:reference dbpedia:Saudade] .
```

Namespace

Lemma

Sense

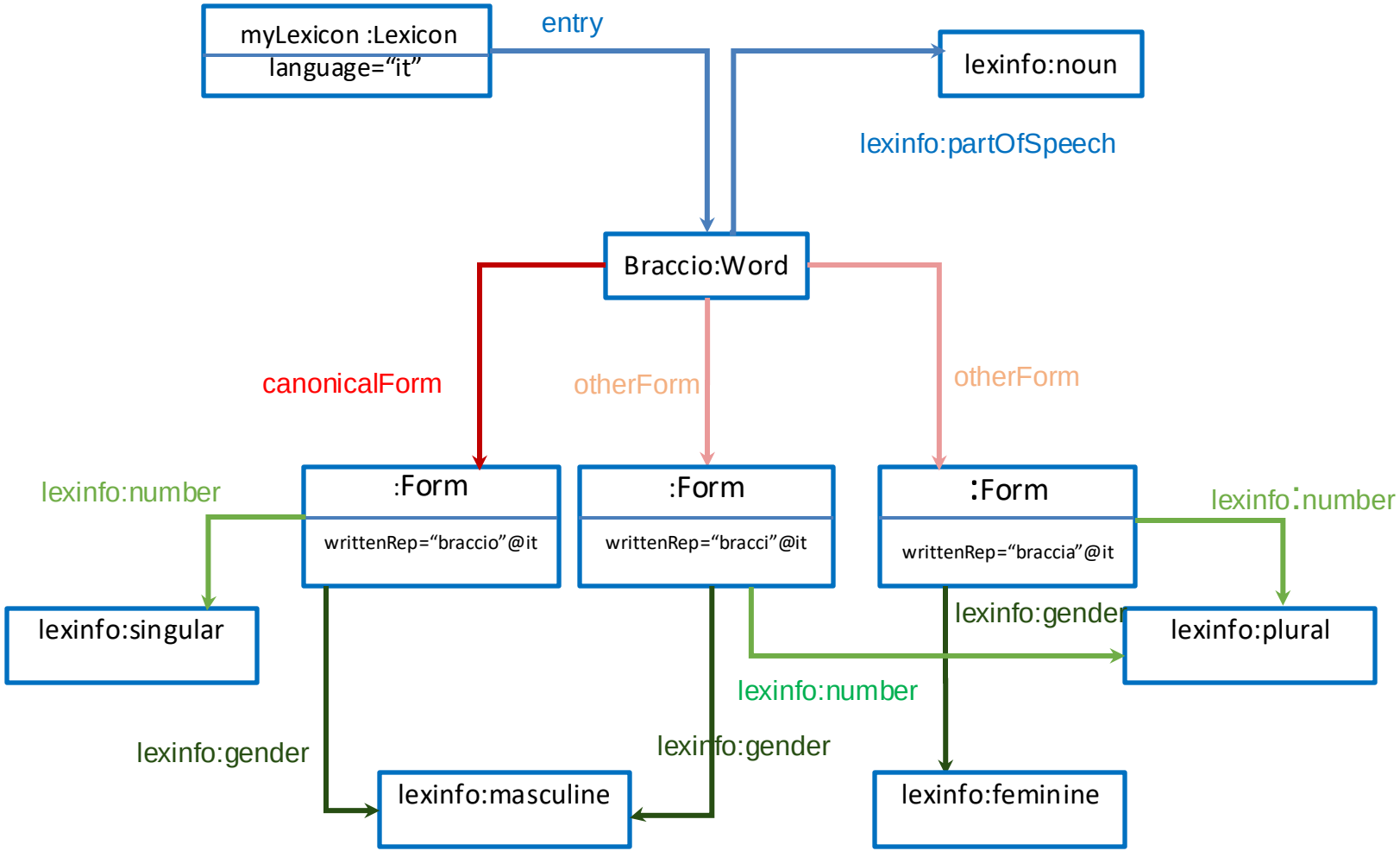
[Slide based on Gracia and Khan @ SDLLOD-22]

The Clergyman Example

We can encode the clergyman example in RDF with Ontolex using the Turtle format as follows:

```
:clergymanLE rdf:type ontolex:LexicalEntry ;
        lexinfo:partOfSpeech lexinfo:noun ;
        ontolex:canonicalForm :clergySing ;
        ontolex:otherForm :clergyPlural .
:clergyPlural rdf:type ontolex:Form;
        ontolex:writtenRep "clergymen"@en .
:clergySing rdf:type ontolex:Form ;
        ontolex:writtenRep "clergyman"@en .
```

[Slide based on Gracia and Khan @ SDLLOD-22]



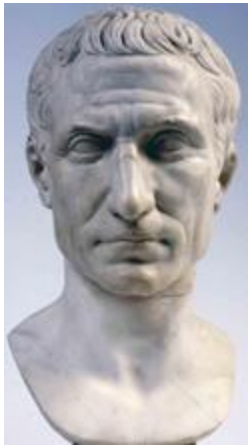
[Slide based on Gracia and Khan @ SDLLOD-22]

Adding Phonetic Information

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .

saudade_entry a ontolex:LexicalEntry;
     ontolex:lexicalForm :saudade_tomato.

:saudade_form a ontolex:Form;
     ontolex:writtenRep "saudade"@pt;
     ontolex:phoneticRep "sɐwˈða.ðɨ"@pt-PT-fonipa;
     ontolex:phoneticRep "saʊˈ̯da.d͡ʒi"@pt-BR-fonipa.
```

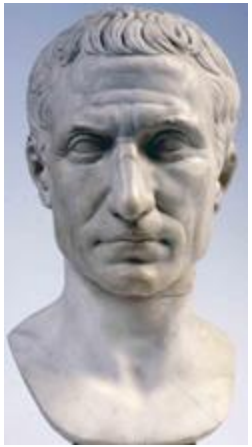
Basic Morphological Information

```
:venio ontolex:morphologicalPattern :latin_fourth_conjugation ;
  ontolex:canonicalForm :venio_form ;
  ontolex:otherForm :veni_form .
:venio_form ontolex:writtenRep "veniō"@la;
:veni_form ontolex:writtenRep "vēnī"@la .
```



[Slide based on Gracia and Khan @ SDLLOD-22]

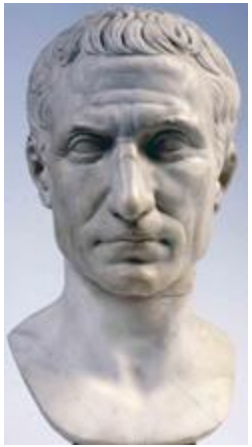
Basic Morphological Information



```
:venio ontolex:morphologicalPattern :latin_fourth_conjugation ;
  ontolex:canonicalForm :venio_form ;
  ontolex:otherForm :veni_form .
:venio_form ontolex:writtenRep "veniō"@la;
:veni_form ontolex:writtenRep "vēnī"@la .

:video ontolex:morphologicalPattern :latin_second_conjugation ;
  ontolex:canonicalForm :video_form ;
  ontolex:otherForm :vidi_form .
:video_form ontolex:writtenRep "videō"@la ;
:vidi_form ontolex:writtenRep  "vīdī"@la.
```

[Slide based on Gracia and Khan @ SDLLOD-22]

Basic Morphological Information



```
:venio ontolex:morphologicalPattern :latin_fourth_conjugation ;
  ontolex:canonicalForm :venio_form ;
  ontolex:otherForm :veni_form .
:venio_form ontolex:writtenRep "veniō"@la;
:veni_form ontolex:writtenRep "vēnī"@la .

:video ontolex:morphologicalPattern :latin_second_conjugation ;
  ontolex:canonicalForm :video_form ;
  ontolex:otherForm :vidi_form .
:video_form ontolex:writtenRep "videō"@la ;
:vidi_form ontolex:writtenRep  "vīdī"@la.

:vinco ontolex:morphologicalPattern :latin_third_conjugation ;
  ontolex:canonicalForm :vinco_form ;
  ontolex:otherForm :vinco_form.
:vinco_form ontolex:writtenRep "vincō"@la
:vici_form ontolex:writtenRep "vīcī"@la  .
```

[Slide based on Gracia and Khan @ SDLLOD-22]

Basic Semantic Information

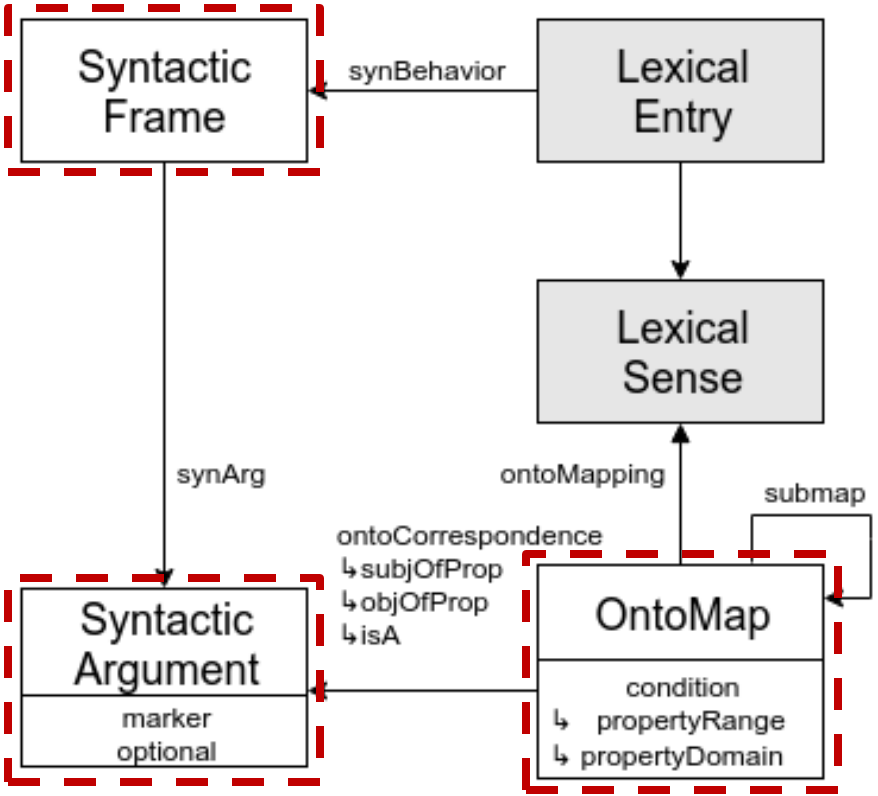
```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbo: <http://dbpedia.org/ontology/> .

:trem a ontolex:Word ;
  ontolex:sense [
    ontolex:reference dbpedia:Train ;
    ontolex:usage [ rdf:value “Brazilian Portuguese” ] ] ;
  ontolex:denotes dbpedia:Train .

:comboio a ontolex:Word ;
  ontolex:sense [
    ontolex:reference dbpedia:Train ;
    ontolex:usage [ rdf:value “European Portuguese” ] ] ;
  ontolex:denotes dbpedia:Train .
```

sense ∘ reference
=
denotes

Restriction on
Lexical Sense

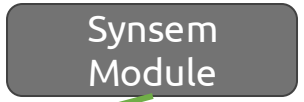
[Slide based on Gracia and Khan @ SDLLOD-22]

Syntax and Semantics



[Slide based on Gracia and Khan @ SDLLOD-22]

Syntactic Frames




```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix synsem: <http://www.w3.org/ns/lemon/synsem#> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .

:know a ontolex:Word ;
  synsem:synBehavior :know_transitive .

:know_transitive a synsem:SyntacticFrame, lexinfo:TransitiveFrame
;
  lexinfo:subject :know_subject ;
  lexinfo:directObject :know_directObject .
```

[Slide based on Gracia and Khan @ SDLLOD-22]

Syntactic and Semantic Frames

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix synsem: <http://www.w3.org/ns/lemon/synsem#> .
@prefix lexinfo: <http://www.lexinf                /2.0/lexinfo#> .
@prefix foaf: <http://xmlns.com/foa
@prefix rdfs: <http://www.w3.org/20                ema#> .

:know a ontolex:Word ;
  ontolex:sense :know_sense ;
  synsem:synBehavior :know_transitive .

:know_sense a ontolex:LexicalSense , synsem:OntoMap ;
  synsem:ontoMap :know_sense ;
  ontolex:reference foaf:knows ;
  synsem:subjOfProp :know_subject ;
  synsem:objOfProp :know_directObject .

