Universität Bielefeld

Modelling the ontology-lexicon interface with lemon

John McCrae

Semantic Computing Group, University of Bielefeld

CLT Seminar at University Gotheburg, 28th February 2013

CITEC

Universität Bielefeld

Outline

CITEC

Universität Bielefeld

Outline

CITEC

About me

- From S.E. England (Muswell Hill, Tonbridge, Fulham)
- Studied Mathematics and Computer Science at Imperial College London
- PhD work at the National Institute of Informatics, Tokyo, Japan
  - "Automatic extraction of logically consistent ontologies from text corpora"
  - Supervisor: Nigel Collier
- Joined the Semantic Computing Group in October 2009
- Worked on FP7 Project "Monnet"

Bielefeld

- 20th largest city in Germany (323,076 inhabitants)
- University founded in 1969
- Approx 20,000 undergraduates
- Famous for "Bielefeld Conspiracy"
  - "Bielefeld does not actually exist. Rather, its existence is merely propagated by an entity known only as THEM"
- Fun fact: Bielefeld produces more pizzas daily than any other city in Europe

Semantic Computing

- Semantic Computing group headed by Prof. Dr. Philipp Cimiano
- Part of the Cognitive Interaction Technology Excellence Cluster (CITEC)
  - Focused on Cognition, Artifical Intelligence and Robotics
- Topics covered in the group:
  - Linked Data
  - Machine Learning
  - Question Answering
  - Social Media
  - Language Resources
  - Machine Translation

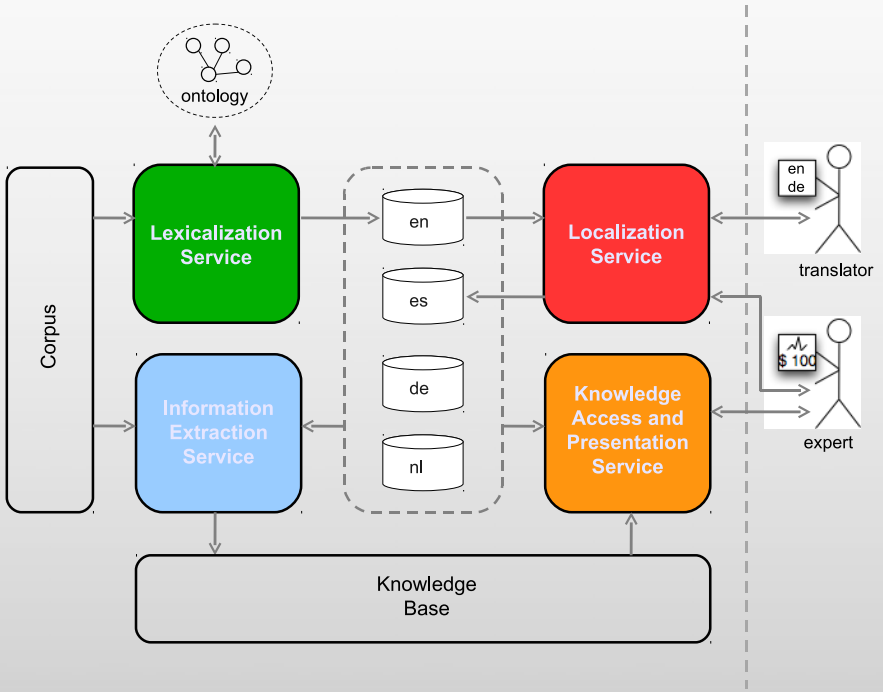
Monnet

- 3-year FP7 project on "Multilingual Ontologies for Networked Knowledge"
- Partners:
  - DERI, National University of Galway (Lead)
  - Polytechnic University of Madrid (UPM)
  - German Research Center for Artifical Intelligence (DFKI)
  - University Bielefeld
  - SAP AG
  - BeInformed (The Netherlands)
  - XBRL Europe
- Started in March 2010, ends today!

CITEC

Monnet Overview

- Lexicalization service takes existing ontologies and adds lexical information
- Localization service translates these resources to other languages
- Lexica used to extract information from text and extended/populate ontologies
- Presentation framework to enable automatic localization



Universität Bielefeld

Outline

CITEC

Outline

Problem

- Ontologies have become popular
- Use several formalisms: RDFS, OWL, F-Logic, etc.
- Ontologies do not have much linguistic information

```
:Cat a owl:Class ;
   rdfs:label "cat"@eng ;
   rdfs:label "katt"@swe .
```

- What is the plural? Easy for English, not for Swedish

Ontologies

Take a word:

"edema"

And it means something, so we put it in an ontology and give it an identifier (URI):

```
http://www.dbpedia.org/resource/Edema
```

In fact it (already) has lots of identifiers linked on the web

```
mesh:D004487

                                        icd10:R60.9

            umls:C0013604
```

Ontologies

- We can describe the entity with axioms
- Relationships to entities in other ontologies
- Use reasoning to infer equivalence
- All done with the “Web Ontology Language” (OWL)
  - Published by W3C in 2002; version 2 in 2008

