

Integrating WordNet and Wiktionary with lemon

John M^cCrae¹, Elena Montiel-Ponsoda² and
Philipp Cimiano¹

¹ Cognitive Interaction Technology Exzellenzcluster, Universität Bielefeld

² Ontology Engineering Group, Universidad Politécnica de Madrid



NUI Galway
OÉ Gaillimh



Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH



Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

- ▶ Much lexical data is in “data silos”
 - ▶ Proprietary formats
 - ▶ Restricted access
- ▶ The *Linking Open Data* project fosters:
 - ▶ Publication using RDF
 - ▶ Linking between resources
- ▶ We need **open** and **RDF-native** formats for language resources
 - ▶ **lemon** - **L**exicon **M**odel for **O**ntologies
 - ▶ Development under W3C OntoLex community group

Introduction

From Data Silos to Linked Data

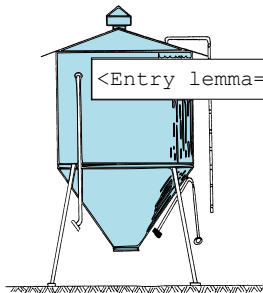
Lemon

WordNet to lemon

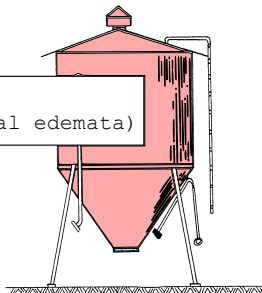
Wiktionary to lemon

Linking

Conclusion



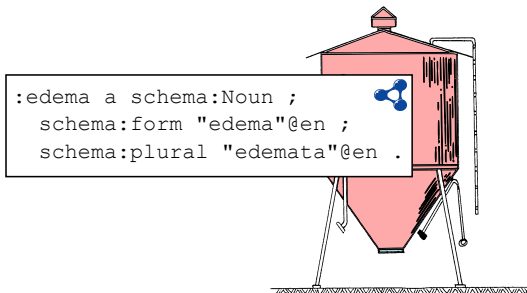
```
<Entry lemma="edema" pos="NP"/>
```



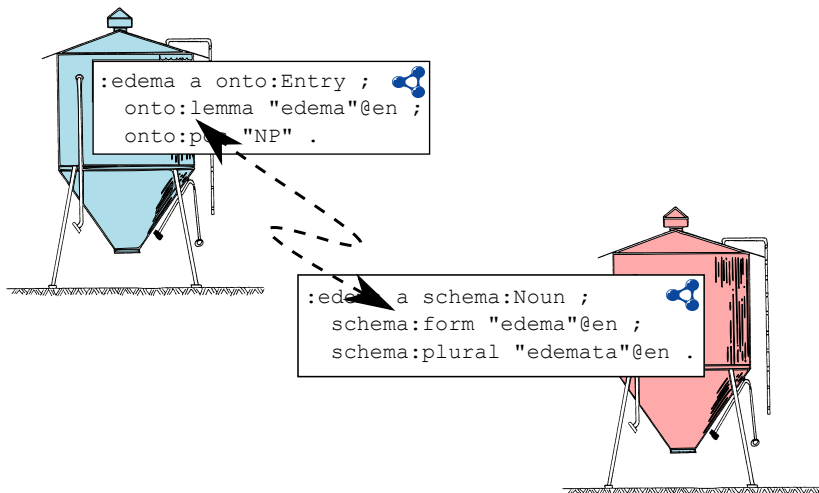
Noun:
edema (plural edemata)

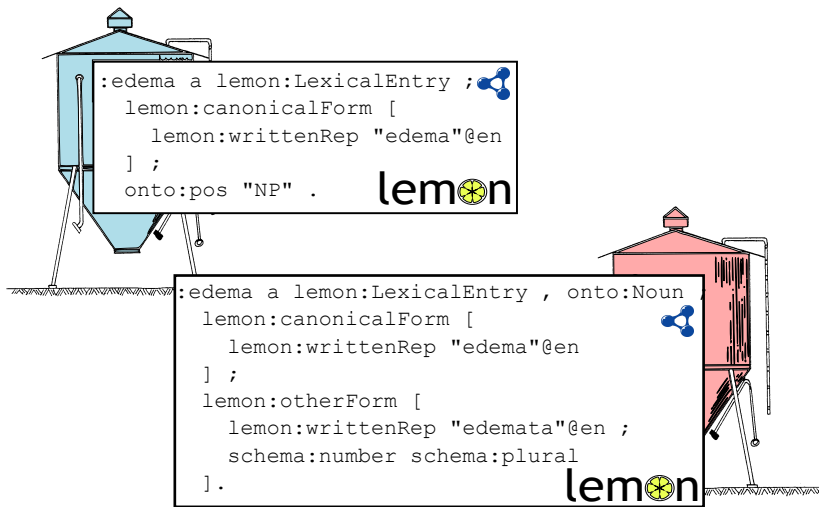


```
:edema a onto:Entry ;  
onto:lemma "edema"@en ;  
onto:pos "NP" .
```



```
:edema a schema:Noun ;  
schema:form "edema"@en ;  
schema:plural "edemata"@en .
```





