



TEI and LEMON: a comparative study on the lexical encoding and interoperability

Jack Bowers, Thierry Declerck

► To cite this version:

Jack Bowers, Thierry Declerck. TEI and LEMON: a comparative study on the lexical encoding and interoperability. TEI conference and members' meeting 2016, Sep 2016, Vienna, Austria. hal-02921955

HAL Id: hal-02921955

<https://inria.hal.science/hal-02921955>

Submitted on 16 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TEI and LEMON: a comparative study on the lexical encoding and interoperability

Jack T. Bowers¹, Thierry Declerck²

1: Austrian Centre for Digital Humanities;

2: DFKI GmbH, Saarland University

Components Covered

2 Core

- 2.1 Lexical Entries
- 2.2 Forms
- 2.3 Semantics
- 2.4 Lexical Sense & Reference
- 2.5 Usage
- 2.6 Lexical Concept

3 Syntax and Semantics (synsem)

- 3.1 Syntactic Frames
- 3.2 Ontology Mappings
- 3.3 Complex ontology mappings / submappings
- 3.4 Conditions

4 Decomposition (decomp)

- 4.1 Subterms
- 4.2 Components
- 4.3 Phrase structure

5 Variation & Translation (vartrans)

- 5.1 *Lexico-Semantic Relations*
- 5.2 Translation
 - 5.2.1 Translation as shared reference
 - 5.2.2 Translation as a relation between lexical senses
 - 5.2.3 Translatable As

7 Linguistic Description

- 7.1 Morphosyntactic Description
- 7.2 Pragmatic & Paradigmatic Description
- 7.3 Arguments
- 7.4 Frames

8 Lexical Nets

- 8.1 Lexical nets in lemon

-Core: (list specifics)

-Decomposition

-VarTrans

-(SynSem)