Ontology labels

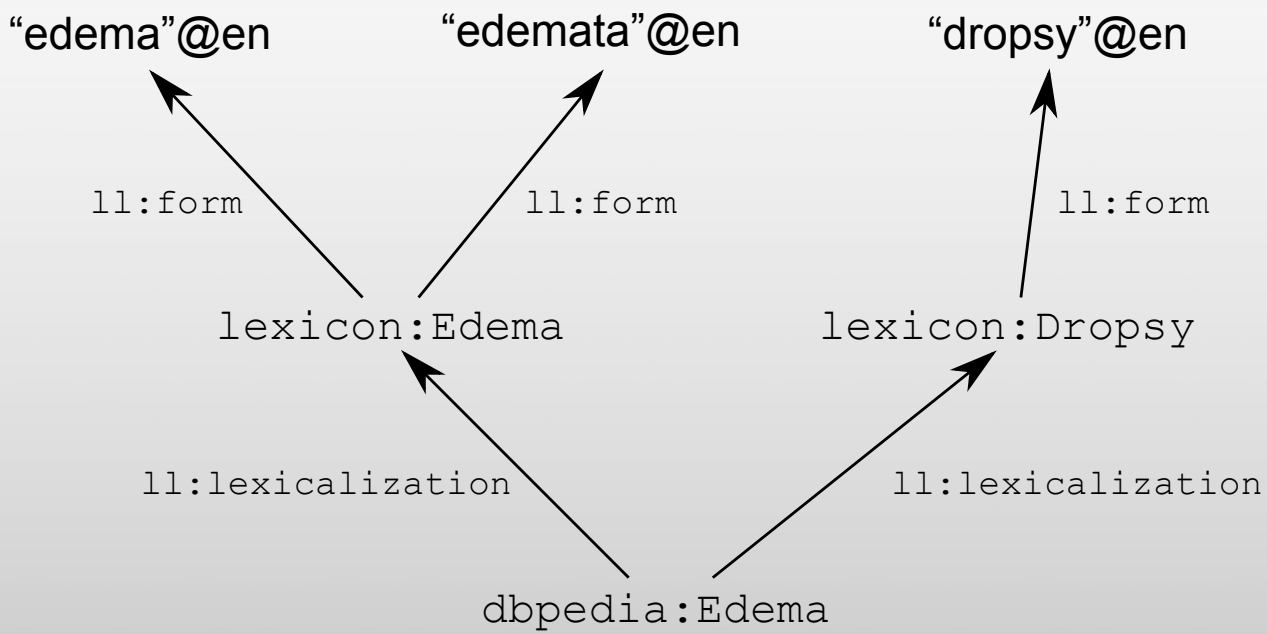
Concepts may be identified by many words

"edema" "edemata" "dropsy"

- These are all labels for the same ontology concept
- No differentiation
- Cannot say which are plural, which not

Inflection and Synonyms

We could introduce an element for each word:



SKOS-XL

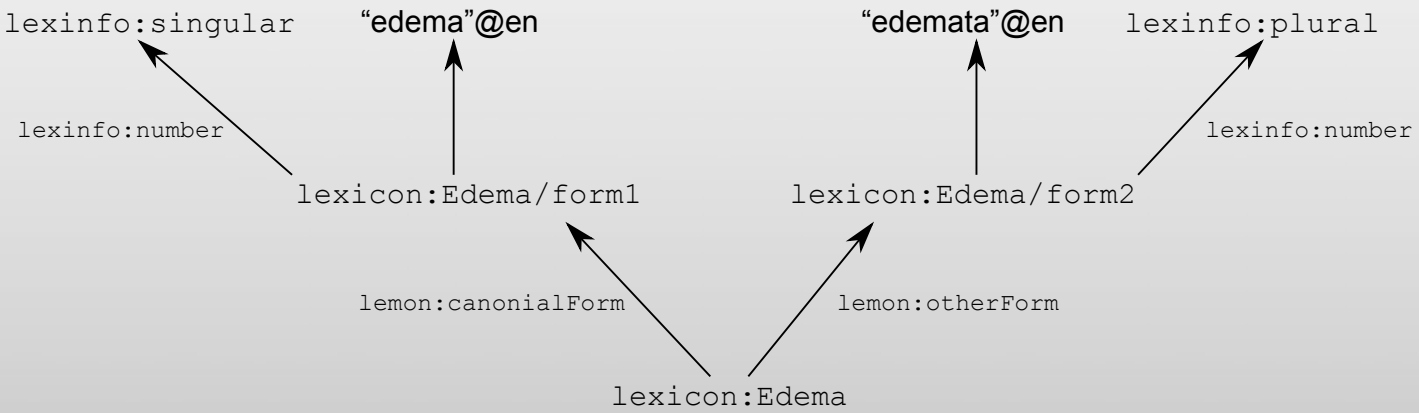
- Similar to proposed model SKOS-XL
  - eXtended Labels for the Simple Knowledge Organization System
  - W3C Recommendation since 2009
- SKOS-XL does not allow multiple forms of the same label
  - No grouping of "edema" and "edemata"
- "We [TopQuadrant] have yet to hear a use case that cannot be supported by SKOS alone" (Polikoff, 2013)

Forms

But such a distinction is only useful if we can say why:

"edema" (singular) "edemata" (plural)

Hence, forms are also nodes:

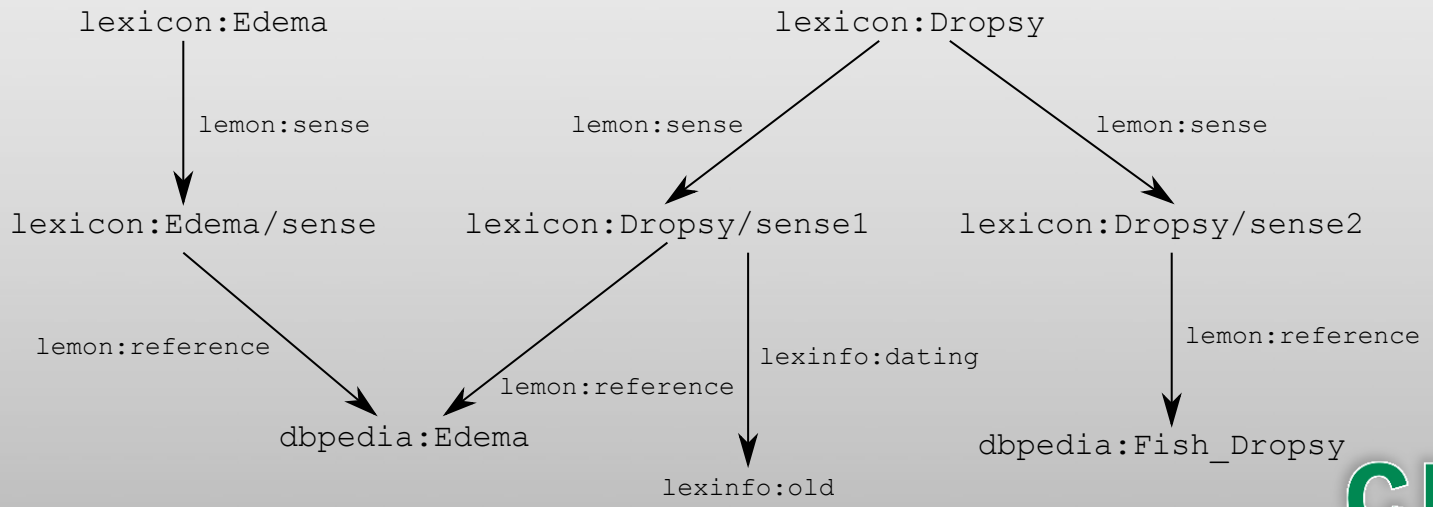


Senses

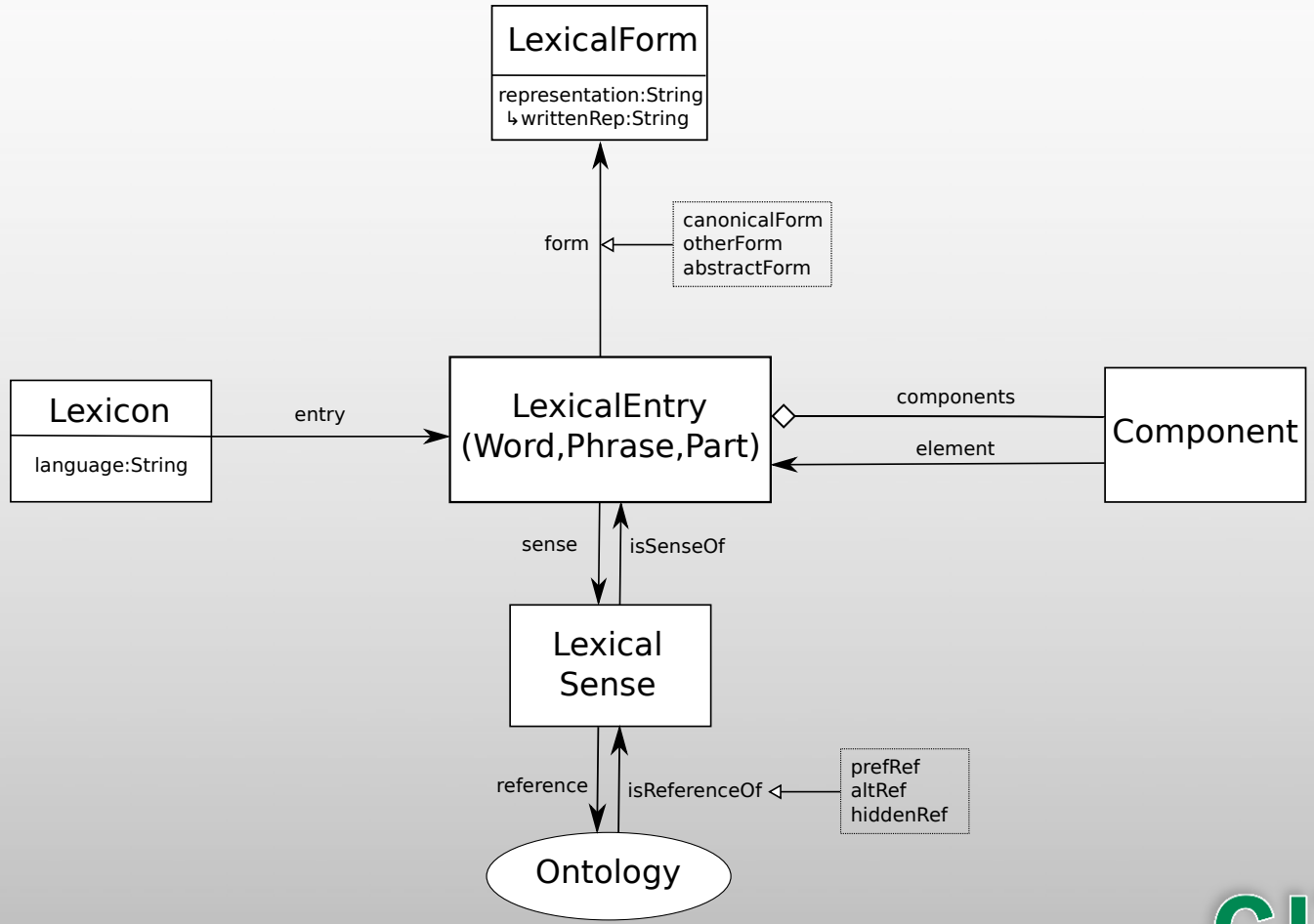
Sometimes we wish to say something about why a particular word is used

"edema" (modern) "dropsy" (antiquated)