foaf:knows a rdf:Property ;
  rdfs:domain foaf:Person ;
  rdfs:range foaf:Person .
```

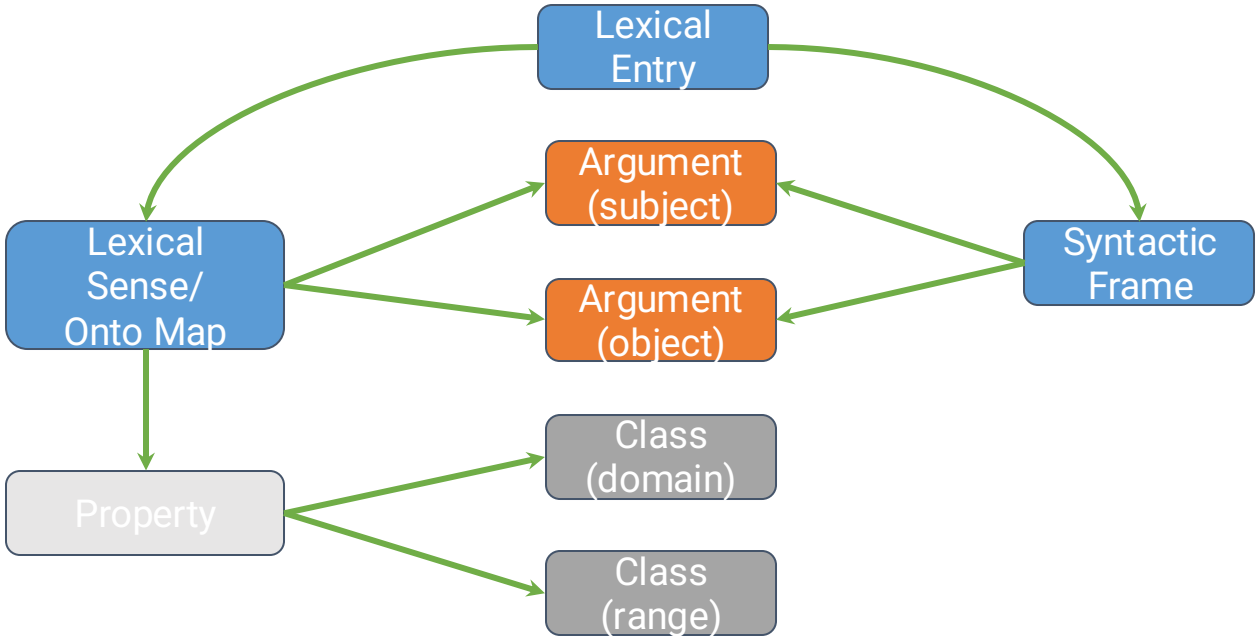
Lexical sense is an ontology mapping

Identifiers from syntactic frame

Ontological definition of semantic frame

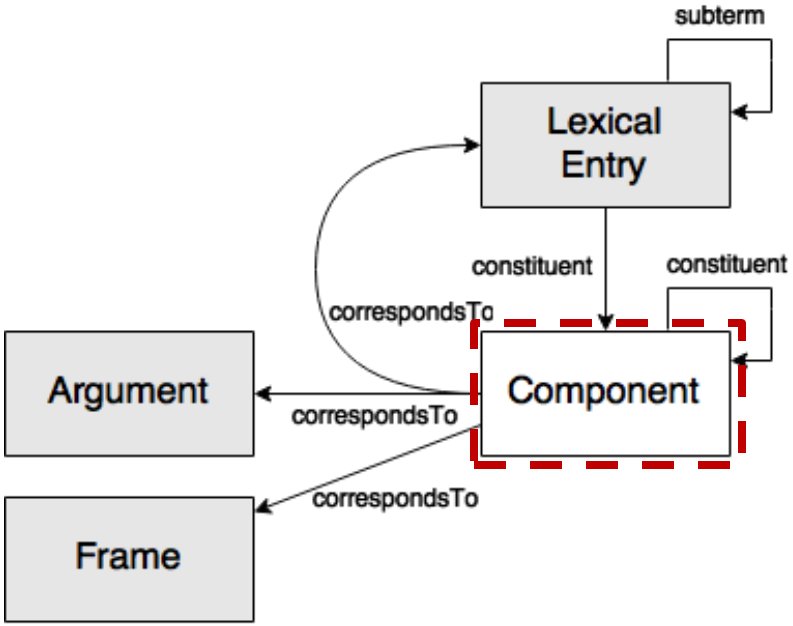
[Slide based on Gracia and Khan @ SDLLOD-22]

Syntactic-Semantic Mapping



[Slide based on Gracia and Khan @ SDLLOD-22]

Decomposition



[Slide based on Gracia and Khan @ SDLLOD-22]

Decomposition

constituent ∘ correspondsTo
=
subterm

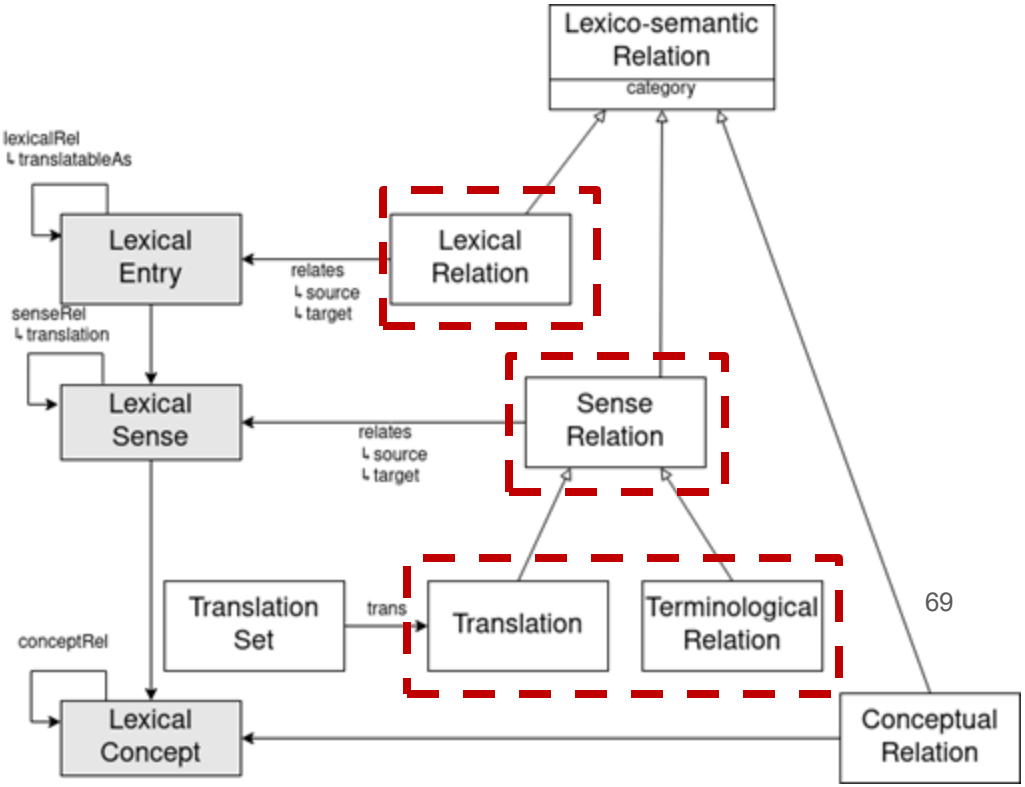
```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix decomp: <http://www.w3.org/ns/lemon/decomp#> .

:summer_school a ontolex:MultiWordExpression ;
  decomp:subterm :summer, :school .

:curso_de_verão a ontolex:MultiWordExpression ;
  decomp:constituent :curso_de_verão_curso_comp ,
                     :curso_de_verão_de_comp ,
                     :curso_de_verão_verão_comp ;
  rdf:_1 :curso_de_verão_curso_comp ;
  rdf:_2 :curso_de_verão_de_comp ;
  rdf:_3 :curso_de_verão_verão_comp ;