-Linguistic Description

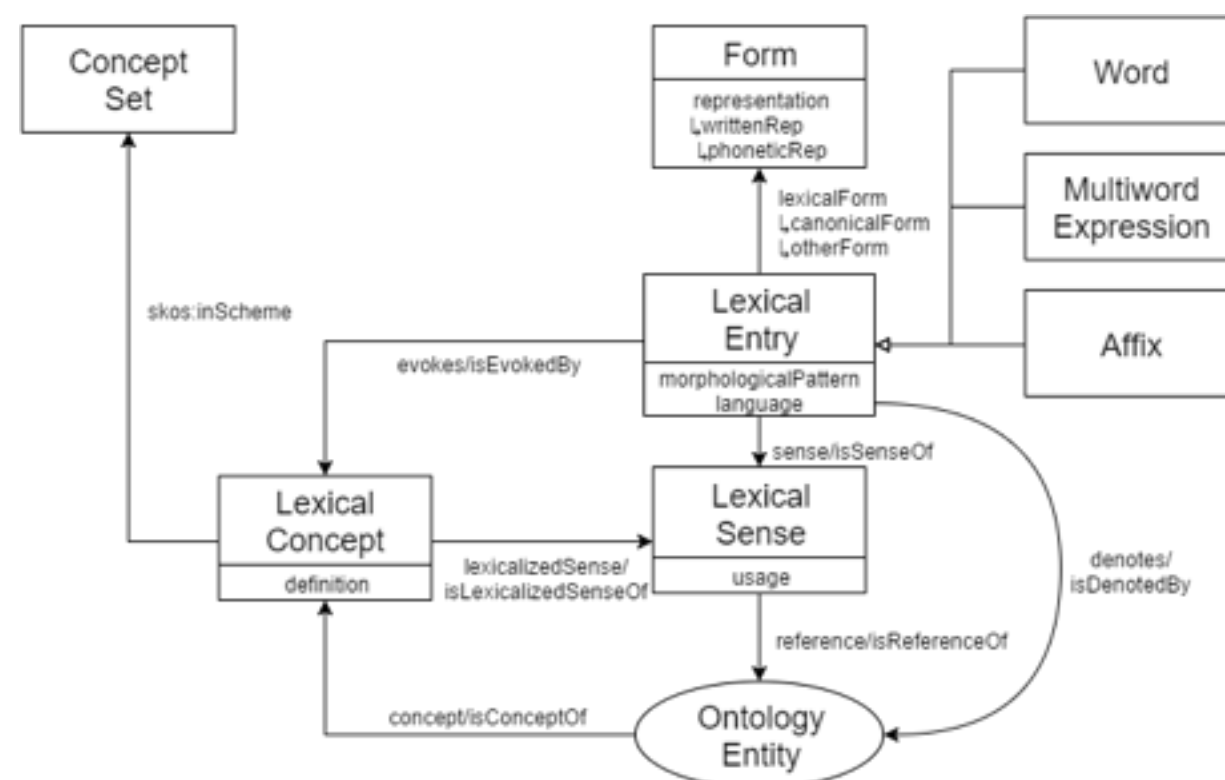
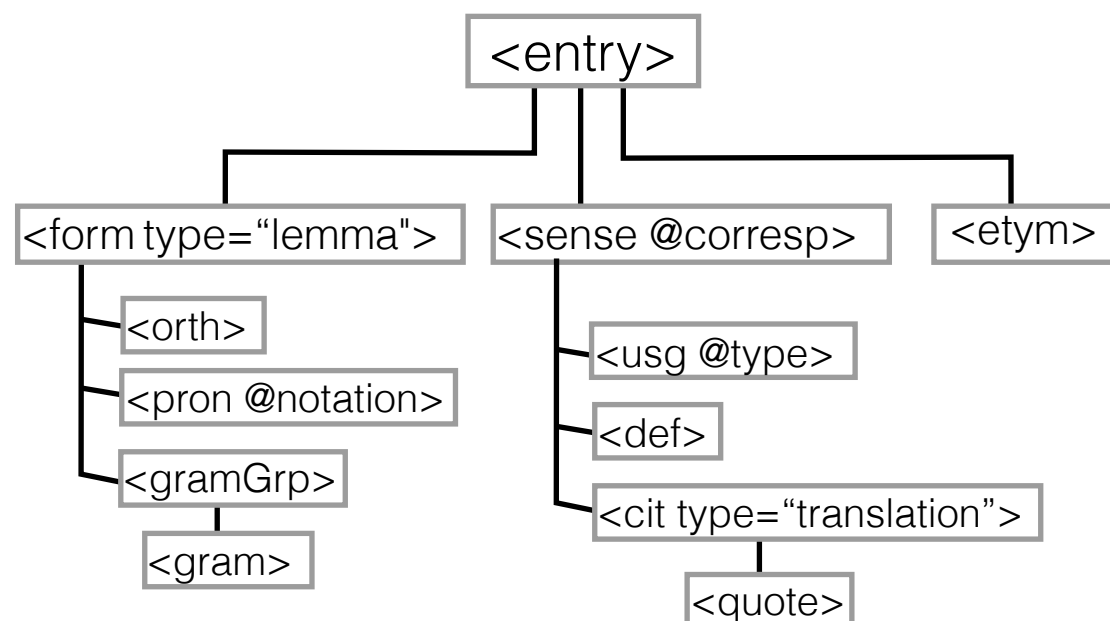
- Not covering LIME:

Fundamental Differences

- TEI is a hierarchical structure (reflecting XML basis)
- LEMON (Lexicon Model in Ontologies) is non-hierarchical, modular series of linked nodes/edges (*reflecting RDF basis*)
 - ONTOLEX (Core module)
 - Decomp
 - SynSem
 - VarTrans
 - Lime (Metadata)
- Whereas LEMON was originally made for LOD, TEI is adapting to it adhoc mostly within existing constructs

Fundamental Differences

LEMON/ONTOLEX Core Model



https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

Basic Parallels

Cannonical Form:

```
<entry xml:id="rabenschwarz">  
:Stachel_form  
rdf:type ontolex:Form ;  
ontolex:writtenRep "Stachel"^^rdf:langString ;
```

Lexical Entry type:

```
<form type="MultiWordExpression">  
rdf:type ontolex:MultiWordExpression ;
```

ISO language tag:

```
xml:lang="de"  
dcterms:language <http://id.loc.gov/vocabulary/iso639-2/de> ;
```

Definition:

```
<def>  
skos:definition
```

Phonetic & Orthographic Representations

Ontolex/Lemon in .ttl syntax

```
:lex_privacy a ontolex:LexicalEntry;  
ontolex:form :form_privacy.
```