Hence we introduce a sense to describe the usage of a word with a given meaning



The core of lemon



Universität Bielefeld

Outline

CITEC

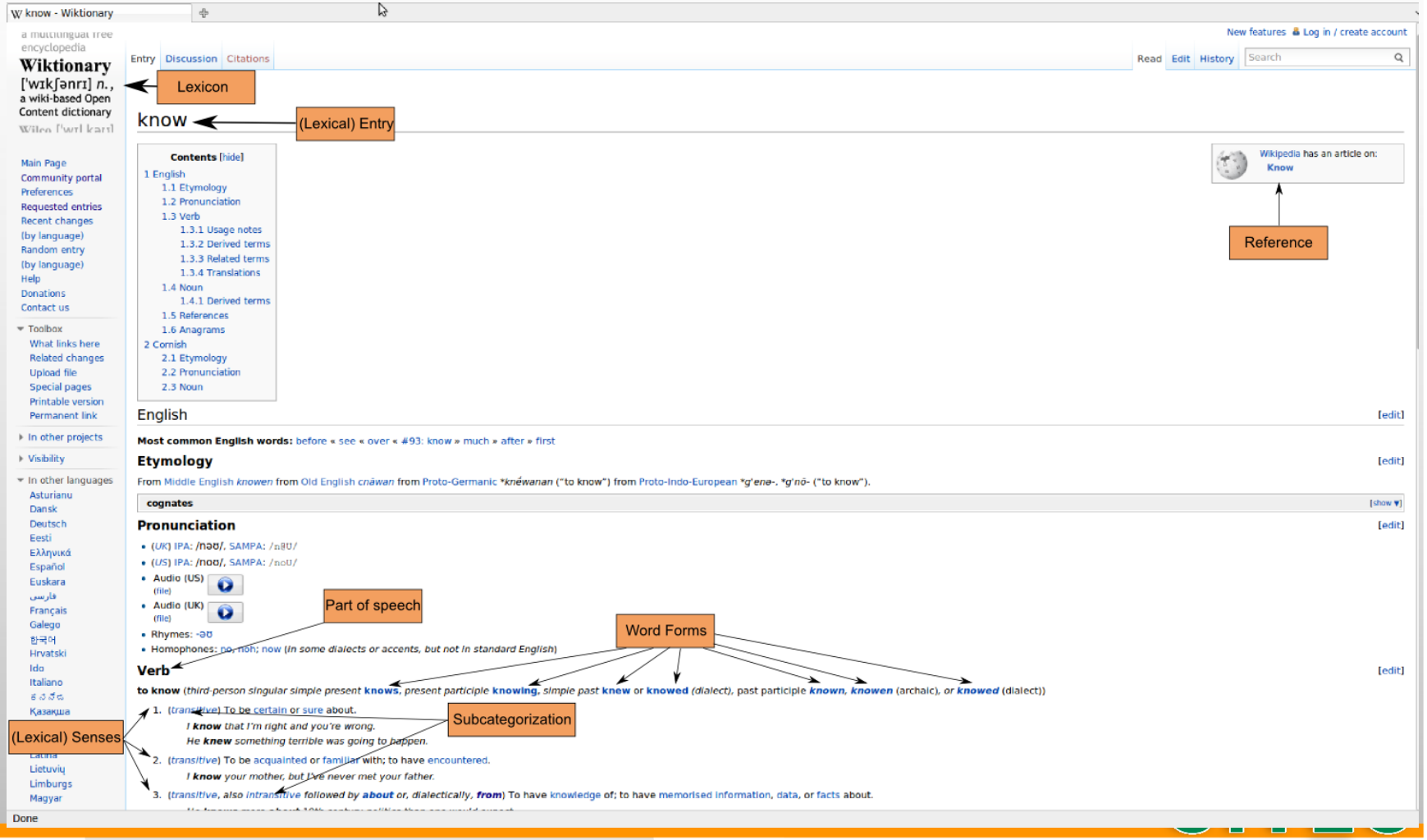
So..., what is a lexicon?

- A lexicon is a collection of lexical information
- We do not need to define semantics within the lexicon
- "An ontology-based semantic lexicon would leave the semantics to the ontology, focusing instead on providing domain-specific terms and object descriptions in the ontology." (Buitelaar, 2010)

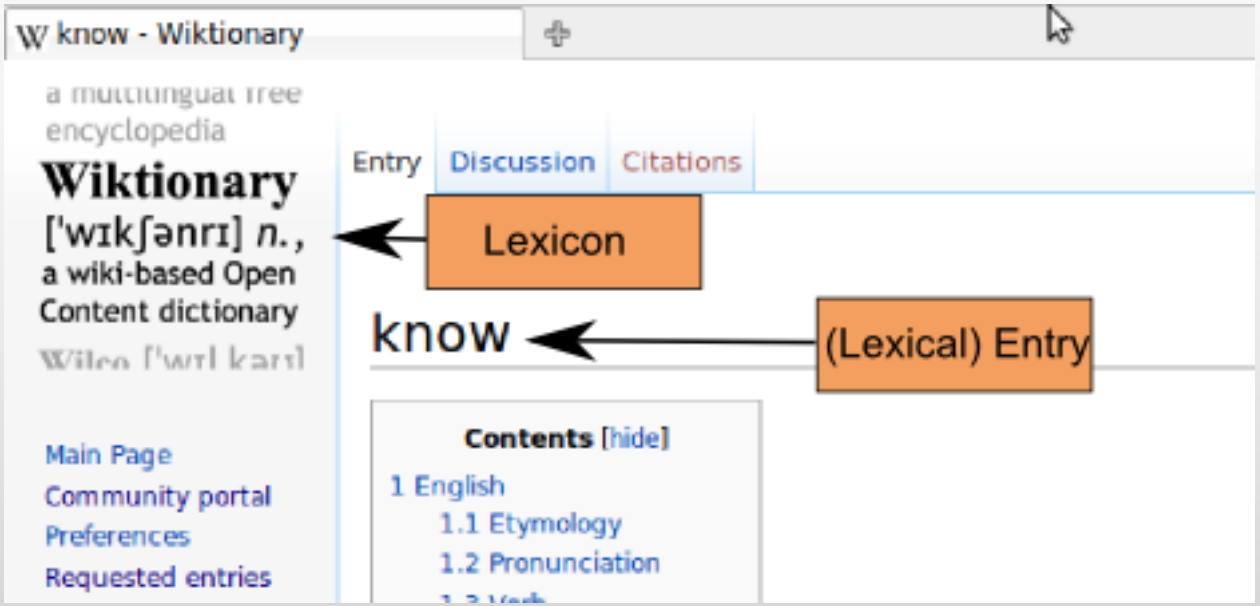
Dictionaries as lexica

- In fact, a lexicon represents much of the information already found in a dictionary
- That is words, their forms and their meaning
- Must be machine-readable
- Take Wiktionary as an example