:curso_de_verão_de_comp a decomp:Component ;
  decomp:correspondsTo :de .
```

Order

Component
Properties

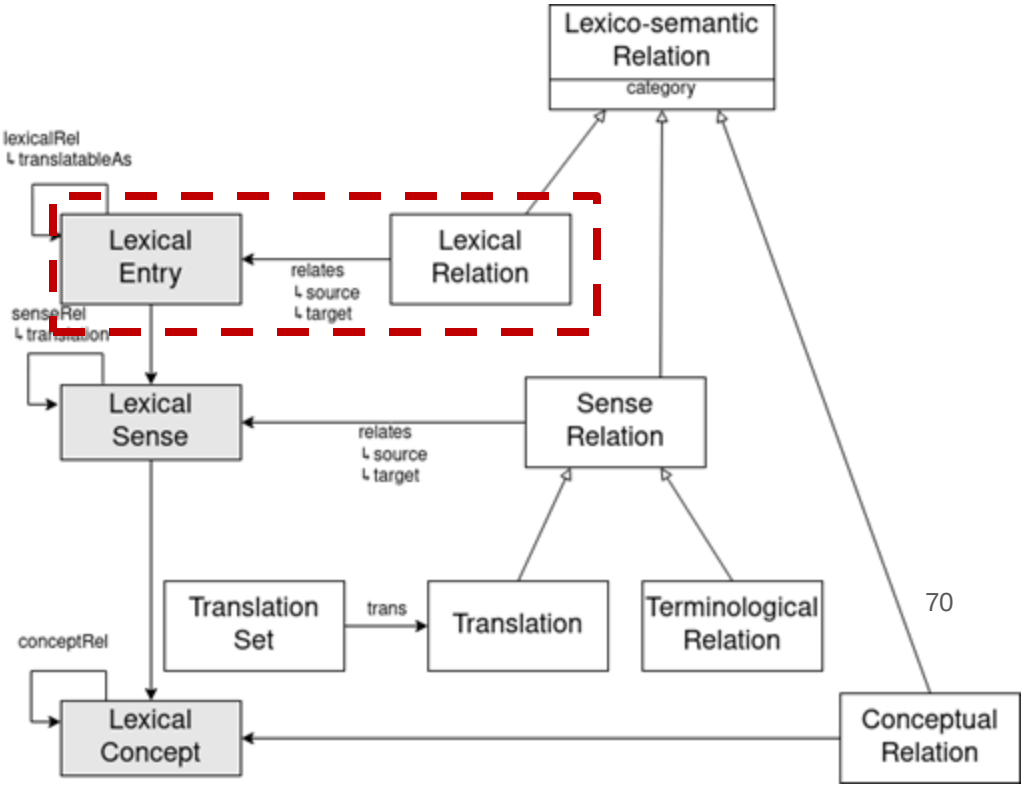
[Slide based on Gracia and Khan @ SDLLOD-22]

Vartrans Module



[Slide based on Gracia and Khan @ SDLLOD-22]

Vartrans Module

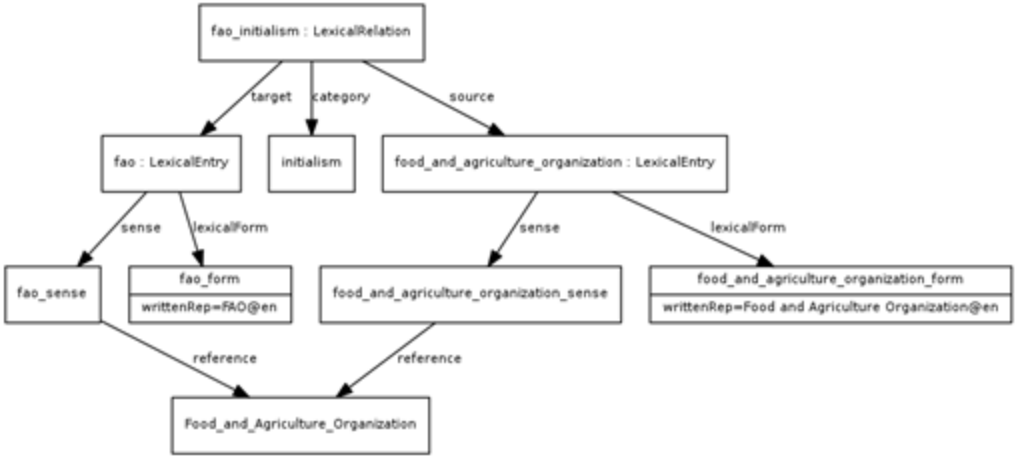


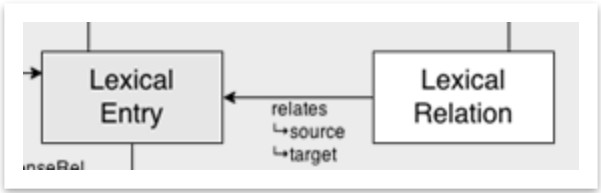
[Slide based on Gracia and Khan @ SDLLOD-22]

Lexical Relations

Examples of **lexical relations** include the following:

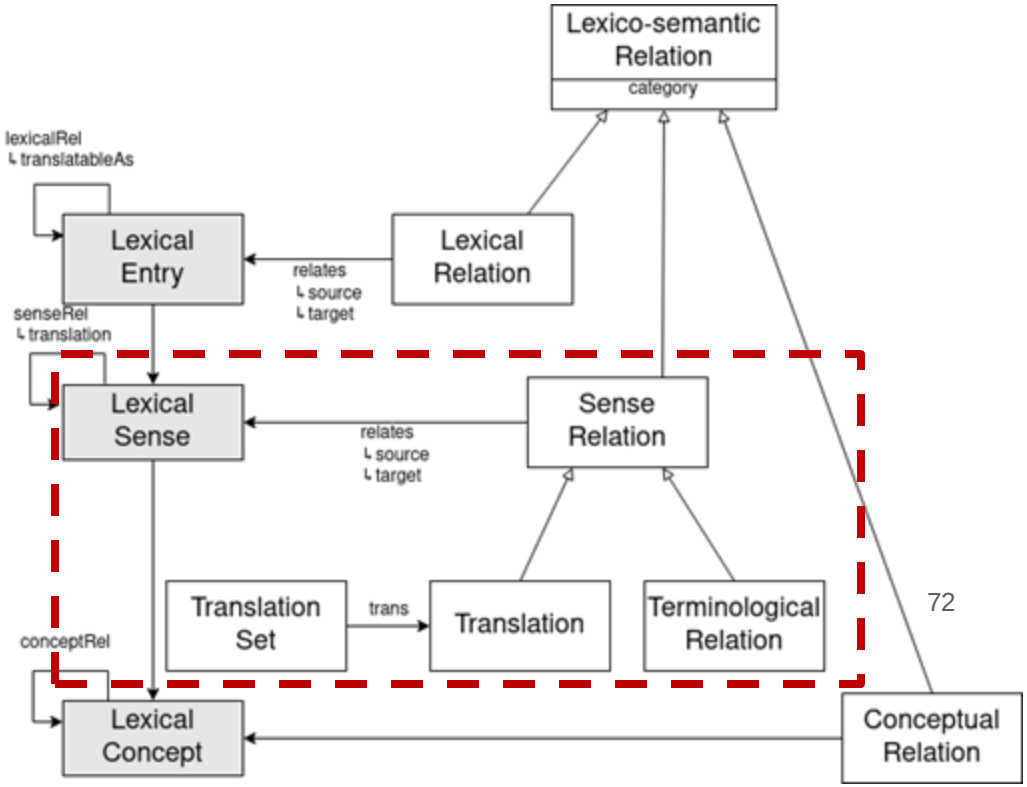
- **Derivational** relations (e.g., adjective → adverb variation: quick vs. quickly)
- **Morphosyntactic** relations (e.g. ecological tourism vs. eco-tourism)
- **Abbreviation** relations (including acronyms, e.g., peer to peer and p2p; WYSWYG, FAO, UNO)





[Slide based on Gracia and Khan @ SDLLOD-22]

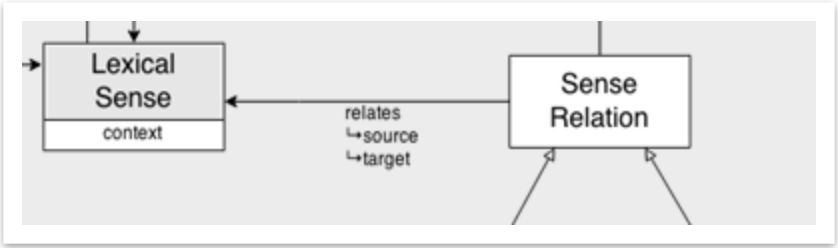
Vartrans Module



[Slide based on Gracia and Khan @ SDLLOD-22]

Semantic relations

- Examples of **semantic relations** are the **equivalence** relation between two senses, **hypernymy** and **hyponymy** relations, **synonymy**, **antonymy**, **translations**, etc..

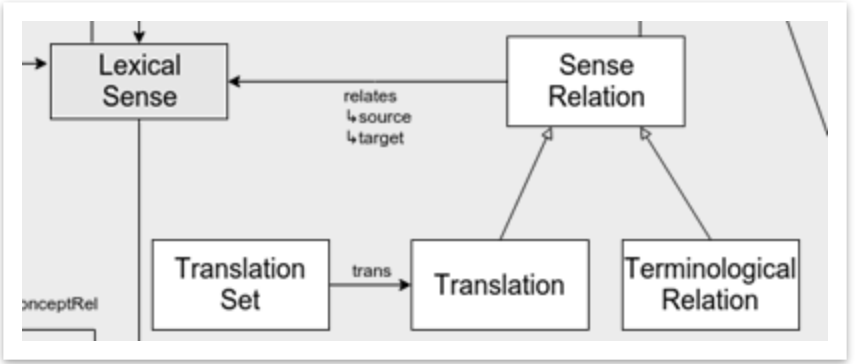


[Slide based on Gracia and Khan @ SDLLOD-22]

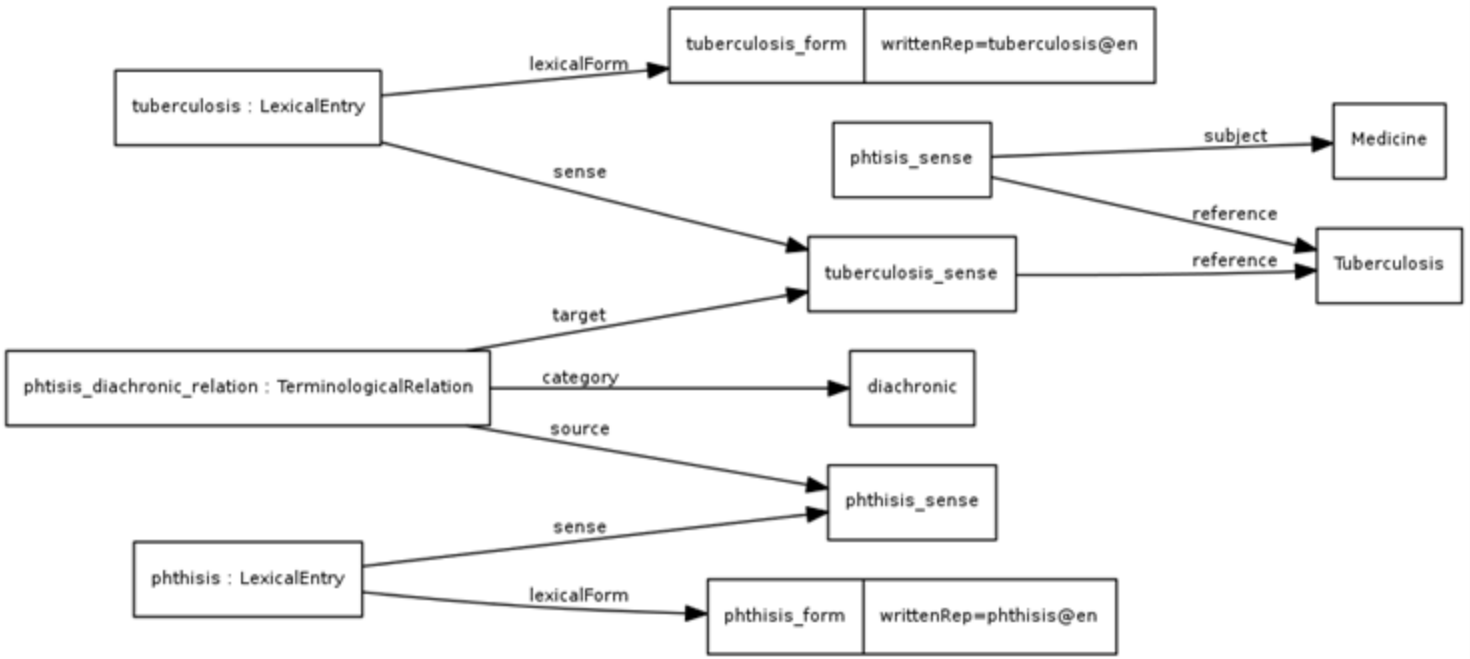
Terminological variants

Examples of categories of **terminological variants (terminological relations)** include:

- **Diatopic** (dialectal or geographical variants) (e.g., gasoline vs. petrol)
- **Diaphasic** (register) (e.g., headache vs. cephalalgia; swine flu vs. pig flu vs. H1N1 vs. Mexican pandemic flu)
- **Diachronic** (or chronological variants) (e.g., tuberculosis vs. phthisis)
- **Diastratic** (discursive or stylistic variants) (e.g., man vs. bloke)
- **Dimensional** variants: the terms point to the same concept but highlight a different property or dimension of the concept (e.g., bio-sanitary waste vs. hospital waste)

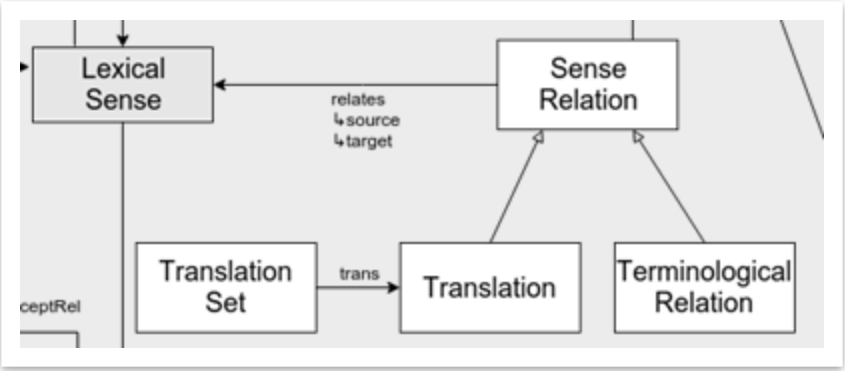


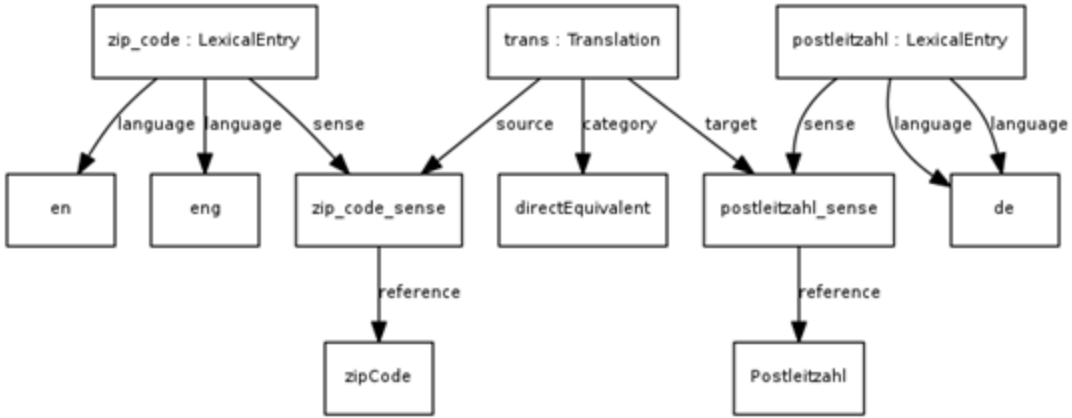
Terminological variants



[Slide based on Gracia and Khan @ SDLLOD-22]

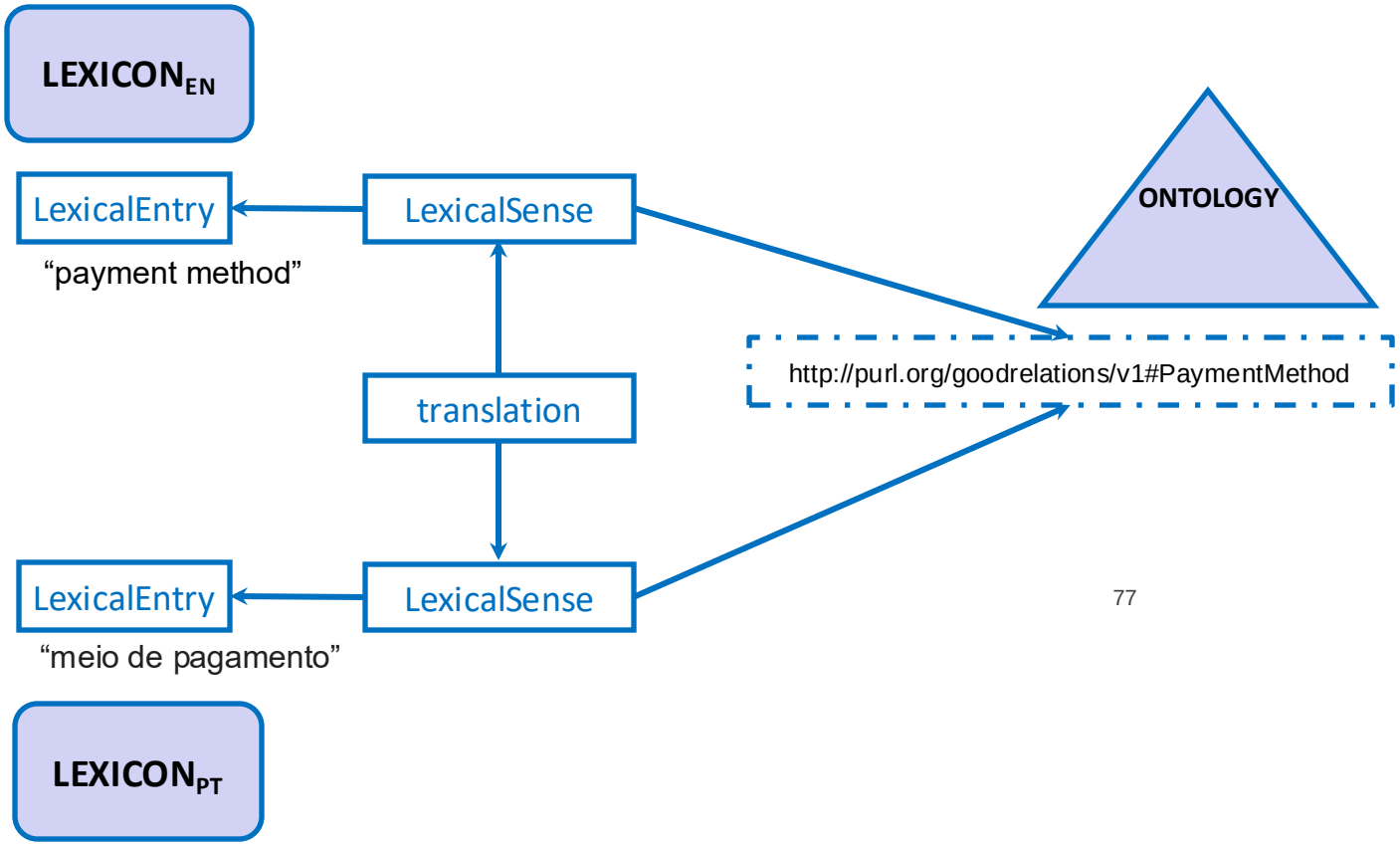
Translations





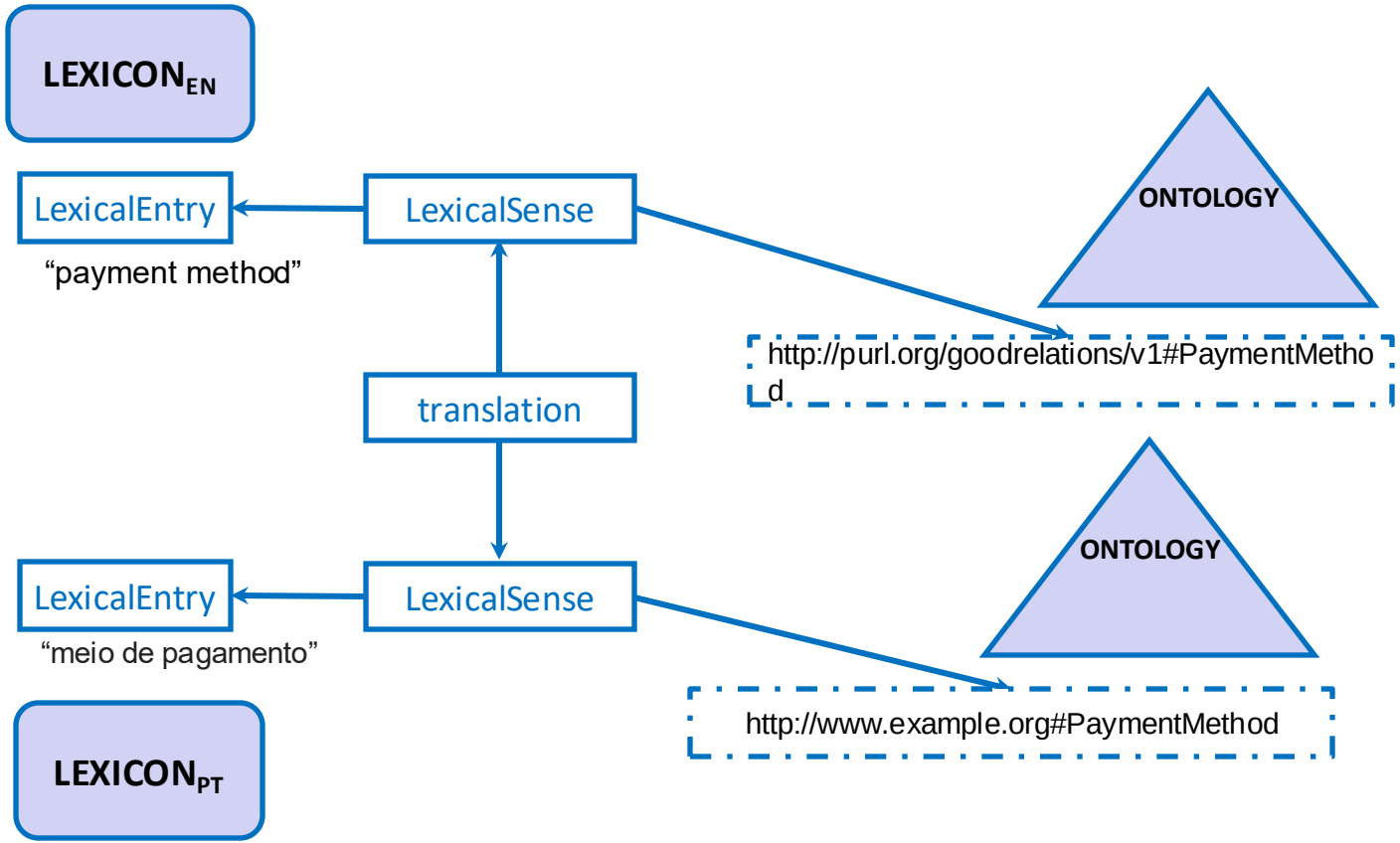
[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: translations



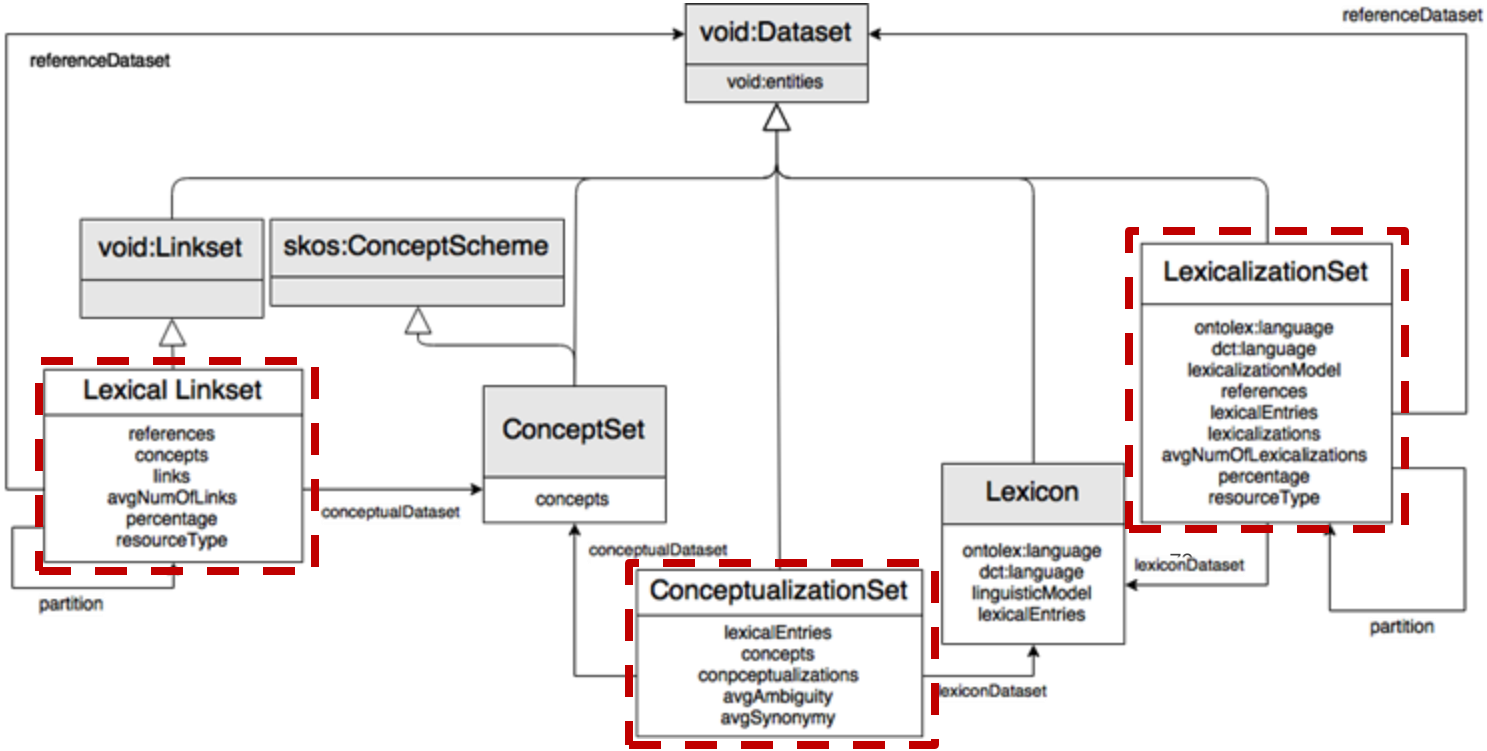
[Slide based on Gracia and Khan @ SDLLOD-22]

LLD – Ontolex-Lemon: translations



[Slide based on Gracia and Khan @ SDLLOD-22]

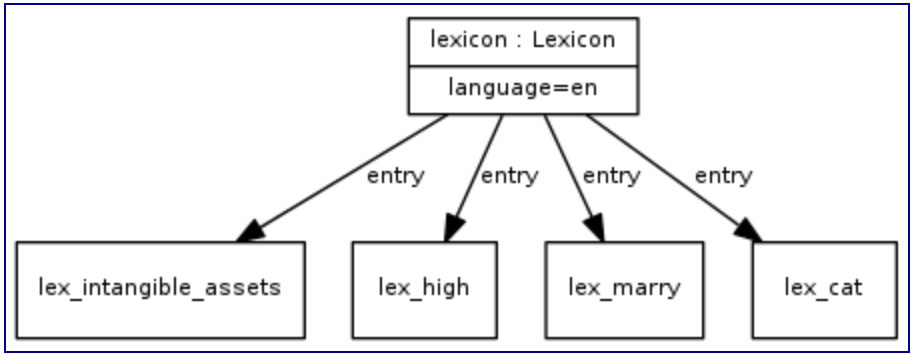
LIME



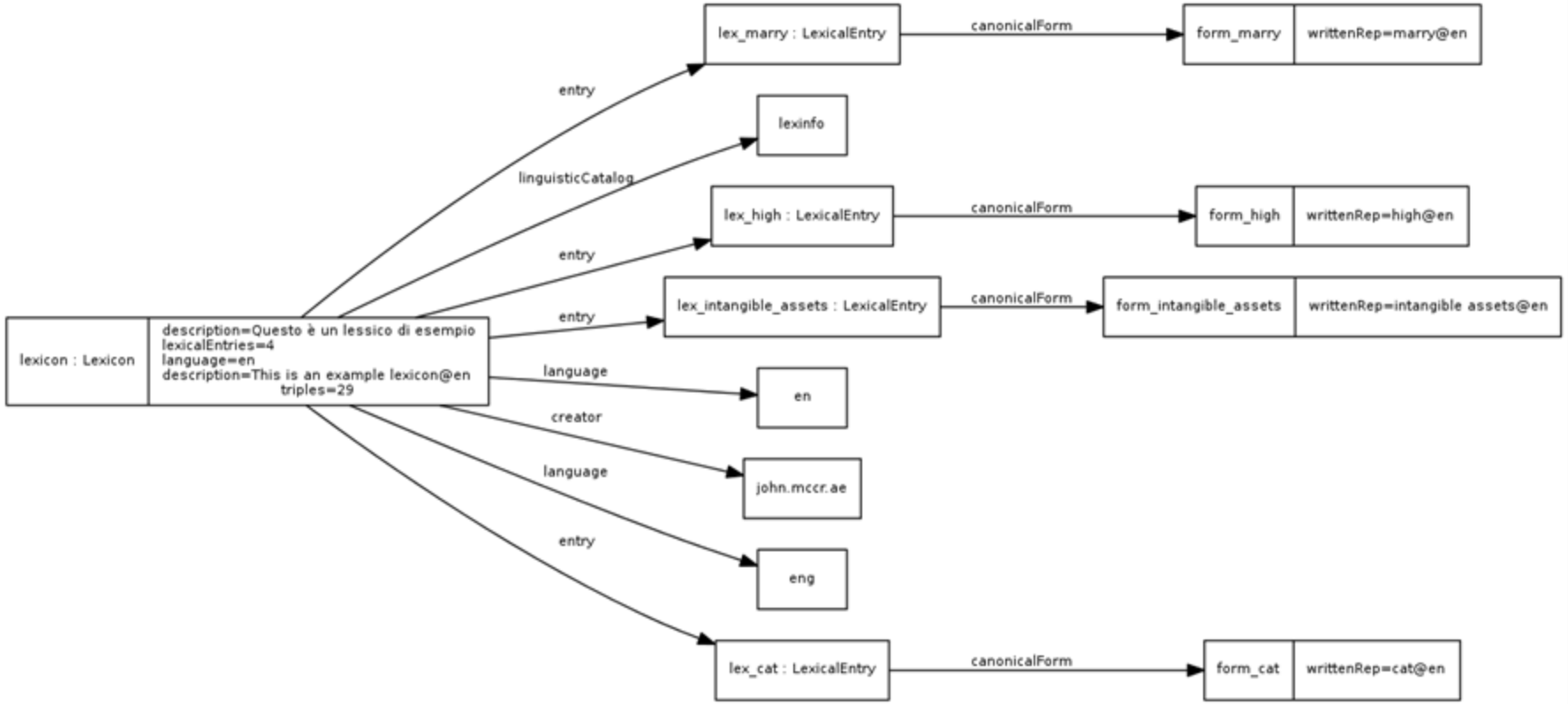
[Slide based on Gracia and Khan @ SDLLOD-22]

LIME