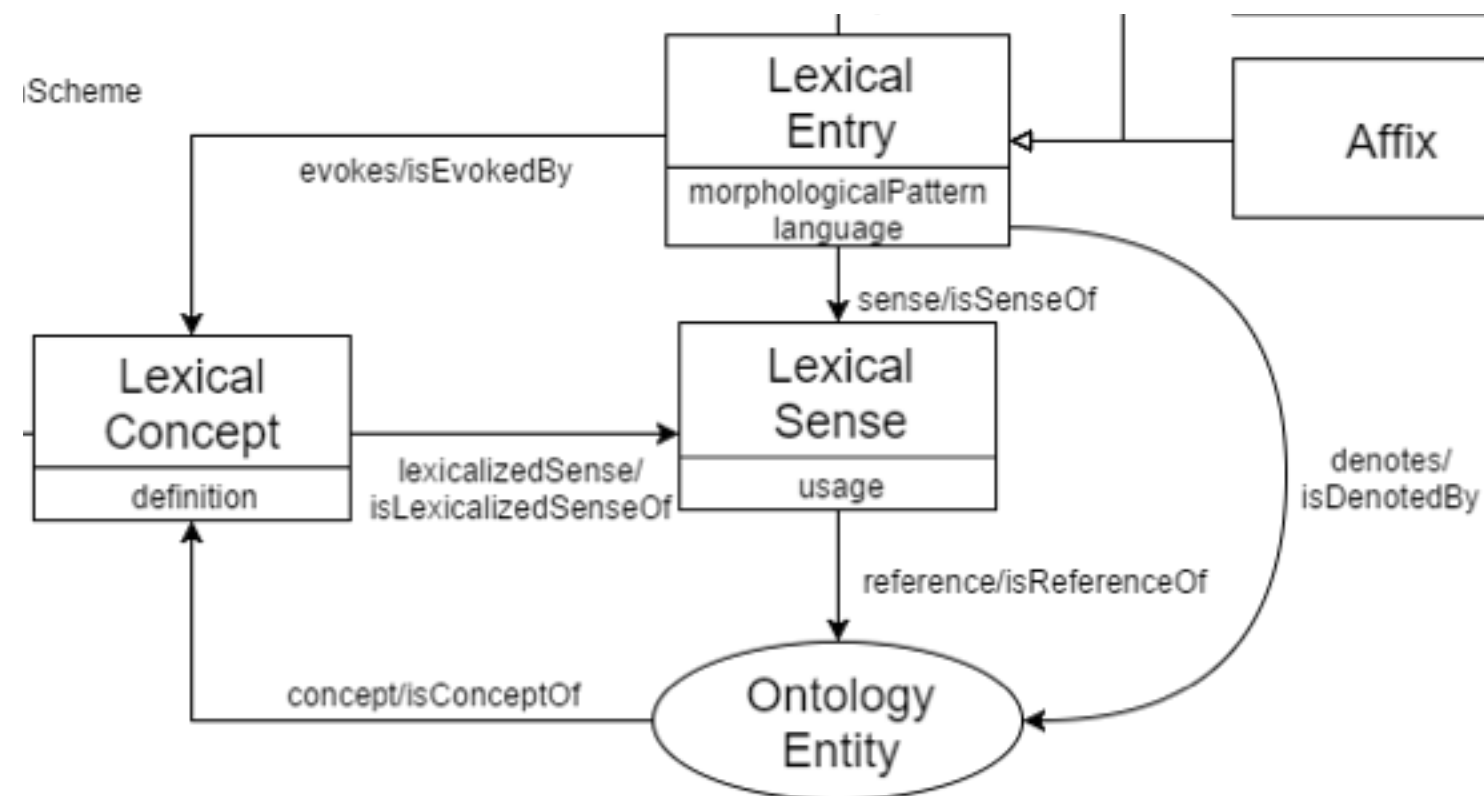
```
:form_privacy a ontolex:Form;  
ontolex:writtenRep "privacy"@en;  
ontolex:phoneticRep "pɹɪv.ə.si"@en-US-fonipa;  
ontolex:phoneticRep "pɹaɪ.və.si"@en-GB-fonipa.
```

```
<entry xml:id="privacy-id1">  
  <form type="lemma">  
    <pron xml:lang="en-US-fonipa">'pɹaɪ.və.si</pron>  
    <pron xml:lang="en-GB-fonipa">'pɹɪv.ə.si</pron>  
    <orth xml:lang="en">privacy</orth>  
  </form>  
</entry>
```

```
<!-- TEI option 2 -->  
<entry xml:id="privacy-id2">  
  <form type="lemma">  
    <pron xml:lang="en-US" notation="ipa">'pɹaɪ.və.si</pron>  
    <pron xml:lang="en-GB" notation="ipa">'pɹɪv.ə.si</pron>  
    <orth>privacy</orth>  
  </form>  
</entry>
```