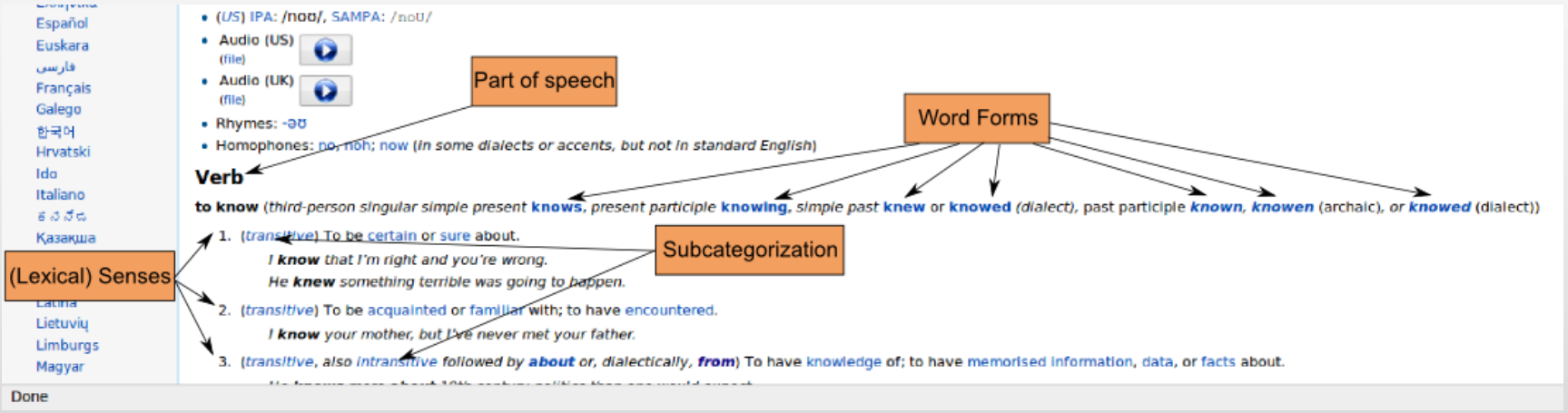
Wiktionary as a lexicon



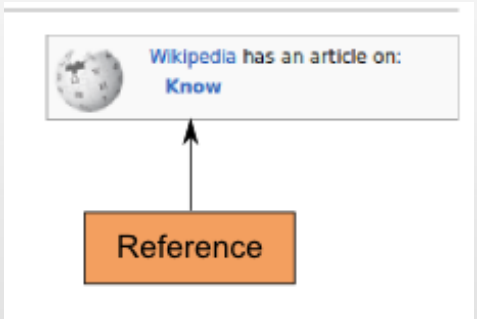
Wiktionary as a lexicon



Wiktionary as a lexicon

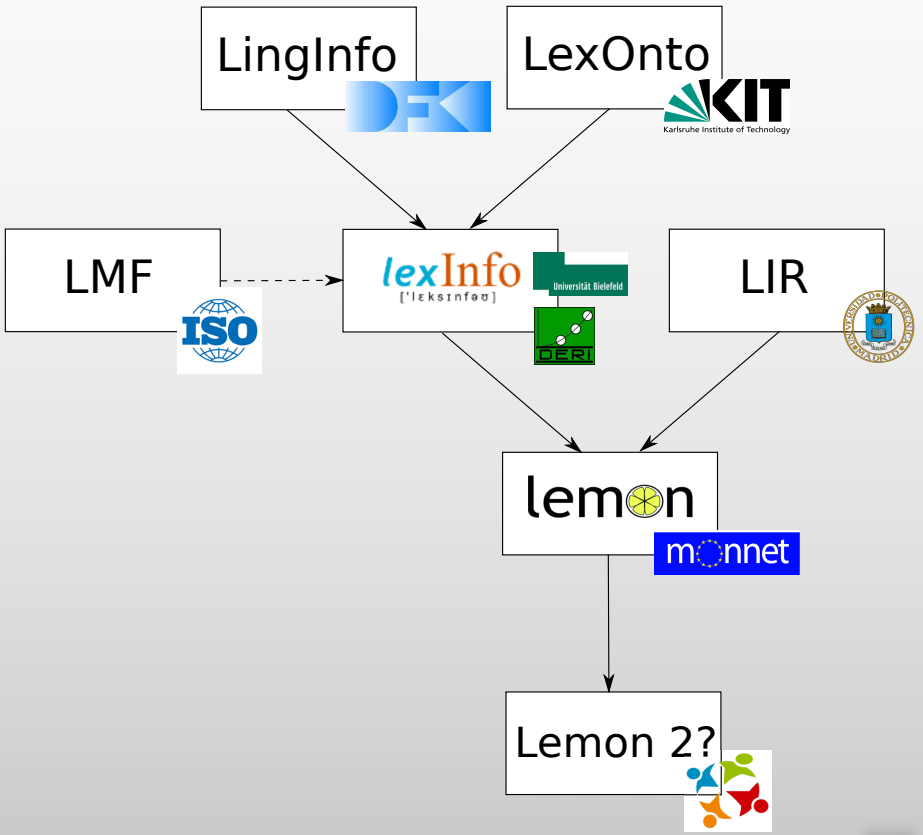


Wiktionary as a lexicon



lemon's origins

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



Design goals

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

Why lemon: RDF(S)

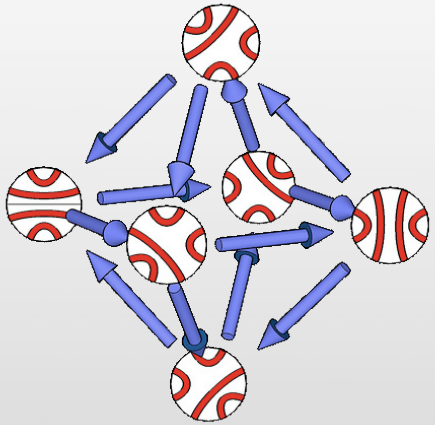
- RDF models are labelled directed graphs
  - Allows for smarter representation
- Each entry has a URI
  - Queriable on the web using standards
  - Clear responsibility for data
- Linking possible between different lexica
  - Reuse of lexicon data
- Some induction possible (subproperties, classes etc.)