lemon

OLIA



lemon

Introduction

From Data Silos to Linked Data

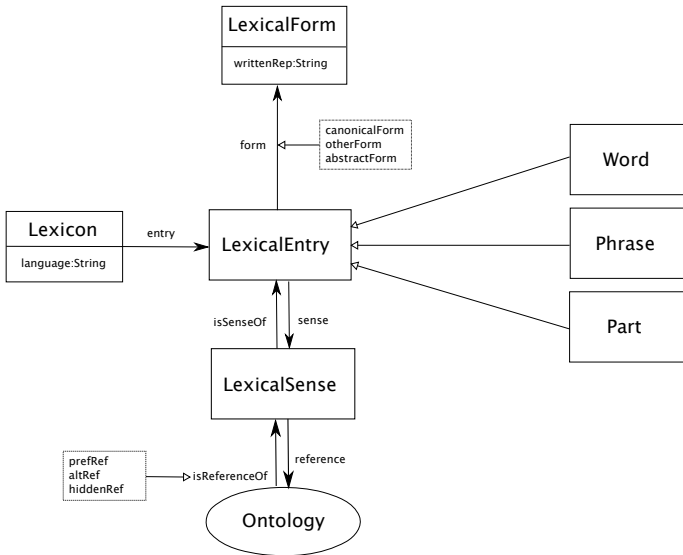
Lemon

WordNet to lemon

Wiktionary to lemon

Linking

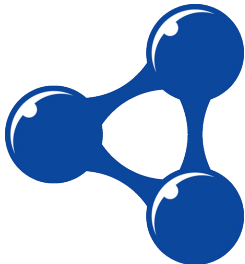
Conclusion



- ▶ Lexical Markup Framework (ISO 24613)
 - ▶ Standard for representing lexicons
 - ▶ XML, UML (primarily)
- ▶ LexInfo, LIR
 - ▶ Represent lexical information relative to an ontology
 - ▶ OWL
- ▶ SKOS (W3C Standard)
 - ▶ Designed for Taxonomy/Vocabulary representation
 - ▶ RDF

- ▶ RDF(S)
- ▶ Conciseness
- ▶ Not prescriptive
 - ▶ i.e., uses data categories
- ▶ Semantics by reference
 - ▶ i.e., uses ontologies
- ▶ Extensible

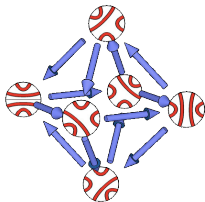
- ▶ RDF models are labelled directed graphs
 - ▶ Better representation
- ▶ Each entry has a URI
 - ▶ Queriable on the web using standards
 - ▶ Clear ownership of data
- ▶ Linking possible between different lexica
 - ▶ Reuse of lexicon data
- ▶ Some induction possible (subproperties, classes etc.)



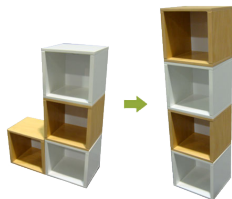
- ▶ Small models (i.e., fewer links, fewer kB)
- ▶ Easier to understand
- ▶ “Open-world”: Not necessary to state all facts
 - ▶ Multiple points of view



- ▶ The web of data is full of ontologies in OWL, RDFS, RIF...
- ▶ Meaning of a word given by reference
- ▶ Reference (generally an ontology) capable of representing more complex semantic information
- ▶ Disambiguation is performed relative to the ontology
- ▶ No (traditional) word senses
 - ▶ No clashing of word senses in cross-lingual mappings



- ▶ RDF(S) extensibility allows representation of
 - ▶ Subtle differences
 - ▶ Unexpected data categories
- ▶ Modularity
 - ▶ Different modules for different user requirements
 - ▶ New modules can be added later without affecting core



Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

- ▶ Start with RDF-WordNet 2.0
- ▶ Mapped synsets to references
 - ▶ Hence synsets are treated as ontology classes
- ▶ Sense and Word correspond to **lemon**
- ▶ Canonical form introduced as new node, other forms extracted from WordNet files (not in RDF!)
- ▶ Part-of-Speech tags mapped to LexInfo

```
lwn:marmoset-noun-entry rdf:type lemon:LexicalEntry ;  
  lexinfo:partOfSpeech lexinfo:noun ;  
  lemon:sense lwn:sense-marmoset-noun-1 ;  
  lemon:canonicalForm lwn:word-marmoset-canonicalForm .
```

```
lwn:sense-marmoset-noun-1  
  lemon:reference wn20:synset-marmoset-noun-1 .
```

```
lwn:word-marmoset-canonicalForm  
  lemon:writtenRep "Marmoset"@en .
```

Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

Wiktionary [ˈwɪkʃənəri] n., a wiki based Open Content dictionary

Lexicon (Lexical) Entry

Contents [hide]

- English
 - 1.1 Etymology
 - 1.2 Pronunciation
 - 1.3 Verb
 - 1.3.1 Usage notes
 - 1.3.2 Derived terms
 - 1.3.3 Related terms
 - 1.3.4 Translations
 - 1.4 Noun
 - 1.4.1 Derived terms
 - 1.5 References
 - 1.6 Anagrams
- Conish
 - 2.1 Etymology
 - 2.2 Pronunciation
 - 2.3 Noun

English

Most common English words: before • see • over • #93 know • much • after • first

Etymology

From Middle English *knownen* from Old English *cnawan* from Proto-Germanic **kneṡan* ("to know") from Proto-Indo-European **ǵnē-*, **ǵnō-* ("to know").

cognates

Pronunciation

- (UK) IPA: /nəʊ/, SAHPA: /nɔʊ/
- (US) IPA: /noʊ/, SAHPA: /noʊ/
- Audio** (NS)
- Audio** (UK)
- Rhymes**: -oʊ
- Homophones**: paw, now (in some dialects or accents, but not in standard English)

Part of speech

Word Forms

Verb

to know (third person singular simple present **knows**, present participle **knowing**, simple past **knew** or **knawed** (dialect), past participle **known**, **knownen** (archaic), or **knawed** (dialect))

Subcategorization

- (transitive) To be certain or sure about.
 - I know that I'm right and you're wrong.*
 - He **knows** something terrible was going to happen.*
- (transitive) To be acquainted or familiar with; to have encountered.
 - I **know** your mother, but I've never met your father.*
- (transitive, also intransitive followed by **about** or, dialectally, **from**) To have knowledge of; to have memorised information, data, or facts about.

(Lexical) Senses

Done

Wikimedia has an article on: **Know**

Reference

The image shows a screenshot of the Wiktionary entry for the word "know". The browser tab is titled "W know - Wiktionary". The page content includes the Wiktionary logo, the word "know" with its IPA transcription [ˈwɪkʃənəri], and a list of contents. Annotations are present: an orange box labeled "Lexicon" with an arrow pointing to the word "know", and another orange box labeled "(Lexical) Entry" with an arrow pointing to the word "know".

W know - Wiktionary

a multilingual free encyclopedia

Wiktionary
[ˈwɪkʃənəri] *n.*,
a wiki-based Open
Content dictionary
Wileo [ˈwɪl kəri]

[Main Page](#)
[Community portal](#)
[Preferences](#)
[Requested entries](#)



[Entry](#) [Discussion](#) [Citations](#)

know

Contents [hide]

- 1 English
 - 1.1 Etymology
 - 1.2 Pronunciation
 - 1.3 Verb

Español
 Euskara
 العربية
 Français
 Galego
 한국어
 Hrvatski
 Ido
 Italiano
 ភាសាខ្មែរ
 Қазақша
 Лезги
 Lietuvių
 Limburgs
 Magyar

- (US) IPA: /nɒʊ/, SAMPA: /nɒU/
- Audio (US) 
- Audio (UK) 
- Rhymes: -ɒʊ
- Homophones: *now, nob; now* (in some dialects or accents, but not in standard English)

Part of speech

Verb

Word Forms

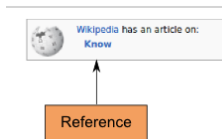
(Lexical) Senses

Subcategorization

to know (third-person singular simple present **knows**, present participle **knowing**, simple past **knew** or **knower** (dialect), past participle **known**, **knownen** (archaic), or **knower** (dialect))

1. (transitive) To be certain or sure about.
*I **know** that I'm right and you're wrong.*
*He **knew** something terrible was going to happen.*
2. (transitive) To be acquainted or familiar with; to have encountered.
*I **know** your mother, but I've never met your father.*
3. (transitive, also intransitive followed by **about** or, dialectally, **from**) To have knowledge of; to have memorised information, data, or facts about.
*He **knows** more about 19th-century politics than any other student.*