```
:lexicon a lime:Lexicon;
   lime:language "en";
   lime:entry :lex_high;
   lime:entry :lex_cat;
   lime:entry :lex_marry;

    lime:entry
:lex_intangible_assets.
```



LIME


lexicon : Lexicon
description=Questo è un lessico di esempio
lexicalEntries=4
language=en
description=This is an example lexicon@en
triples=29
entry
linguisticCatalog
entry
entry
language
creator
language
entry
lex_marry : LexicalEntry
canonicalForm
form_marry
writtenRep=marry@en
lexinfo
lex_high : LexicalEntry
canonicalForm
form_high
writtenRep=high@en
lex_intangible_assets : LexicalEntry
canonicalForm
form_intangible_assets
writtenRep=intangible assets@en
en
john.mccr.ae
eng
lex_cat : LexicalEntry
canonicalForm
form_cat
writtenRep=cat@en

Encoding an Example from Wiktionary

- Although there exists a linked data version of **Wiktionary**, **DBnary** (which we will look at soon), it doesn't encode all of the information in entries.
- We will look at the OntoLex model via the conversion of **a Wiktionary example to show the different facets of the model**
- We will take a non-European language, Urdu as an example, the word زبان(zaban) which means both 'tongue' and 'language'

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānĕ)
oblique	زبان (zubān)	زبانوں (zubānŏ)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Example from https://en.wiktionary.org/wiki/%D8%B2%D8%A8%D8%A7%D9%86#Urdu, January 2022

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

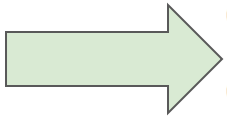
   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)



```
:زبان a ontolex:Word;
  lexinfo:gender lexinfo:feminine;
  lexinfo:partOfSpeech lexinfo:noun;
  lime:language "ur"^^xsd:language;
  ontolex:canonicalForm :زبان_lemma;
  ontolex:denotes <http://dbpedia.org/resource/Language>,
                  <http://dbpedia.org/resource/Tongue>;
  ontolex:otherForm :زبان_dir_pl,
                    :زبان_obl_pl,
                    :زبان_obl_sing,
                    :زبان_voc_pl,
                    :زبان_voc_sing;
  ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

```
:زبان a ontolex:Word;
    lexinfo:gender lexinfo:feminine;
    lexinfo:partOfSpeech lexinfo:noun;
    lime:language "ur"^^xsd:language;
    ontolex:canonicalForm :زبان_lemma;
    ontolex:denotes <http://dbpedia.org/resource/Language>,
                    <http://dbpedia.org/resource/Tongue>;
    ontolex:otherForm :زبان_dir_pl,
                      :زبان_obl_pl,
                      :زبان_obl_sing,
                      :زبان_voc_pl,
                      :زبان_voc_sing;
    ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Lexinfo properties and classes

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Lime module used for the language (the value of which is a string)

```
:زبان a ontolex:Word;
  lexinfo:gender lexinfo:feminine;
  lexinfo:partOfSpeech lexinfo:noun;
  lime:language "ur"^^xsd:language;
  ontolex:canonicalForm :زبان_lemma;
  ontolex:denotes <http://dbpedia.org/resource/Language>,
                  <http://dbpedia.org/resource/Tongue>;
  ontolex:otherForm :زبان_dir_pl,
                    :زبان_obl_pl,
                    :زبان_obl_sing,
                    :زبان_voc_pl,
                    :زبان_voc_sing;
  ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Lemma and other forms

```
:زبان a ontol
  lexinfo:gender  xinfo:feminine;
  lexinfo:partOfSpeech lexinfo:noun;
  dime:language "ur"^^xsd:language;
  ontolex:canonicalForm :زبان_lemma;
  ontolex:denotes <http://dbpedia.org/resource/Language>,
                  <http://dbpedia.org/resource/Tongue>;
  ontolex:otherForm :زبان_dir_pl,
                    :زبان_obl_pl,
                    :زبان_obl_sing,
                    :زبان_voc_pl,
                    :زبان_voc_sing;
  ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Semantics information

```
:زبان a ontol
  lexinfo:gender lexinfo:feminine;
  lexinfo:partOfSpeech lexinfo:noun;
  lime:language "ur"^^xsd:language;
  ontolex:canonicalForm :زبان_lemma;
  ontolex:denotes <http://dbpedia.org/resource/Language>,
                  <http://dbpedia.org/resource/Tongue>;
  ontolex:otherForm :زبان_dir_pl,
                    :زبان_obl_pl,
                    :زبان_obl_sing,
                    :زبان_voc_pl,
                    :زبان_voc_sing;
  ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Encoding of a morphological variant and sense

```
:زبان_dir_pl_form a ontolex:Form;
  lexinfo:number lexinfo:plural;
  ontolex:writtenRep "zubānē", "زبانیں"@ur .

:زبان_sense2 a ontolex:LexicalSense;
  ontolex:reference <https://dbpedia.org/resource/Language> .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