Why lemon: Conciseness

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

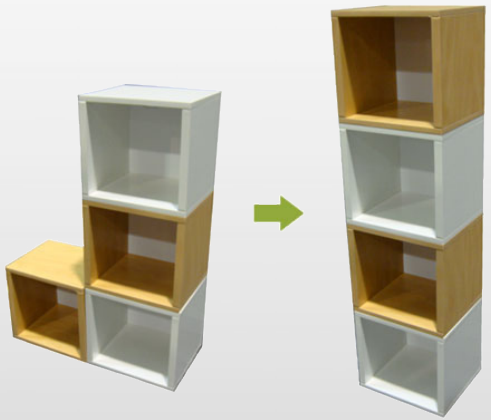
Why lemon: Semantics by Reference

- 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



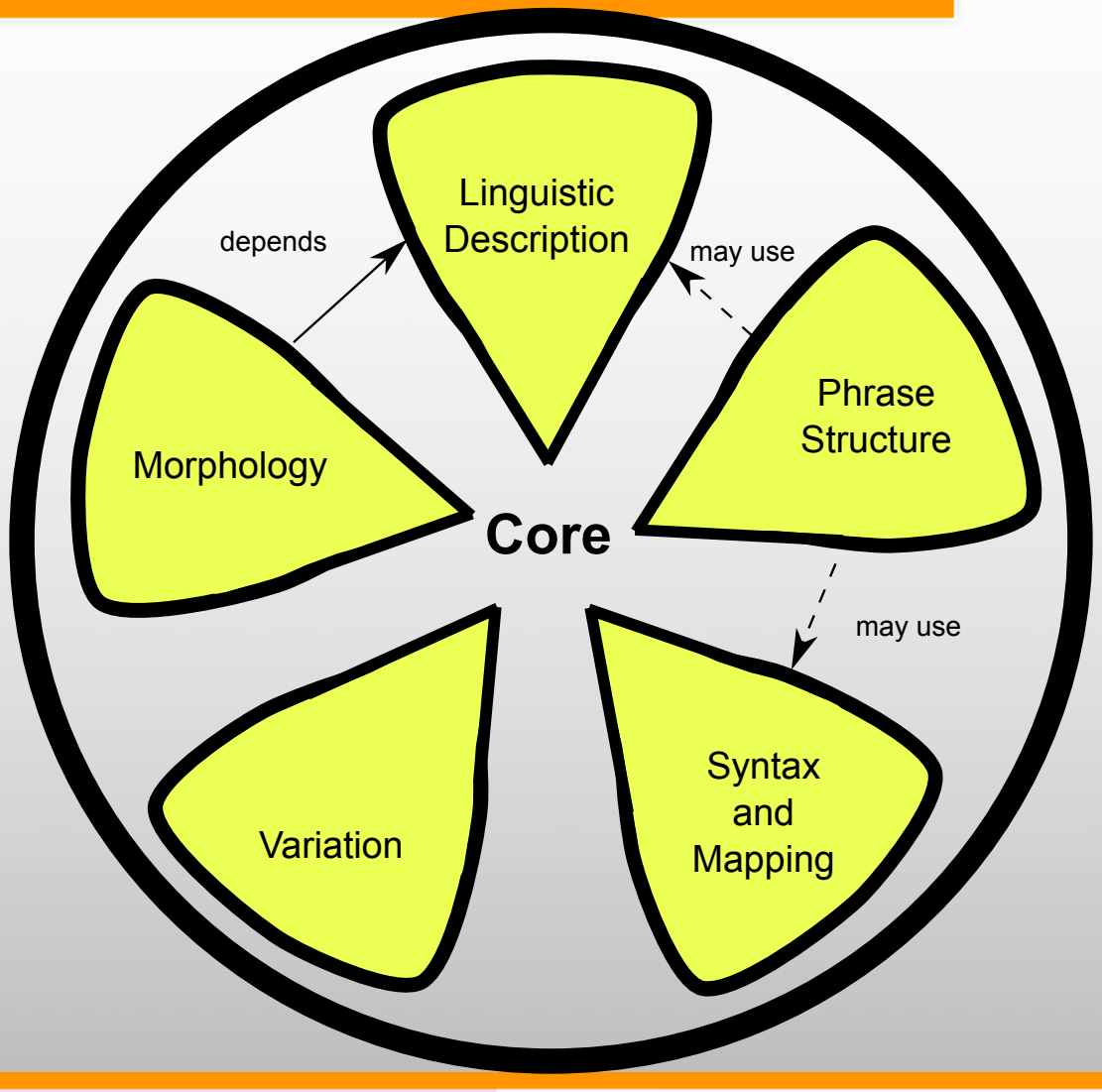
Why lemon: Modular and extensible

- 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



Outline

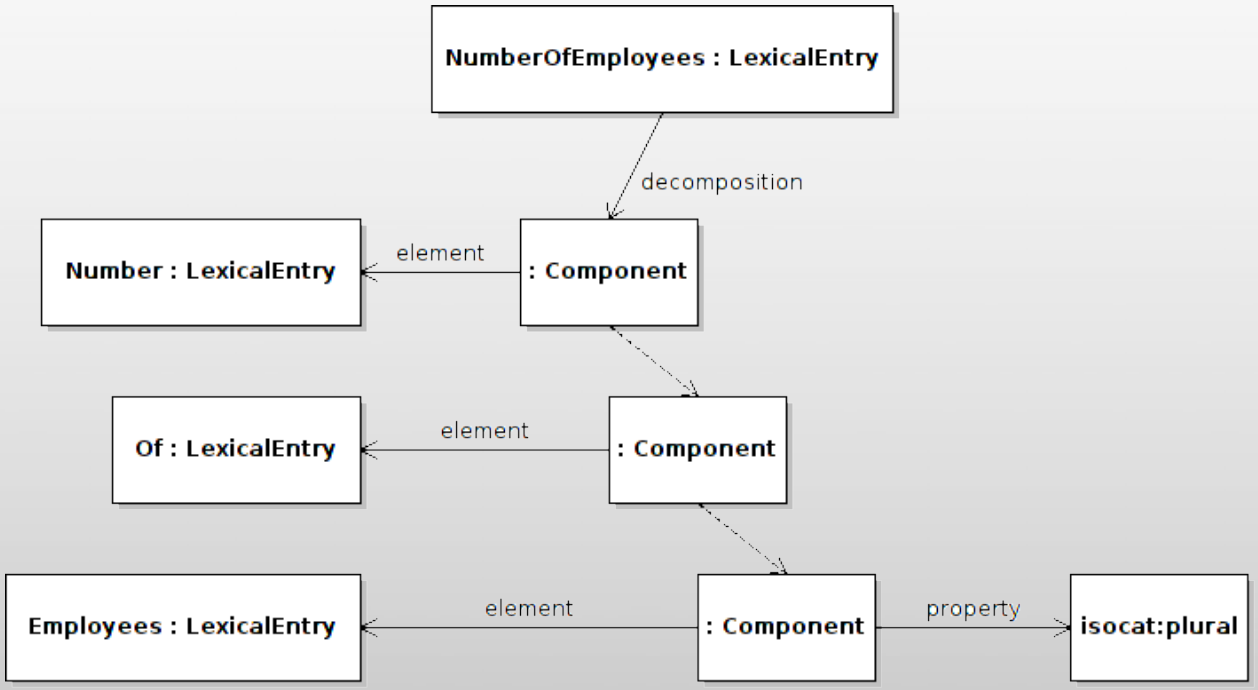
Modules


Linguistic Description
depends
may use
Morphology
Phrase Structure
Core
may use
Variation
Syntax and Mapping

Decomposition

- Entries marked as Word,Phrase or Part (of word)
- Decomposed into sub-entries
  - Phrase $\longrightarrow$ Words
  - Word $\longrightarrow$ Words and/or Parts
- Implemented by RDF list
  - Ordered