Linking Sense to Ontologies: Lemon/Ontolex

One of the most useful aspects of the system is the dynamic ability to represent semantic information with regard to an ontology



Linking to Ontologies in TEI

Defining the lexical sense in terms of an external ontology can be done in several ways:

- `<sense @corresp>`, `<ref @target>`, `<usg type="dom" @corresp>`

e.g. `<sense corresp="http://dbpedia.org/resource/Porcupine">`
`<usg type="dom" corresp="http://dbpedia.org/resource/Animal">Tier</usg>`
`<ref type="sense" corresp="http://dbpedia.org/resource/Pig"/>`

- OR using feature structures as external URI concept repository, can be referenced using tag defined in `<prefixDef>`

e.g.

```

<encodingDesc>
  <listPrefixDef>
    <prefixDef ident="concept"
      matchPattern="([A-Z]+)"
      replacementPattern="concept-features-tei.xml#$1">
      ....
    </prefixDef>
  
```

←

```

<fs type="concept" xml:id="Schwein">
  <f name="term">
    <string xml:lang="de">Schwein</string>
  </f>
  <f name="conceptURL" corresp="http://dbpedia.org/resource/Pig"/>
</fs>

```

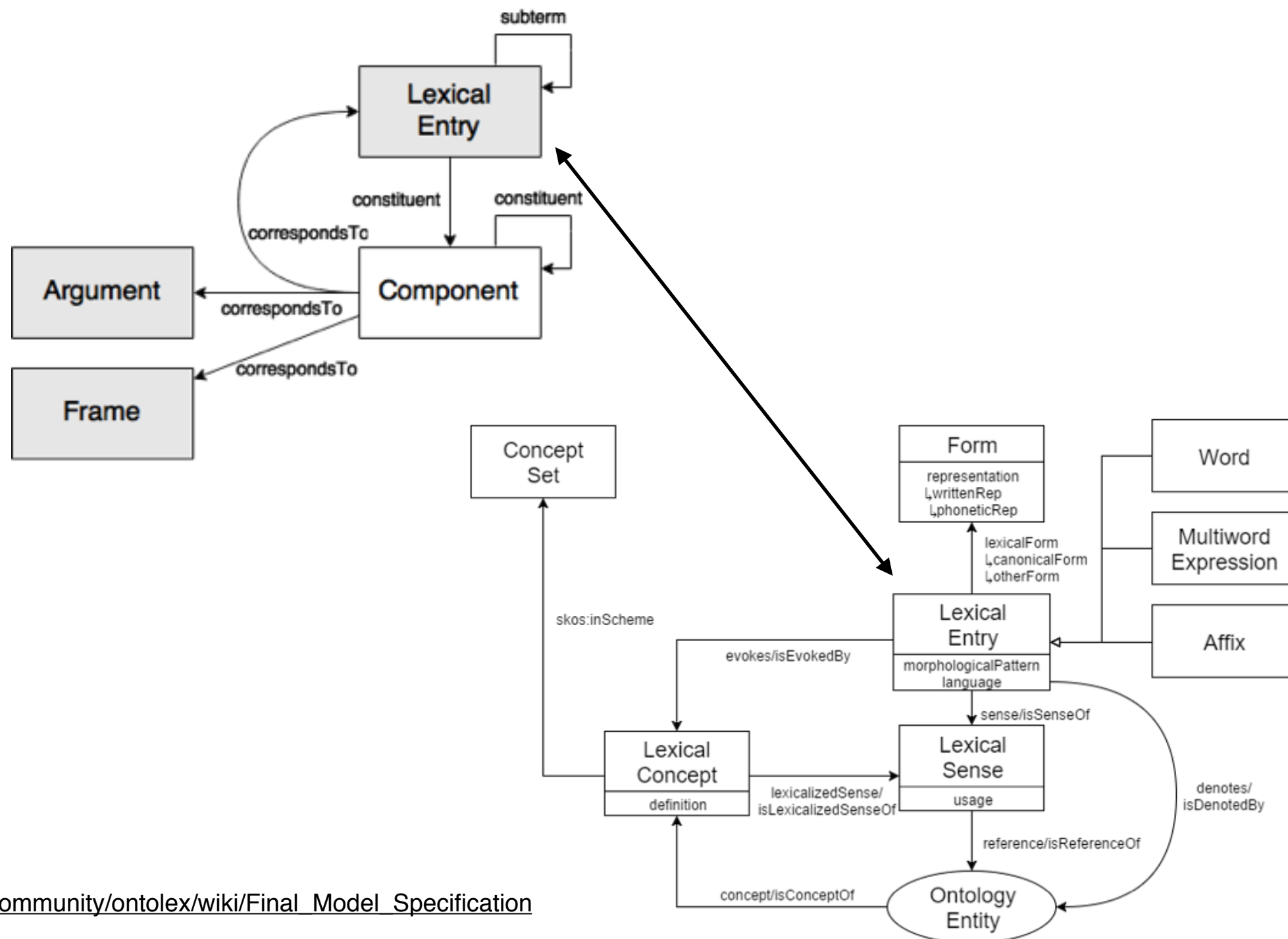
←

```

<sense corresp="concept:Schwein">