```
:MWE_مادری_زبان a ontolex:MultiWordExpression;
  lime:language "ur"^^xsd:language ;
  decomp:subterm :زبان,
                 :مادری .
```

Urdu [edit]

Etymology [edit]

Borrowed from Classical Persian زبان (*zubān*).

Pronunciation [edit]

- (*Standard Urdu*) IPA(key): /zʊ.bɑ:n/, /zə.bɑ:n/
- Rhymes: -ɑ:n

Noun [edit]

زُبان • (zubān) *f* (*Hindi spelling* ज़ुबान)

1. tongue (body part) [synonyms ▲]

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language [synonyms ▲]

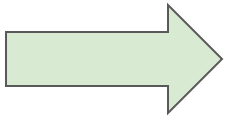
   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زبان (zubān)	زبانیں (zubānē)
oblique	زبان (zubān)	زبانوں (zubānō)
vocative	زبان (zubān)	زبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)



```
:translationRelation a vartrans:Translation;
  vartrans:category <http://purl.org/net/translation-categories#directEquivalent>;
  vartrans:source :زبان_sense2;
  vartrans:target :language_sense_1 .
```

```
:senseRelation a vartrans:SenseRelation;
  vartrans:category lexinfo:synonym;
  vartrans:source :بھاشا_sense1;
  vartrans:target :زبان_sense2 .
```

OntoLex - How Good is its Coverage?

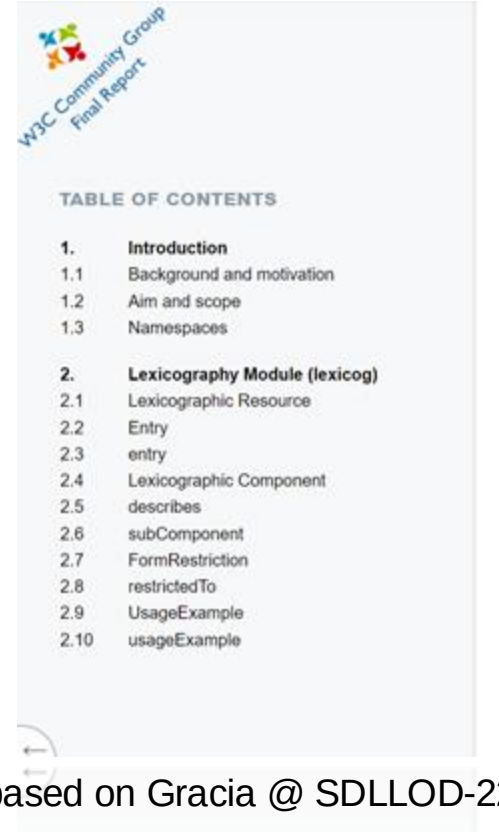
- Ontolex covers a wide range of use-cases but obviously can’t cover everything. For instance:
  - ***It is missing a module for more detailed descriptions of morphological processes such as word formation***
  - ***It is missing classes and properties for describing the relationships between lexicons and corpora, relating to e.g., attestations***
  - ***It lacks specialised vocabularies for modelling terminologies***
- In response to these use cases the W3C group is working on new follow up modules to the OntoLex Core: a specialised ***Morphpology*** module and a ***Frequency Corpus and Attestations*** module, both due to be published this year.
- Moreover, work towards a terminology module has also begun recently.

OntoLex - How Good is its Coverage?

- It also has some constraints that are problematic for covering dictionary resources, e.g., **OntoLex imposes one part of speech per entry.** In general it also **focuses on the content or a resource (the TEI lexical view) rather than its visual appearance or organisation**.
- This led to the publication of the **first extension** to the original OntoLex specifications to capture this kind of ‘lexicographic’ data

Ontolex lexicog module

https://www.w3.org/2019/09/lexicog/



The OntoLex Lemon Lexicography Module

Final Community Group Report 17 September 2019

Editors:

Julia Bosque-Gil (Ontology Engineering Group, Universidad Politécnica de Madrid)
Jorge Gracia (Aragon Institute of Engineering Research, University of Zaragoza)

Authors:

Julia Bosque-Gil (Ontology Engineering Group, Universidad Politécnica de Madrid)
Jorge Gracia (Aragon Institute of Engineering Research, University of Zaragoza)
John McCrae (Insight Centre for Data Analytics, National University of Ireland, Galway)
Philipp Cimiano (Cognitive Interaction Technology Excellence Center, Bielefeld University)
Sander Stolk (Centre for the Arts in Society, Leiden University)
Fahad Khan (Istituto di Linguistica Computazionale "Zampolli", CNR, Pisa)
Katrien Depuydt (Institute for Dutch Lexicology, Leiden, Netherlands)
Jesse de Does (Institute for Dutch Lexicology, Leiden, Netherlands)
Francesca Frontini (Paul-Valéry University, Montpellier III)
Ilan Kernerman (K Dictionaries)

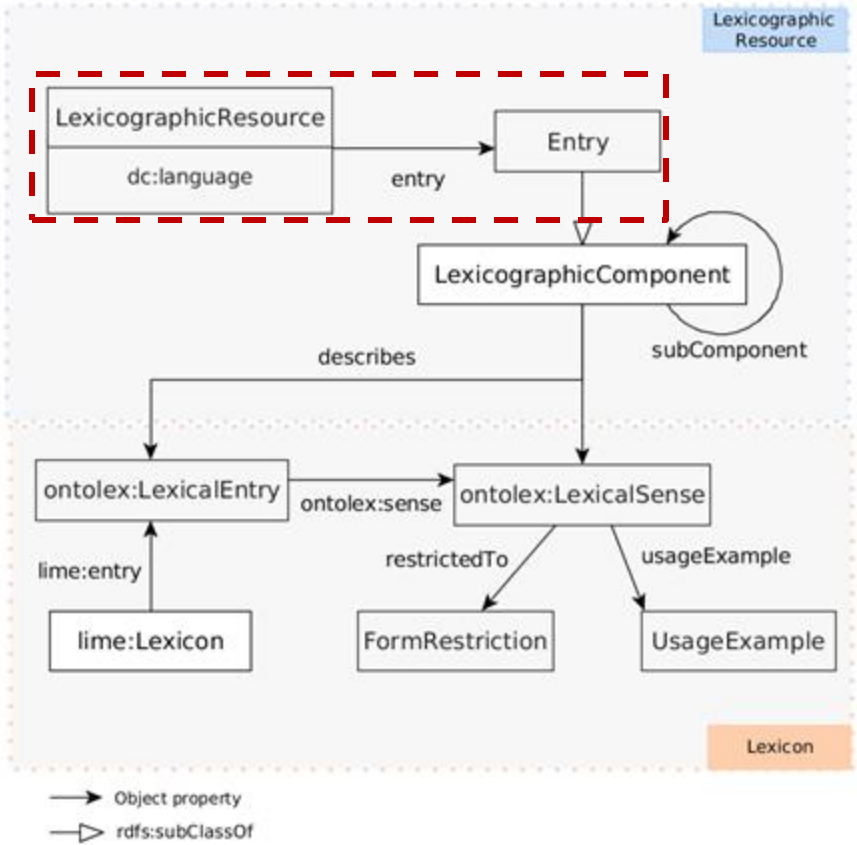
Abstract

This document describes the *lexicography module* of the Lexicon Model for Ontologies (*lemon*) as a result of the work of the Ontology Lexicon community group (OntoLex). The module is targeted at the representation of

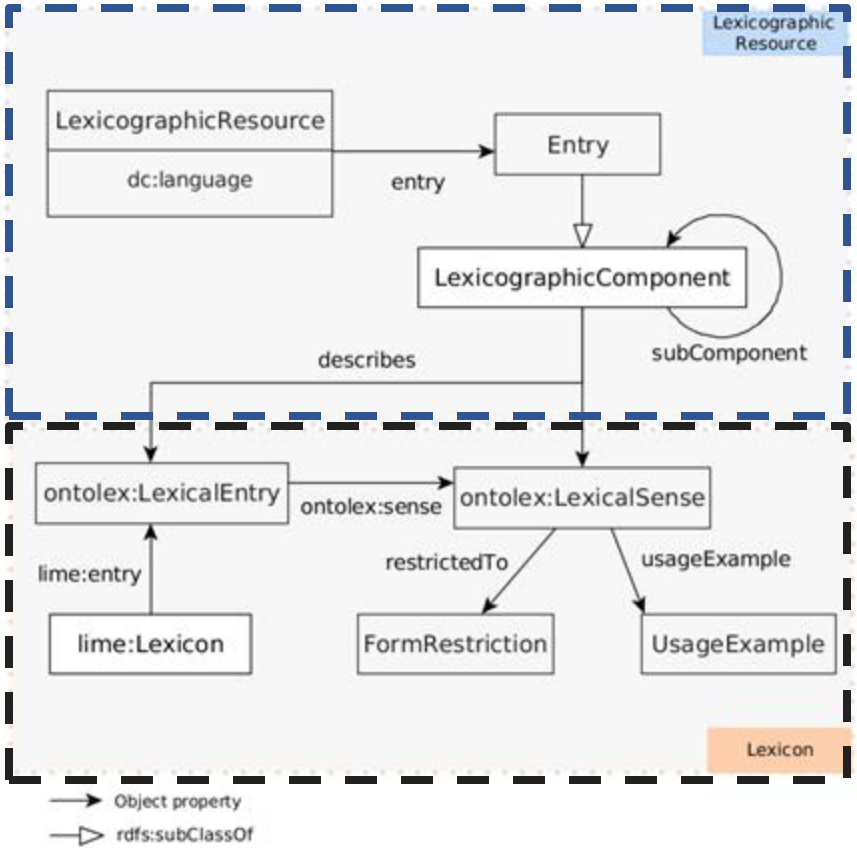
[Slide based on Gracia @ SDLLOD-22]

Lexicog

- The **OntoLex-Lemon Lexicography Module (lexicog)** developed by the W3C OntoLex group to represent some of the structural information “lost” in OntoLex-Lemon.
- It defines new classes such as **Lexicographic Resource** (complementing OntoLex **Lexicon**) which consists of single **Entry** individuals which represent lexicographic articles and which can be realised by OntoLex **Lexical Entry** elements.
- **Entry** is a subclass of **Lexicographic Component** which represents elements which describe the structuring of lexicographic articles.



[Slide based on Gracia @ SDLLOD-22]



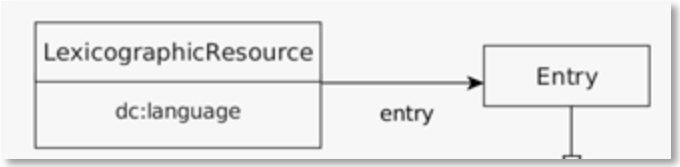
[Slide based on Gracia @ SDLLOD-22]

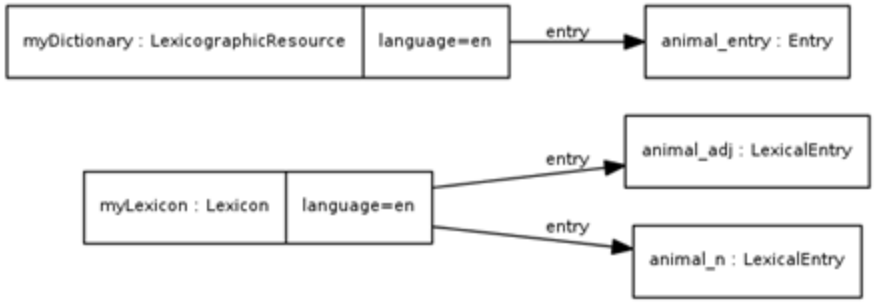
Example extracted from the American Heritage Dictionary

> **an·i·mal**
>
> n.
>
> 1. Any of numerous multicellular eukaryotic organisms of the kingdom Metazoa (or Animalia) [...]
>
> 2. An animal organism other than a human, especially a mammal.
>
> [...]
>
> adj.
>
> 1. Relating to, characteristic of, or derived from an animal or animals, especially when not human: animal cells; animal welfare.
>
> 2. Relating to the physical as distinct from the rational or spiritual nature of people: animal instincts and desires.

[Slide based on Gracia @ SDLLOD-22]

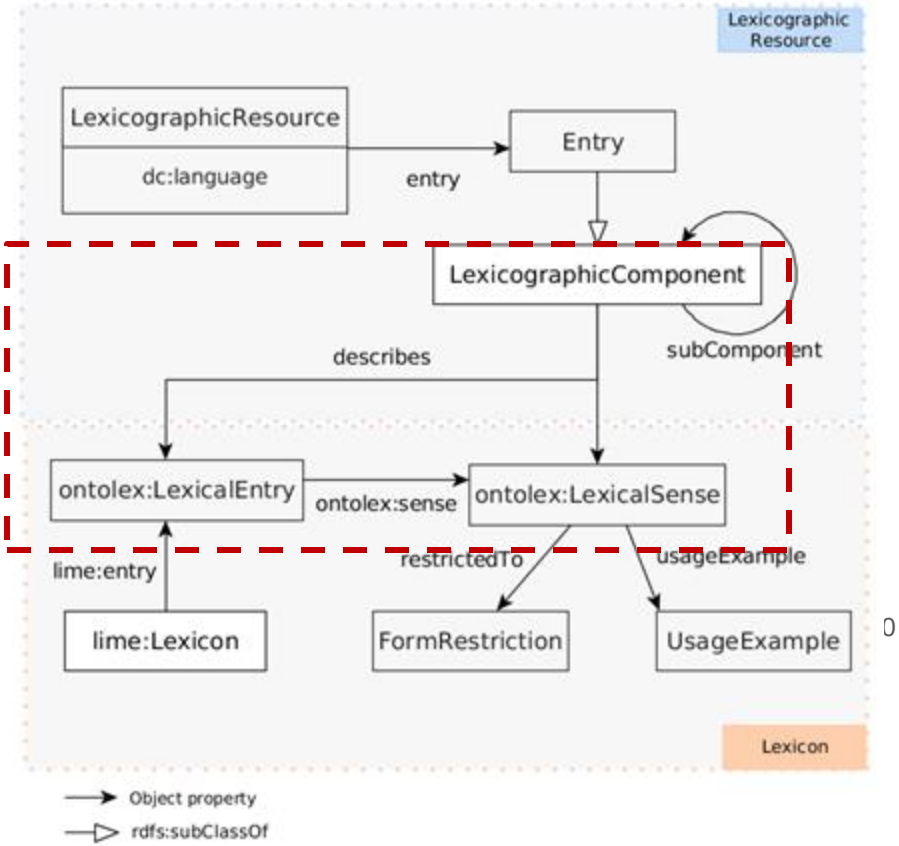
Ontolex lexicog module






```
# LEXICOGRAPHIC RESOURCE

:myDictionary a lexicog:LexicographicResource
;
        dc:language "en" ;
        lexicog:entry :animal_entry .