- Components may be marked to show necessary form properties

Decomposition: example



Properties

- Any element in the lexicon may have properties
- All propeties are stated as subproperties of lemon's `property`
- lemon does not have any such properties or values. A separate ontology is required
  - e.g., ISOcat, GOLD, LexInfo

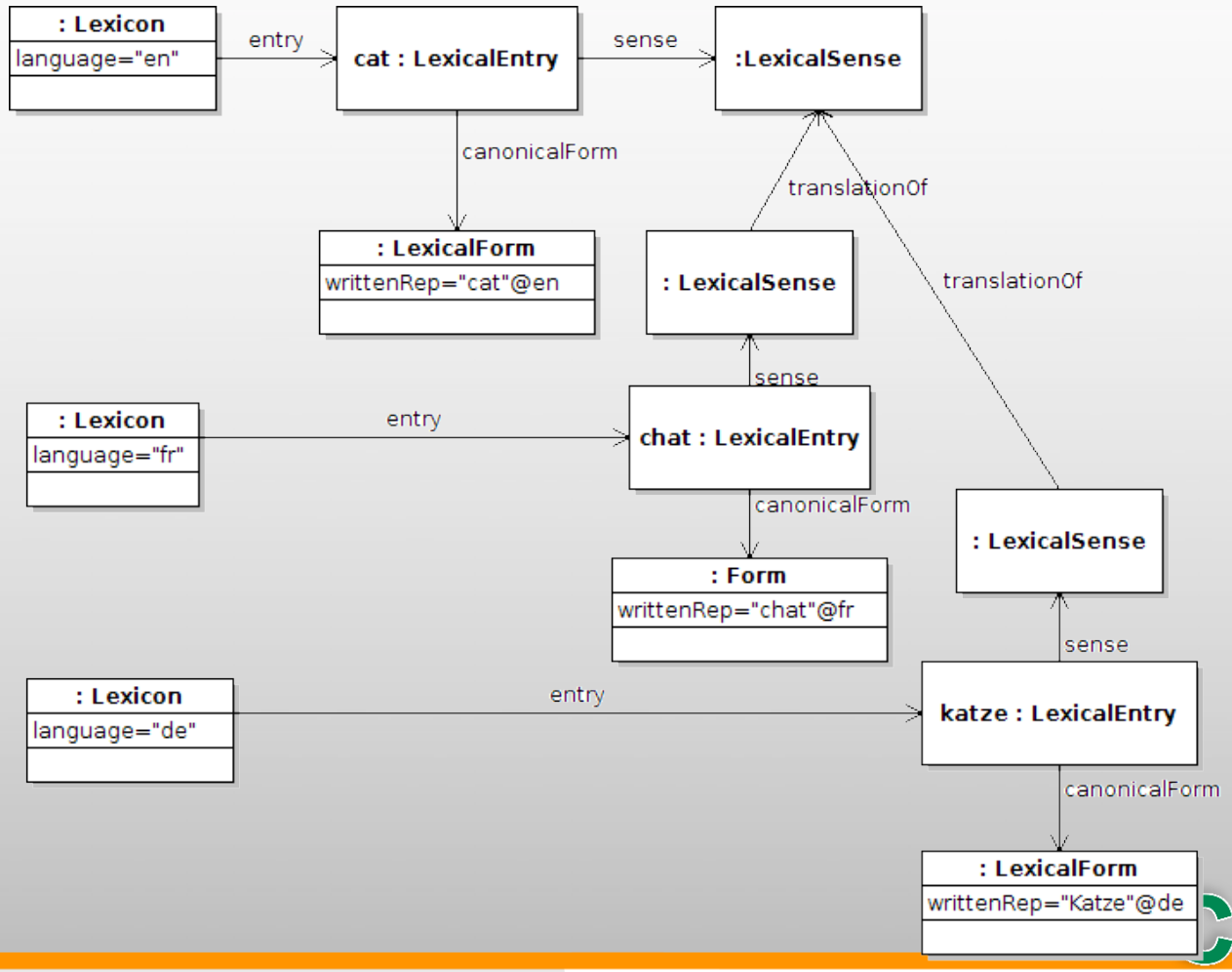
Properties: example

```
@prefix isocat: <http://www.isocat.org/datcat/> .

:katt a lemon:Word ;
  lemon:canonicalForm [
    lemon:writtenRep "katt"@swe ;
    isocat:DC-251 isocat:DC-252 ] ; # number=singular

  lemon:otherForm [
    lemon:writtenRep "katter"@swe ;
    isocat:DC-251 isocat:DC-253 ] . # number=plural

isocat:DC-251 rdfs:subPropertyOf lemon:property .
```