Done



Wiktionary:

```
<page>
<title>free</title>
<text>
==English==
===Adjective===
{{en-adj}}

# Not [[imprisoned]] or [[enslaved]].
# Obtainable without any [[payment]].

====Synonyms====
* {{sense|obtainable without payment}}:
  [[free of charge]], [[gratis]]

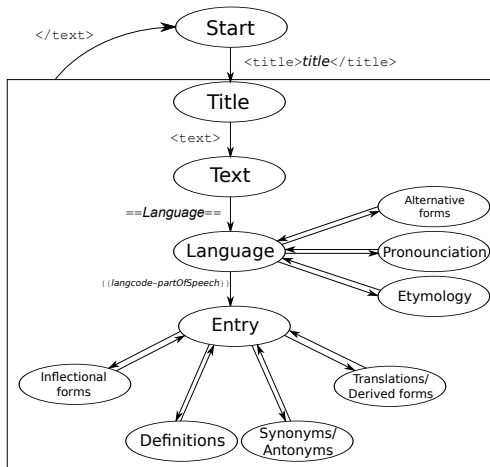
====Translations====
{{trans-top|not imprisoned}}
* German: {{t+|de|frei}}
{{trans-bot}}
</text>
</page>
```

lemon:

```
:free_en_adj lemon:canonicalForm [
  lemon:writtenRep "free"@en ] ;
lexinfo:partOfSpeech lexinfo:adjective ;
lemon:sense :free_en_adj_sense0 ;
lemon:sense :free_en_adj_sense1 ;
lemon:sense :free_en_sense_def .

:free_en_adj_sense0 lemon:definition [
  lemon:value "Not imprisoned or enslaved"@en ] ;
lemon:reference
  <http://en.wiktionary.org/wiki/free> ;
lexinfo:translation :frei_de_sense_def .

:free_en_adj_sense1 lemon:definition [
  lemon:value "Obtainable without any payment"@en ] ;
lemon:reference
  <http://en.wiktionary.org/wiki/free> ;
lexinfo:synonym :free_of_charge_en_sense_def .
```



- ▶ (English) Wiktionary uses different glosses to link pages
 - ▶ “Not imprisoned or enslaved” vs. “Not imprisoned”
 - ▶ “Obtainable without any payment” vs. “Obtainable without payment”
- ▶ We merge information on the same Wiktionary page
 - IF The secondary gloss is a substring of the primary gloss
 - OR The Levenshtein distance between the glosses exceeds some λ
 - AND The Levenshtein distance is maximal among candidates

λ	Merged	Coverage	Precision	Harmonic Mean
Substring	36595	37.8%	99.5%	54.8%
0.9	6842	44.9%	100%	62.0%
0.8	3398	48.4%	99%	65.0%
0.7	2669	51.2%	99%	67.5%
0.6	3243	54.5%	97%	69.8%
0.5	7128	61.9%	97%	75.6%
0.4	4612	66.6%	98%	79.3%
0.3	6295	73.1%	91%	81.1%
0.2	7983	81.4%	92%	86.4%
0.1	6934	88.5%	73%	80.0%
0.0	3862	92.5%	71%	80.3%

Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

- ▶ We used the following criteria:
 - ▶ The canonical (lemma) form is equivalent
 - ▶ Part-of-speech is the same
 - ▶ Do not assert different values for the same property
 - ▶ Do not have a different non-canonical form with the same properties
 - ▶ e.g., German: “Banken” versus “Bänke”
- ▶ Results:

	#Entries	Percent (WN)	Percent (Wikt)
Linked	63,478	21.0%	26.9%
Not Linked (Wiktionary)	172,674	-	73.1%
Not Linked (WordNet)	238,408	79.0%	-
Ambiguous	1,741	0.6%	0.7%

(in Wiktionary not in WordNet)

- ▶ 28: In WordNet
- ▶ 9 (“polysemic”, “abaciscus” (pictured)): Omissions
- ▶ 10 (“false friend”, “apples and pears”): Idioms not covered by WordNet
- ▶ 2 (“raven” (adj), “to minute” (verb)): Not with same part-of-speech
- ▶ 1 (“wares”): Other



Introduction

From Data Silos to Linked Data

Lemon

WordNet to lemon

Wiktionary to lemon

Linking

Conclusion

- ▶ Conversion of WordNet easy due to model interoperability (... even stage 1 helps!)
- ▶ Wiktionary much harder
- ▶ **lemon** is an adequate model for representing Wiktionary and WordNet
- ▶ Wiktionary's data model is flawed!
- ▶ Overlap between WordNet and Wiktionary quite low (~25%)
- ▶ Linking these resources can create a “virtual” resource with much higher coverage

- ▶ <http://monnetproject.deri.ie/lemonsource>: Data sets from the presentation
- ▶ <http://www.lexinfo.net/lemon-cookbook.pdf>: The **lemon** cookbook (technical manual)
- ▶ <http://www.w3.org/community/ontolex>: OntoLex Community group
- ▶ <http://www.monnet-project.eu/lemon>: **lemon** Ontology