:animal_entry a lexicog:Entry .

# LEXICON

:myLexicon a lime:Lexicon;
        lime:language "en" ;
        lime:entry  :animal_n, :animal_adj .

:animal_n a ontolex:LexicalEntry .
:animal_adj a ontolex:LexicalEntry .
```

[Slide based on Gracia @ SDLLOD-22]



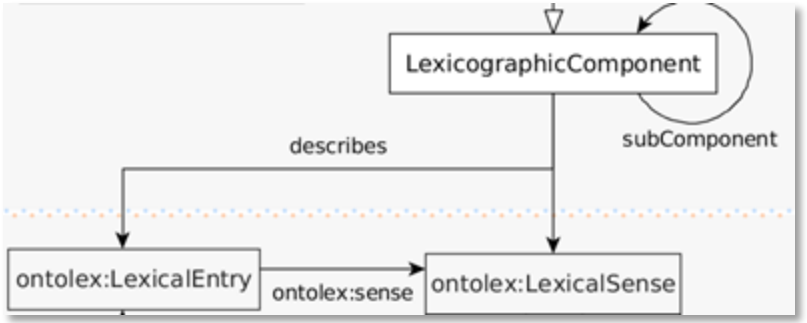
[Slide based on Gracia @ SDLLOD-22]

Ontolex lexicog module

A **lexicographic component** is a structural element that represents the (sub-)structures of lexicographic articles providing information about entries, senses or sub-entries. If desired, lexicographic components can be arranged in a specific order and/or hierarchy.

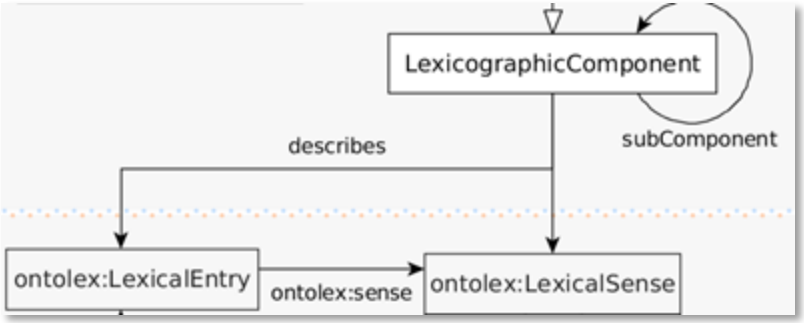
The property **`describes`** relates a lexicographic component to an element that represents the actual information provided by that component in the lexicographic resource. In most cases, this information will be lexical, and hence the object of the property will be an instance of `ontolex:LexicalEntry` or `ontolex:LexicalSense`.

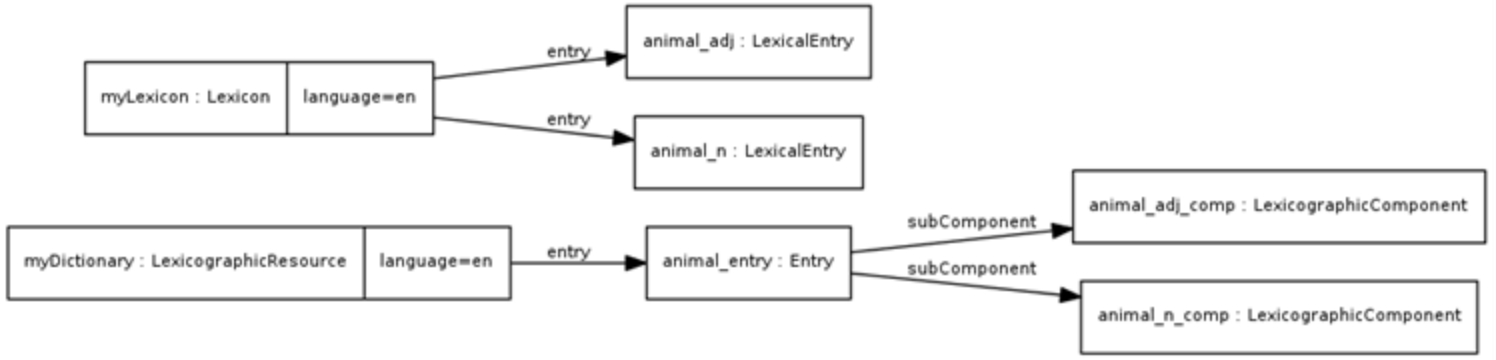
The property **`subComponent`** encodes a hierarchical relation between two lexicographic components



[Slide based on Gracia @ SDLLOD-22]

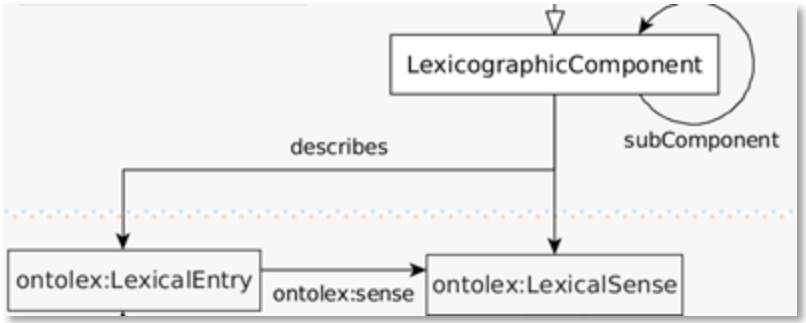
Ontolex lexicog module

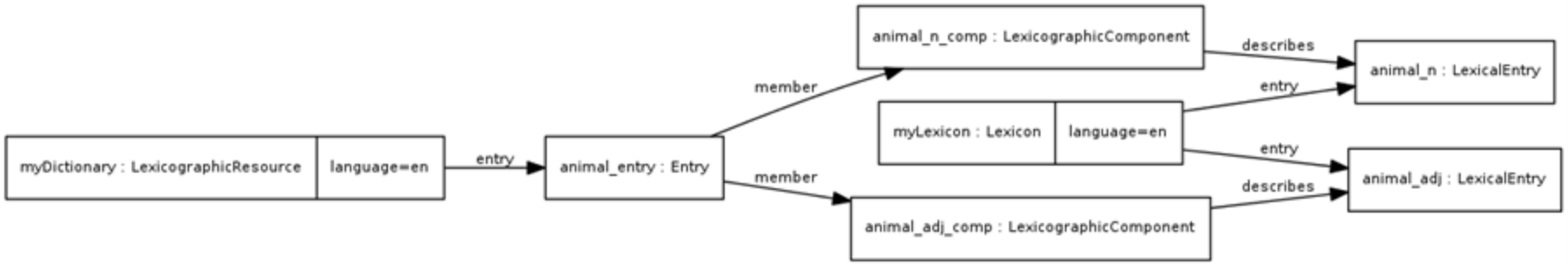




[Slide based on Gracia @ SDLLOD-22]

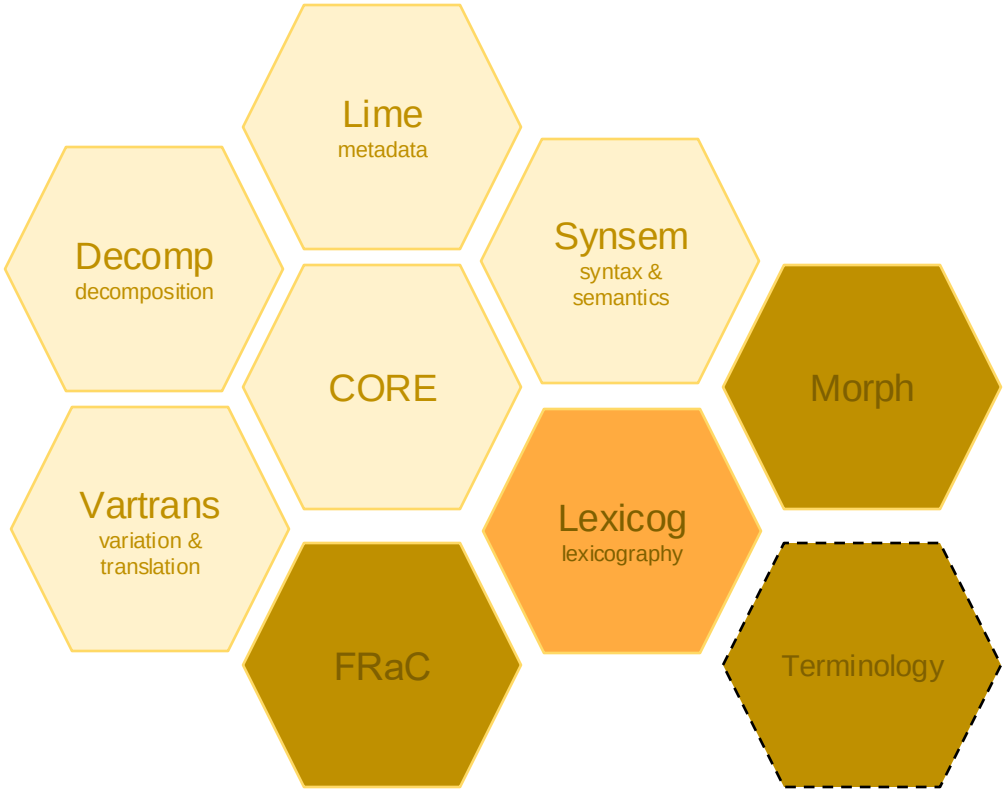
Ontolex lexicog module





[Slide based on Gracia @ SDLLOD-22]

Coming Soon



[Slide based on Gracia and Khan @ SDLLOD-22]

SPARQL

- The Wikidata schema for lexicographic data is based on OntoLex.
- We will now therefore look at how to write SPARQL queries to query Wikidata for lexicographic data.
- You can see an updated set of statistics lexicographic data contained in on
  - https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Statistics
- Note the mix of classes and properties from OntoLex and others which are Wikidata native

First Query

How many words in Wikidata in Italian ending in ‘a’ are feminine

```
SELECT DISTINCT ?l ?lemma ?word WHERE {
  ?l a ontolex:LexicalEntry ;
       dct:language wd:Q652 ;
       wikibase:lemma ?lemma ;
       wikibase:lexicalCategory wd:Q1084 .
  ?l wdt:P5185 wd:Q1775415.
  FILTER regex (?lemma, "a$").
}
```

[https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries]

Second Query

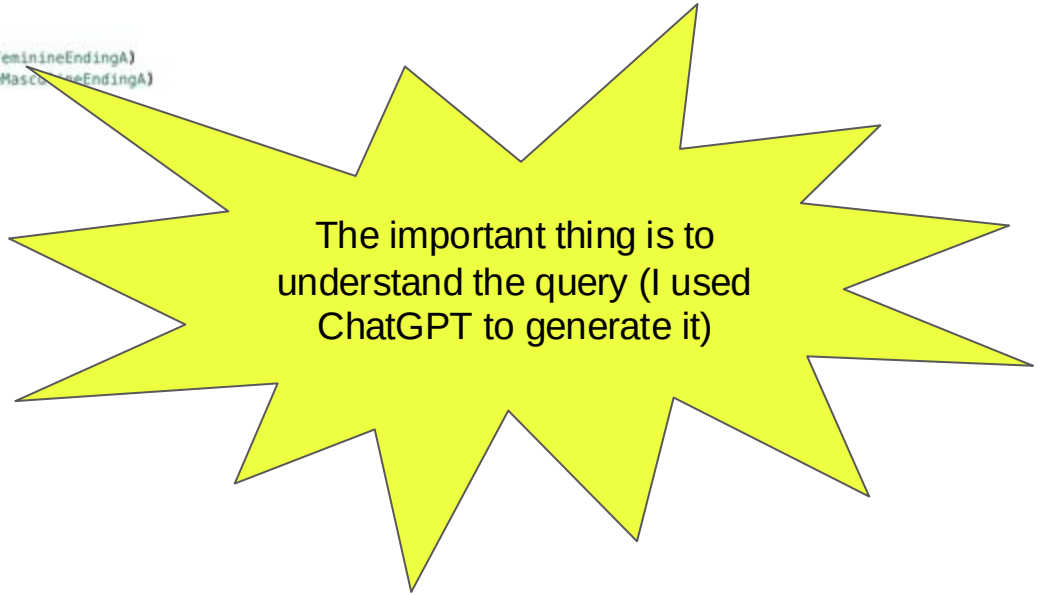
Modify the previous query to give the percentage of masculine nouns and feminine nouns which end in ‘a’ in Italian…

```
SELECT ?totalNouns ?feminineNounsEndingA ?masculineNounsEndingA
       ((?feminineNounsEndingA * 100.0) / ?totalNouns AS ?percentageFeminineEndingA)
       ((?masculineNounsEndingA * 100.0) / ?totalNouns AS ?percentageMasculineEndingA)
WHERE {
  {
    # total amount of nouns
    SELECT (COUNT(?l) AS ?totalNouns)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lexicalCategory wd:Q1084 .
    }
  }
  {
    SELECT (COUNT(?l) AS ?feminineNounsEndingA)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lemma ?lemma ;
         wikibase:lexicalCategory wd:Q1084 .
      ?l wdt:P5185 wd:Q1775415 . # Feminine gender
      FILTER regex (?lemma, "a$").
    }
  }
  {
    SELECT (COUNT(?l) AS ?masculineNounsEndingA)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lemma ?lemma ;
         wikibase:lexicalCategory wd:Q1084 .
      ?l wdt:P5185 wd:Q499327 . # Masculine gender
      FILTER regex (?lemma, "a$").
    }
  }
}
```

Second Query

Modify the previous query to give the percentage of masculine nouns and feminine nouns which end in ‘a’ in Italian…