```

Decomposition: ONTOLEX & DECOMP



Decompositon of Compounds LEMON / DECOMP

Each component of compound identified and linked to individual lexical entry and sense

:Stachelschwein_lex

```
rdf:type ontolex:MultiWordExpression ;  
dcterm:language <http://id.loc.gov/vocabulary/iso639-2/de> ;  
lexinfo:gender lexinfo:neuter ;  
lexinfo:partOfSpeech lexinfo:noun ;  
rdf:_1 :Stachel_comp ;  
rdf:_2 :schwein_comp ;  
<http://www.w3.org/ns/lemon/decomp#constituent> :Stachel_comp ;  
<http://www.w3.org/ns/lemon/decomp#constituent> :schwein_comp ;  
<http://www.w3.org/ns/lemon/decomp#subterm> :Schwein_lex ;  
<http://www.w3.org/ns/lemon/decomp#subterm> :Stachel_lex ;  
ontolex:canonicalForm :Stachelschwein_form ;  
ontolex:sense :Porcupine_sense_1 ;
```

:Stachel_comp

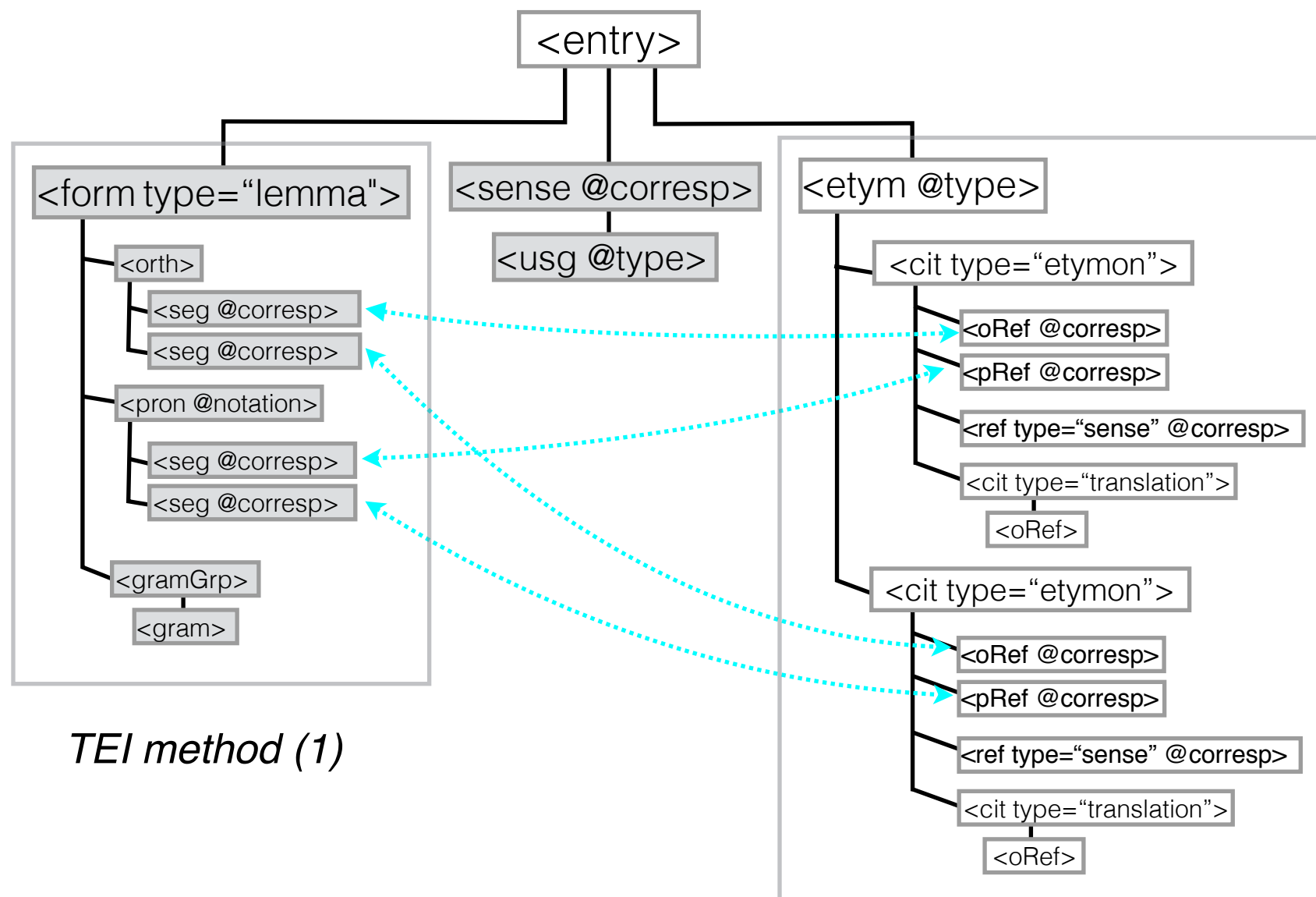
```
rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;  
lemon:sense :Spine_sense_1 ;  
<http://www.w3.org/ns/lemon/decomp#correspondsTo> :Stachel_lex ;
```

:schwein_comp

```
rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;  
lemon:sense :Pig_sense_1 ;  
<http://www.w3.org/ns/lemon/decomp#correspondsTo> :Schwein_lex ;
```

Decomposition of Compounds in TEI

**etymology structure according to proposals from Bowers & Romary (2016)*



TEI method (1)

TEI method (2)

`<p/oRef> @corresp` points to entry of etymon/compound component

`<ref type="sense"> @target` points to ontological entry corresponding to conceptual sense of etymon

Decomposition of Compounds in TEI

TEI: 2 ways

1) Segmentation of <orth> or <pron>;
e.g. using <seg>

```
<form type="lemma">
  <orth>
    <seg xml:id="d1e155" corresp="#Stachel">Stachel</seg>
    <seg xml:id="d1e157" corresp="#Schwein">schwein</seg>
  </orth>
  <gramGrp>
    <pos>noun</pos>
    <gen>neut</gen>
  </gramGrp>
</form>
```

./entry[@xml:id="Stachel"]

./entry[@xml:id="Schwein"]