```
SELECT ?totalNouns ?feminineNounsEndingA ?masculineNounsEndingA
       ((?feminineNounsEndingA * 100.0) / ?totalNouns AS ?percentageFeminineEndingA)
       ((?masculineNounsEndingA * 100.0) / ?totalNouns AS ?percentageMasculineEndingA)
WHERE {
  {
    # total amount of nouns
    SELECT (COUNT(?l) AS ?totalNouns)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lexicalCategory wd:Q1084 .
    }
  }
  {
    SELECT (COUNT(?l) AS ?feminineNounsEndingA)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lemma ?lemma ;
         wikibase:lexicalCategory wd:Q1084 .
      ?l wdt:P5185 wd:Q1775415 . # Feminine gender
      FILTER regex (?lemma, "a$").
    }
  }
  {
    SELECT (COUNT(?l) AS ?masculineNounsEndingA)
    WHERE {
      ?l a ontolex:LexicalEntry ;
         dct:language wd:Q652 ;
         wikibase:lemma ?lemma ;
         wikibase:lexicalCategory wd:Q1084 .
      ?l wdt:P5185 wd:Q499327 . # Masculine gender
      FILTER regex (?lemma, "a$").
    }
  }
}
```



Anatomy of the Query

Modify the previous query to give the percentage of masculine nouns and feminine nouns which end in ‘a’ in Italian…

```
SELECT ?totalNouns ?feminineNounsEndingA ?masculineNounsEndingA
    ((?feminineNounsEndingA * 100.0) / ?totalNouns AS ?percentageFeminineEndingA)
    ((?masculineNounsEndingA * 100.0) / ?totalNouns AS ?percentageMasculineEndingA)
WHERE {
{
SUBQUERY TO FIND THE COUNT OF ALL NOUNS IN ITALIAN -> ?totalNoun
SUBQUERY TO FEMININE NOUNS IN ITALIAN -> ?feminineNounsEndingA
SUBQUERY TO MASCULINE NOUNS IN ITALIAN -> ?percentageMasculineEndingA
}
```

First subquery

SUBQUERY TO FIND THE COUNT OF ALL NOUNS IN ITALIAN -> ?totalNoun

```
SELECT (COUNT(?l) AS ?totalNouns)
  WHERE {
    ?l a ontolex:LexicalEntry ;
       dct:language wd:Q652 ;
       wikibase:lexicalCategory wd:Q1084 .
  }
```

Second subquery

SUBQUERY TO FIND THE COUNT OF ALL NOUNS IN ITALIAN -> ?totalNoun

```
SELECT (COUNT(?l) AS ?feminineNounsEndingA)
  WHERE {
    ?l a ontolex:LexicalEntry ;
       dct:language wd:Q652 ;
       wikibase:lemma ?lemma ;
       wikibase:lexicalCategory wd:Q1084 .
    ?l wdt:P5185 wd:Q1775415 . # Feminine gender
    FILTER regex (?lemma, "a$").
  }
```

Third subquery

SUBQUERY TO MASCULINE NOUNS IN ITALIAN -> ?percentageMasculineEndingA

```
SELECT (COUNT(?l) AS ?masculineNounsEndingA)
   WHERE {
     ?l a ontolex:LexicalEntry ;
        dct:language wd:Q652 ;
        wikibase:lemma ?lemma ;
        wikibase:lexicalCategory wd:Q1084 .
     ?l wdt:P5185 wd:Q499327 . # Masculine gender
     FILTER regex (?lemma, "a$").
   }
```

Fourth Query

Words that are different in American English and British English

```
SELECT  ?l ?english ?american
WHERE {
    ?l wikibase:lemma ?english .  FILTER(LANG(?english)="en-gb")
    ?l wikibase:lemma ?american . FILTER(LANG(?american)="en")
    FILTER(?english!=?american)
}
ORDER BY ?english
```

[https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries]

Fourth Query

Your turn….write a similar query for Brazilian Portuguese and European Portuguese! (or your two favourite language variants)

DBnary

- DBnary is a linguistic linked data resource extracted from Wiktionary, the collaboratively edited online dictionary.
- The project aims to transform the rich lexical information available in Wiktionary into a structured and machine-readable format using RDF (Resource Description Framework).
  - Not all of the information in Wiktionary entries is yet available in DBnary
- Uses OntoLex Lemon as its main ontology
- Updated periodically to keep track of changes in Wiktionary
- SPARQL endpoint @ https://kaiko.getalp.org/sparql

Pattern for a DBnary Query

Find the basic information (part of speech, forms, definitions) for a list of words in a language from DBnary?

```
SELECT DISTINCT ?le AS ?lexical_entry, ?pos AS ?part_of_speech, ?r AS ?form, ?d AS ?definition
WHERE {
  ?le a ontolex:LexicalEntry ;
  dbnary:partOfSpeech ?pos ;
   ontolex:canonicalForm [ontolex:writtenRep ?r] ;
   lime:language LANGUAGE TAG;
   ontolex:sense  [skos:definition ?s].
 ?s rdf:value ?d .
 FILTER( LIST OF WORDS TO EXTRACT    )
 FILTER (lang(?d) = LANGUAGE TAG)
}
```

Portuguese Example

```
SELECT DISTINCT ?le AS ?lexical_entry, ?pos AS ?part_of_speech, ?r AS ?form, ?d AS ?definition

WHERE {

  ?le a ontolex:LexicalEntry ;

  dbnary:partOfSpeech ?pos ;

   ontolex:canonicalForm [ontolex:writtenRep ?r] ;

   lime:language "pt";

   ontolex:sense  [skos:definition ?s].

 ?s rdf:value ?d .

 FILTER(?r = "agosto"@pt || ?r = "aia"@pt || ?r = "alfinete"@pt || ?r = "ananás"@pt || ?r = "armário"@pt || ?r = "balde"@pt || ?r = "baptismo"@pt || ?r = "botelha"@pt
|| ?r = "câmara"@pt || ?r = "carraguem"@pt || ?r = "chave"@pt || ?r = "cristão"@pt || ?r = "cruz"@pt || ?r = "dezembro"@pt || ?r = "espada"@pt || ?r = "estábulo"@pt
|| ?r = "estirar"@pt || ?r = "fita"@pt || ?r = "hospital"@pt || ?r = "igreja"@pt || ?r = "inglês"@pt || ?r = "leilão"@pt || ?r = "martelo"@pt || ?r = "masto"@pt || ?r =
"mestre"@pt || ?r = "outubro"@pt || ?r = "padre"@pt || ?r = "pipa"@pt || ?r = "pistola"@pt || ?r = "pompa"@pt || ?r = "saia"@pt || ?r = "salsaparrilha"@pt || ?r =
"setembro"@pt || ?r = "varanda"@pt)

 FILTER (lang(?d) = 'pt')

}
```

Hindi Example

```
SELECT DISTINCT ?le AS ?lexical_entry, ?pos AS ?part_of_speech, ?r AS ?form, ?d AS ?definition

WHERE {

  ?le a ontolex:LexicalEntry ;

  dbnary:partOfSpeech ?pos ;

   ontolex:canonicalForm [ontolex:writtenRep ?r] ;

   lime:language "hi";

   ontolex:sense  [skos:definition ?s].

 ?s rdf:value ?d .

FILTER(LANG(?d) = "en" )

 FILTER(?r = "अगस्त"@hi || ?r = "आया"@hi || ?r = "आलपीन"@hi || ?r = "अनन्नास"@hi || ?r = "अलमारी"@hi || ?r = "बालटी"@hi || ?r = "बाप्तिस्मा"@hi || ?r = "बोतल"@hi || ?r = "कमरा"@hi || ?r = "किराँची"@hi || ?r = "चाबी"@hi || ?r = "क्रिस्तान"@hi || ?r = "क्रूस"@hi || ?r = "दिसंबर"@hi || ?r = "इस्पात"@hi || ?r = "अस्तबल"@hi || ?r = "इस्तरी"@hi || ?r = "फ़ीता"@hi || ?r = "अस्पताल"@hi || ?r = "गिरजा"@hi || ?r = "अँग्रेज़"@hi || ?r = "नीलाम"@hi || ?r = "मारतौल"@hi || ?r = "मस्तूल"@hi || ?r = "मिस्तरी"@hi || ?r = "अक्तूबर"@hi || ?r = "पादरी"@hi || ?r = "पीपा"@hi || ?r = "पिस्तौल"@hi || ?r = "बंबा"@hi || ?r = "साया"@hi || ?r = "सालसा"@hi || ?r = "सितंबर"@hi || ?r = "बरामदा"@hi)