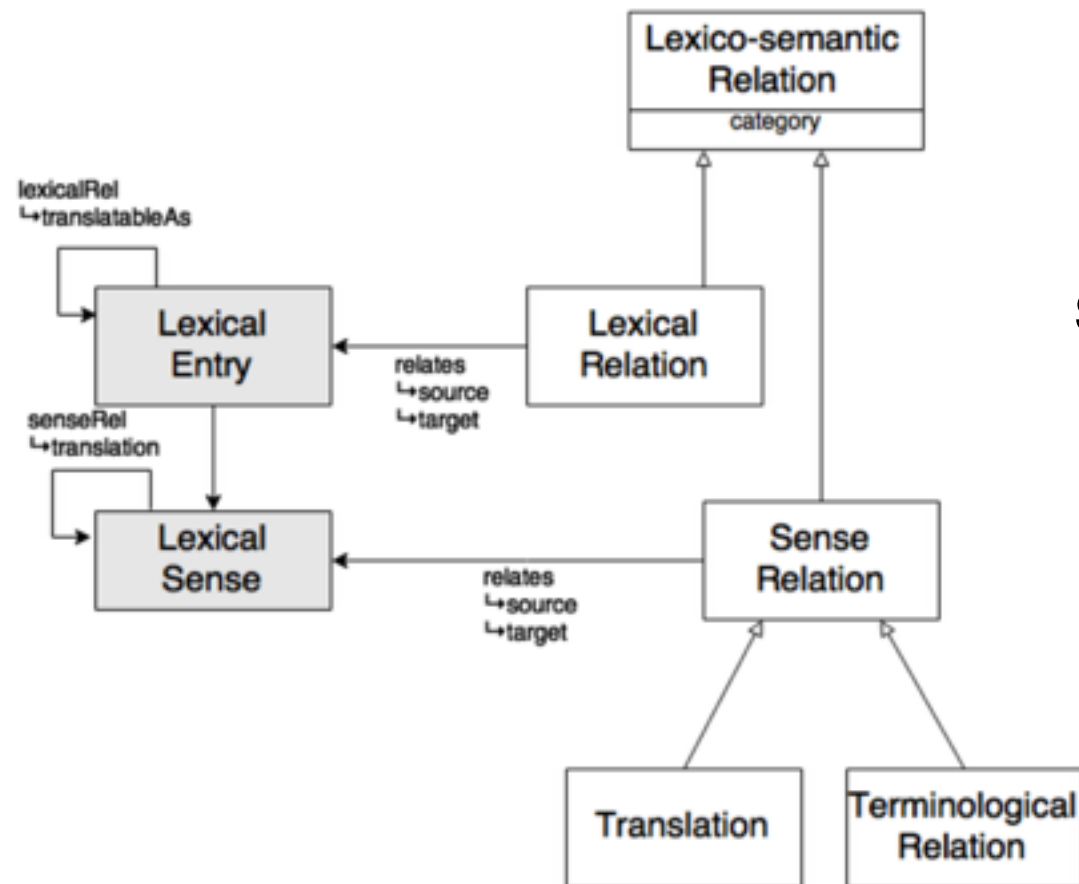
2) Within <etym>; *Bowers & Romary (2016)*

```
<etym type="compounding">
  <etym type="metonymy">
    <cit type="etymon" corresp="#d1e155">
      <oRef>Stachel</oRef>
      <cit type="translation" xml:lang="en">
        <oRef>spike</oRef>
      </cit>
      <ref type="sense" corresp="http://dbpedia.org/page/Spine_(zoology)"/>
    </cit>
    <note xml:lang="en">salient feature of their physical anatomy is their spikes</note>
  </etym>
  <etym type="metaphor">
    <cit type="etymon" corresp="#d1e157">
      <oRef corresp="#Schwein">Schwein</oRef>
      <cit type="translation" xml:lang="en">
        <oRef>pig</oRef>
      </cit>
      <ref type="sense" corresp="http://dbpedia.org/resource/Pig"/>
    </cit>
  </etym></etym>
```

Lemon: Variation & Translation (VarTrans)

lexicalRelation:

e.g. initialisms, derivational, morphosyntactic...



senseRelation:

TerminologicalRelation:

e.g. hypernymy and hyponymy relations, synonymy, antonymy,

```

:Porcupine_sense_1
rdf:type ontolex:LexicalSense ;
lexinfo:hypernym :Animal_sense_1 ;
ontolex:isSenseOf :Stachelschwein_lex ;
ontolex:reference <http://dbpedia.org/page/Porcupine> ;
  
```

https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

TranslationRelation:

e.g. translations

Variation & Translation (VarTrans)

How to express **sense relations** in TEI?

<usg @type="synonym" corresp="#entryid">

<usg @type="antonym" corresp="#entryid">

<usg @type="holonym" corresp="#entryid">

<usg @type="meronym" corresp="#entryid">

Summary / Comparisson

- TEI better for encoding and annotation
- Extended etymological classification
- Better for/closer to scholars

Advantages to ONTOLEX?

- ONTOLEX better for interlinking
- easier reuse of other existing vocabularies
- Better for/closer to programmers/data scientists