}
```

SPARE SLIDES

Excursus: Encoding Dictionaries as Complex Objects using Semantic Web Ontologies

Using Ontologies to Model Texts

Linked data ontologies already used in modeling **cultural heritage data**:

- E.g., **CIDOC-CRM** has been successfully used in several projects including aligning museum catalogues and archaeological datasets

There already exist linked data ontologies/vocabularies for textual metadata which allow for the description of **bibliographic information for textual works**:

- The project "**Mapping the Manuscript Migrations**" is a good example of the impact that linked data + ontologies can have

However, ontologies like CIDOC-CRM offer the possibility of modeling texts **as complex objects** and integrating seemingly contradictory properties.

Using Ontologies to Model Texts

Modelling texts is **challenging** due to their dual nature as physical and as information.

Texts are associated with a physical support, these physical supports can be located in different geographical locations, as well as being subject to various physical processes, such objects can have a fascinating history in their own right (see the MMM project).

On the other hand they also have an (informational) content that can, e.g., be translated into different languages or adapted in different media.

Ontologies provide **a principled way of describing and reasoning about such entities**.

In the world of ontology engineering we call such kinds of multifaceted entities, **informational entities**. These are complex ontological objects that have **a physical form and carry informational content**.

Using Ontologies to Model Texts

Informational entities are related to ***dot objects*** first proposed by the linguist James Pusetejovksy in order to model phenomena such as **co-predication:**

> *"The blue dictionary has more understandable but less comprehensive definitions than the red one, that's why it's lighter!"*
>
> *“The dictionary is outdated and very often incorrect in its etymological analyses but the definitions can be amusing and it makes a nice doorstop.”*

As well as books, other examples of dot objects include **countries**, **institutions**, **diseases**.

Some ontologists argue for the introduction of **separate complex categories** in ontologies to account for dot objects. These categories could be defined using a modified version of the coincidence relation, used to model situations like those described by the clay and statue paradox.

Using Ontologies to Model Texts

Some aspects of texts are difficult to model using already existing ontologies (and formal ontology languages):

> *What are the **arguments** of the text? What is the **plot** of **a literary work**? What are the main **themes** of a novel? What **literary devices** does it make use of?*

Lack of agreement on shared vocabularies and ontologies for describing such properties is a hurdle to modeling texts using linked data ontologies in general.

However certain types of texts can be modeled using already existing ontologies, and **dictionaries/lexicographic resources** are one such example.

Why Lexicographic Resources?

The creation of digital descriptions/versions of ***any*** kind of text confronts us with the distinction between the **content of a text**, and how the **content** is **presented**. Dictionaries are an interesting case: they tend organise **similar kinds** of (**linguistic) information** in **standardised ways**.

Moreover this (linguistic) content can be represented (in a formal way) much more easily than in other cases, e.g., plays, novels, encyclopedias, etc. This makes them **a useful test case in the modelling of texts using ontologies**.

To a large extent we can combine existing vocabularies to model dictionaries **as complex ontological objects**

Encoding Dictionaries as Structured Datasets

What kinds of things can we potentially encode in a linked data edition of a dictionary using ontologies?

- **Metadata** common to other texts can be encoded using existing vocabularies such as **Dublin Core** and **DCAT**.
- Descriptions specific to legacy printed texts, such as **number of pages** and **fonts used**
- Dictionary entries provide information on morpho-syntactic properties of words, **citations**, **examples**, and **etymologies** which can be represented as knowledge graphs

The **extraction** of this information can be done using machine learning methods; ontologies can be used to create **schemas** 'templates' for the information. But the semantics of this information isn't always straightfoward (challenge of what to encode/leave out). In the next few slides we look at some of the complexities that information organisation in dictionaries can present.

Citation: An Anomalous Example

Citations can be used to *attest* various different properties of a lexical entry, e.g., **orthographic**, **semantic**, **phonetic**. But they can also be used for other purposes.

We will look at the entry for **ἀνώμᾰλος** (anomalos) from the hugely influential Liddell-Scott-Jones ancient Greek-English lexicon (made available online by the **Perseus project**).

An Anomalous Example

ἀνώμα^λ-ος , ον, (ἀ- priv., ὁμαλός)

A.*uneven, irregular*, "χώρα" **Pl.Lg.625d**; "φύσις" **Id.Ti.58a**; "τὸ ἀ. τῆς ναυμαχίας" **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. "-λως, κινεῖσθαι"**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

II. of conditions, fortune, and the like , "φεῦ τῶν βροτείων ὡς ἀ. τύχαι" **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; "θέα" **Plot.6.7.34**. Adv. "-λως" **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** ***D***.

III. of persons, *inconsistent, capricious*, "ὁμαλῶς ἀ." **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; "πίθηκος" **Phryn. Com.20**; "τύχη" **AP10.96.** Adv. "-λως" **Isoc. 9.44**.

An Anomalous Example

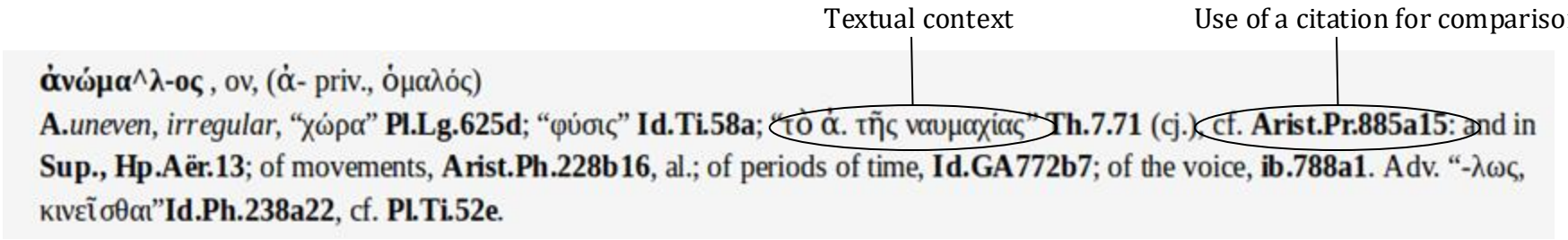
ἀνώμα^λ-ος , ον, (ἀ- priv., ὁμαλός)

A.*uneven, irregular*, "χώρα" **Pl.Lg.625d**; "φύσις" **Id.Ti.58a**; "τὸ ἀ. τῆς ναυμαχίας" **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. "-λως, κινεῖσθαι"**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

II. of conditions, fortune, and the like , "φεῦ τῶν βροτείων ὡς ἀ. τύχαι" **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; "θέα" **Plot.6.7.34**. Adv. "-λως" **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** ***D***.

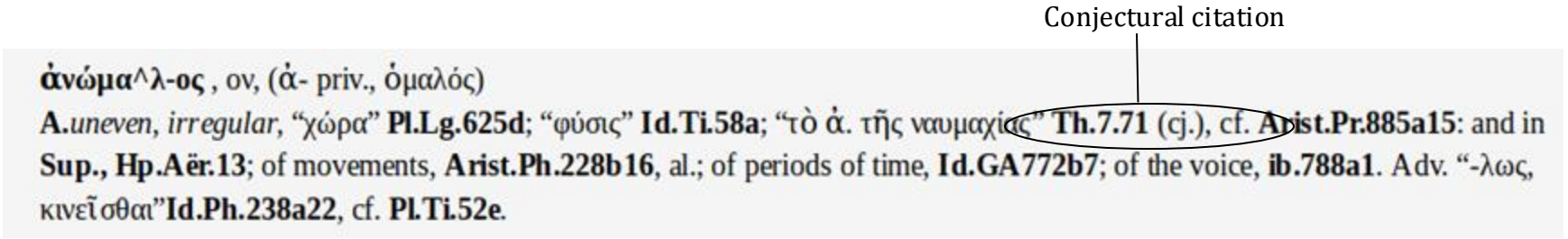
III. of persons, *inconsistent, capricious*, "ὁμαλῶς ἀ." **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; "πίθηκος" **Phryn. Com.20**; "τύχη" **AP10.96.** Adv. "-λως" **Isoc. 9.44**.

An Anomalous Example



Most of the citations in the example are used to *attest* to different shades of meaning of the word in question, with the **textual context** of an attestation **explicitly given** in one case. In other cases citations are used to contrast with other citations: without necessarily attesting to the word sense being dealt with. This use of the citation is annotated by the abbreviation '**cf.**'.

An Anomalous Example



It is also interesting to note that one of the citations, **'Th.7.71'**, is marked with a **'(cj.)'** meaning that it is conjectural -- i.e., it is based on **a reconstruction of the original text**. In this case we can say that the entry cites the text (from the corpus of works attributed to Thucydides) even though the original text might not have actually attested the sense itself.

An example etymological entry

GIRL, a female child, young woman. (E.) ME. gerle, girle, gyrle, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) gerl is a young woman; but in C.T. 666 (A 664), the pl. girles means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means ‘young women;’ in P. Plowman, B.i.33, it means ‘boys;’ cf. B. x. 175. Answering to an AS. form *gyr-el-, Teut. *gur-wil-, a dimin. form from Teut. base *gur-. Cf. NFries. gor, a girl; Pomeran. goer, a child; O. Low G. gor, a child; see Bremen Wortebuch, ii. 528. Cf. Swiss gurre, gurrli,a depriciatory term for a girl; Sanders, G. Dict. i. 609, 641; also Norw. gorre, a small child (Aasen); Swed. dial. garra, guerre (the same). Root uncertain. Der. girl-ish, girlish-ly, girl-ish-ness, girl-hood.

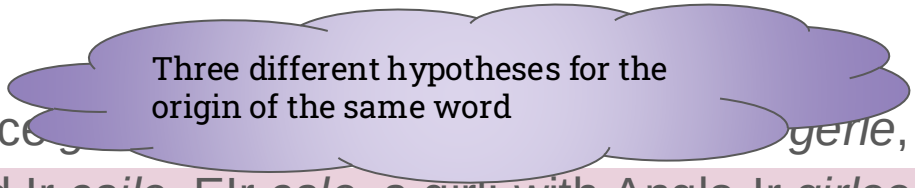
An example etymological entry

GIRL, a female child, young woman. (E.) ME. gerle, girle, gyrle, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) gerl is a young woman; but in C.T. 666 (A 664), the pl. girles means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means ‘young women;’ in P. Plowman, B.i.33, it means ‘boys;’ cf. B. x. 175. Answering to an AS. form *gyr-el-, Teut. *gur-wil-, a dimin. form from Teut. base *gur-. Cf. NFries. gor, a girl; Pomeran. goer, a child; O. Low G. gor, a child; see Bremen Wortebuch, ii. 52 riciatory term for a girl; Sanders, G. Dict. i. 609, ild (Aasen); Swed. dial. garra, guerre (the same). Root uncertain. girl-ish, girlish-ly, girl-ish-ness, girl-hood.

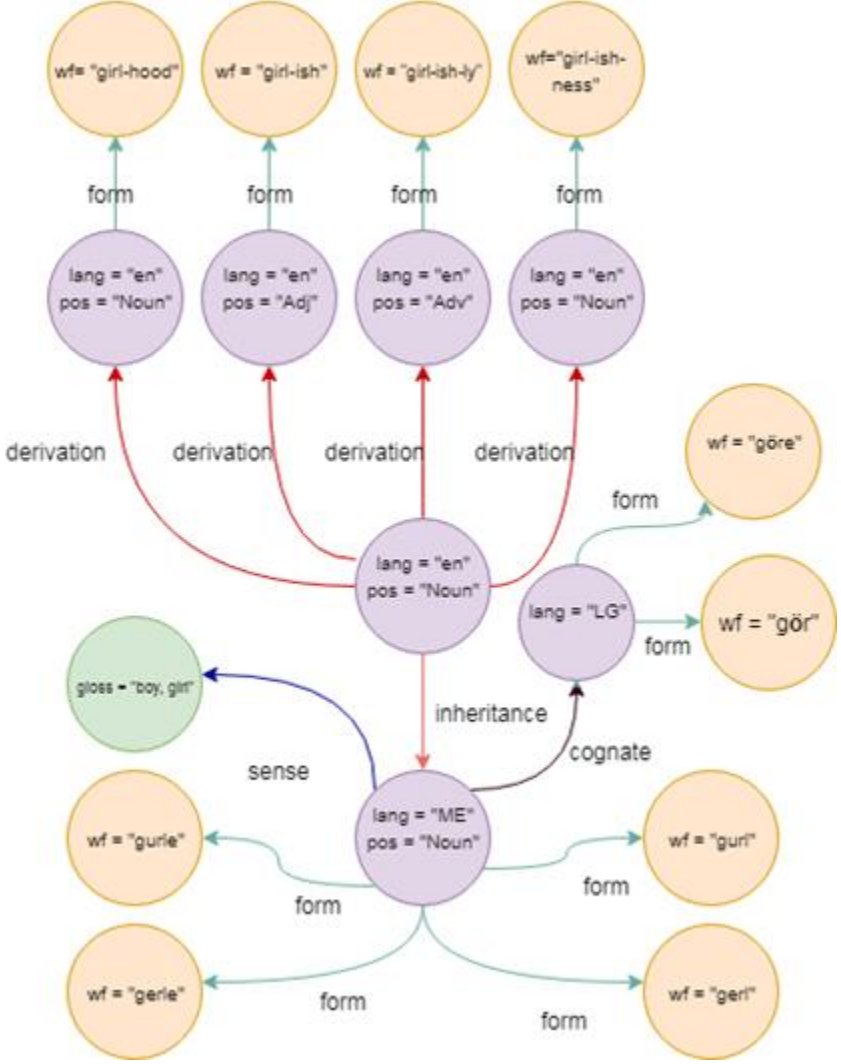
Description of the history and development of the word

An example etymological entry

girl



, whence *gerle*, *gurle*: o.o.o.: perh of C origin: cf Ga and Ir *caile*, EIr *cale*, a girl; with Anglo-Ir *girleen* (dim *-een*), a (young) girl, cf Ga-Ir *cailin* (dim -in), a girl. But far more prob, *girl* is of Gmc origin: Whitehall postulates the OE etymon **gyrela* or **gyrele* and adduces Southern E dial *girls*, primrose blossoms, and *grlopp*, a lout, and tentatively LG *goere*, a young person (either sex). Ult, perh, related to L *puer*, *puella*, with basic idea '(young) growing thing'.


wf= "girl-hood"
wf = "girl-ish"
wf = "girl-ish-ly"
wf="girl-ish-ness"
form
form
form
form
lang = "en"
pos = "Noun"
lang = "en"
pos = "Adj"
lang = "en"
pos = "Adv"
lang = "en"
pos = "Noun"
derivation
derivation
derivation
derivation
wf = "göre"
form
lang = "en"
pos = "Noun"
lang = "LG"
wf = "gör"
form
gloss = "boy, girl"
inheritance
cognate
sense
wf = "gurle"
lang = "ME"
pos = "Noun"
wf = "gurl"
form
form
wf = "gerle"
wf = "gerl"
form
form

Why Linked Data for Lexicographic Resources?

Linked data makes it easier to encode heterogeneous facts about lexicographic resources as well as facilitating querying and data integration using pre-existing tools and standards. On the other hand, TEI-XML has limitations in terms of linking resources with other datasets and describing semantics of descriptors.

Linked data and TEI-XML have different strengths, with linked data being more suitable for representing content and interaction between different views, while TEI-XML being more efficient in representing certain aspects of typographical and editorial views.

Dictionaries as Textual/Material Objects

OntoLex-Lemon + Lexicog however **still aren't** sufficient to represent all the different aspects we might be potentially interested in.

- Who **compiled** the dictionary, is it based on **previous works**?
- What about the **publishing history** of the text itself, its **different editions** (with different entries, definitions, etc), its **translations, manuscripts,** what about **individual copies in libraries**?
- What about the **texts/corpora** that are **cited as attestations**, citations to scholarly works?
- For some of these there already exist generic vocabularies (**Dublin Core**, **Prov-O**, **CITO**) which can provide solutions, others have to be adapted to the dictionary domain.

In fact we still need a conceptual framework for integrating together different levels of description. FRBR will provide this…and this will eventually bring us back to CIDOC-CRM

FRBR

- Stands for **Functional Requirements for Bibliographic Records**: an entity relationship model intended for the classification of intellectual products in **bibliographic databases** and **library catalogues**.
- It introduced an important distinction in terms of how we can describe intellectual products. We can refer to such products at four different levels of description. Namely, at the level of **Work**, **Expression**, **Manifestation**, and **Item**.
- We use the version of this distinction given in the **CIDOC-CRM aligned LRM** ontology.

Work and Expression

- **Work**: *“[C]omprises distinct intellectual ideas conveyed in artistic and intellectual creations such a poems stories or musical compositions. A work is the outcome of an intellectual process of one or more expressions.”*
  - Note that in the case of dictionaries this would encompass the **TEI lexical view**.

- **Expression**: *“[C]omprises the intellectual or artistic realisations of works in the form of identifiable immaterial objects, such as texts, poems [...] or any combination of such forms. The substance of F2 Expression is signs.”*
  - In the case of dictionaries we claim that this description encompasses the **TEI editorial view**.

Manifestation and Item

- **Manifestation**, ”[C]omprises products rendering one or more Expressions. A Manifestation is defined by both the overall content and the form of its presentation. The substance of F3 Manifestation is not only signs, but also the manner in which they are presented to be consumed by users, including the kind of media adopted[...] An instance of F3 Manifestation typically incorporates one or more instances of F2 Expression representing a distinct logical content and all additional input by a publisher such as text layout and cover design”
  - In the case of dictionaries F3 Manifestation encompasses the **TEI typographic view**
- The **Item** class: “*[C]omprises physical objects*” such as specific physical copies of dictionaries kept at libraries or academic institutions.
  - This class is associated with the kind of metadata information that is usually contained within the **TEI header element**.

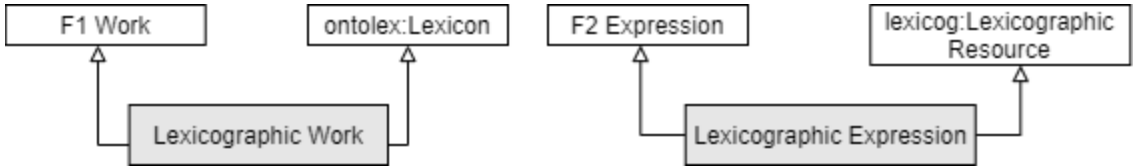
Bridging LRM and OntoLex

- We propose a number of new classes and properties to bridge together LRM (and CIDOC-CRM) and OntoLex-Lemon/Lexicog.
- **Lexicographic Work:** A subclass of the FRBRoo class **F1 Work** and the Ontolex-Lemon class **Lexicon**. It comprises concepts or combinations of concepts for representing/describing the lexicon for a given language community or communities or domain.
    - As **F1 Work** is a subclass of the CIDOC-CRM class **E89 Propositional Object** we can view individuals of **Lexicographic Work** as sets of **propositions about lexemes and related linguistic concepts belonging to a lexicon.**

Bridging LRM and OntoLex

- **Lexicographic Expression:** A subclass of the LRM class **F2 Expression** and the lexicog class **Lexicographic Resource:** The class comprises an intellectual realisation of the description of a lexicon as a structured text.
    - In other words it is a text viewed apart from **a specific typographic realisation**: a sequence of words that has an **additional organisation** in terms of entries, senses (defined as a sub-part of a lexicographical article that discusses a meaning of a lexical unit), forms, etc.

Bridging LRM and OntoLex



Asserting the Lexical View

- In our approach, we view a lexicographic entry as a series of statements making claims about different linguistic phenomena, about the lexicon of a language, as well a structural component of a text. In this we elaborate on previous work in both OntoLex and in CIDOC/FRBRoo.
- By modelling a dictionary as consisting of different levels of information, we can explicitly represent these as **hypotheses** (using named graphs or nanopublications).
- This comes in especially useful when it comes to combining together **etymologies**.

Modelling Citations and Annotations

By forcing us to **explicitly model our data** in terms of Subject-Predicate-Object triples RDF encourages us to think in terms of simple **declarative truth claims**: i.e., they make the preceeding considerations more salient. This is even more true wrt RDFS and OWL as these are much more expressive formal languages (OWL is a of description logic) and enable us/encourage us to make the meanings of our data much more '**explicit**'

The advantage of making this distinction is that it makes these different kinds of information more easily findable and queryable using the Semantic Web Query Language **SPARQL** for example.