Variation

- Forms, Entries and Senses may be marked as variants
- Again, few lemon properties, mostly use external ontology
- Mark links as subproperties of `formVariant`, `lexicalVariant`, `senseRelation`
- Sense Relation does have subproperties `equivalent`, `broader`, `narrower`, `incompatible`
- Sense Relation can be used to model `translationOf`

Variation: example



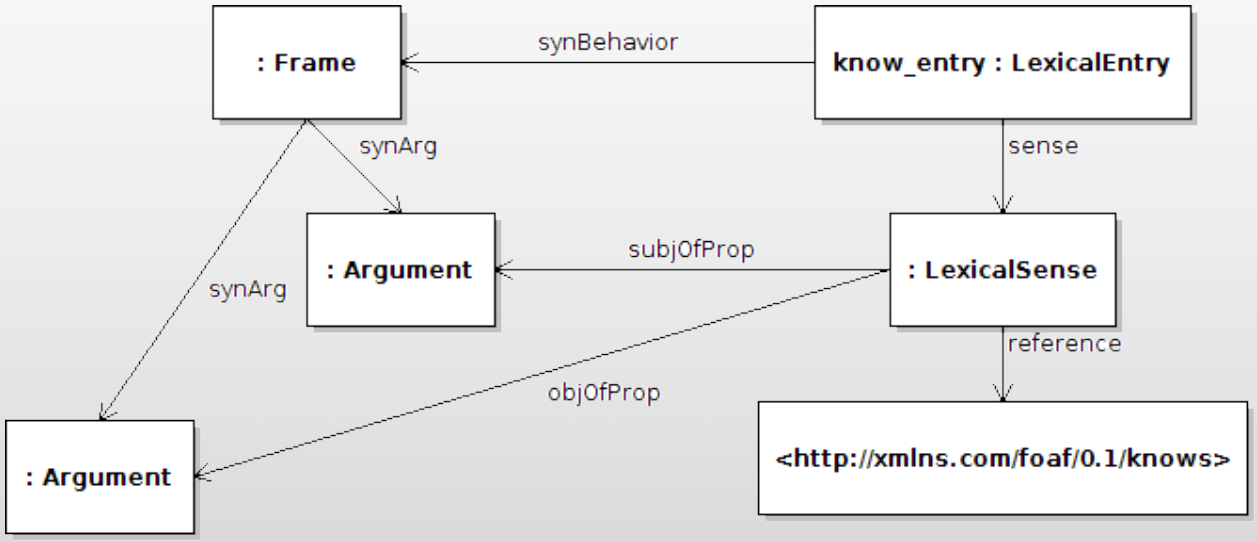
Frames and Correspondence

- The verb "know" is always used in a sentence
  - "John knows Lars"
- Similarly `foaf:knows` is only used in a triple
  - `agsc:jmccrae foaf:knows gu:lars.borin`
- It is necessary to state how these corresponds

Frames and Correspondence

- Linguistically we define each word as having a subcategorization frame
  - e.g., "X knows Y"
- Each RDF property has two arguments
  - Subject
  - Object
- We need to state the correspondence of syntactic arguments and semantic arguments

Frames and Correspondence



Outline

OntoLex Community Group

- W3C setup community groups as "an open forum, without fees, where Web developers and other stakeholders develop specifications"
- OntoLex group was set up with the following goals
  1. Develop model for representation of lexica relative to ontologies
  2. Demonstrate value of representing lexica on the Semantic Web
  3. Best practices for data categories
  4. Show improvement in NLP by means of ontology-lexica
  5. Bring together people working on standards for linguistic information
  6. Build interoperability between existing models
- Chaired by Philipp Cimiano (Uni Bielefeld) and Paul Buitelaar (DERI, Galway)
- 64 Participants across >40 institutes

CITEC

lemon in OntoLex

- lemon is taken as the baseline model for the group
- Agreement on core of model:
  - Semantics by reference
  - Modulated by reified link (senses)
  - Decomposition of terms
  - Properties and Relations
  - Ontology mapping
  - Morphology
- New requirements:
  - Metadata
  - Lexico-syntactic patterns
  - Interface with *Nets

Outline

Outline

Linked Data

- The linked data principles (Berners-Lee, 2006)
  1. Use URIs to identify things.
  2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
  3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF/XML.
  4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the Web.

Linked Data

- Linked data requirements (Cyganiak, 2011)
  1. Resolvable (All URIs valid)
  2. In RDF
  3. >1000 Triples
  4. >50 links to other datasets
  5. Crawlable (Index, dump or search)
  6. Registered (At CKAN)

Why linked data for linguistics? I

- Representation and modelling
  - Directed graph model perfect for LRs (cf. GrAF)
- Structural interoperability
  - RDF as common format
  - Data coexists in same DB
- Federation
  - Resources must not be in same (physical) location

From Chiarcos et al., 2013, “Towards Open Data for Linguistics: Linguistic Linked Data”.

Why linked data for linguistics? II

- Ecosystem
  - SPARQL databases
  - OWL Reasoners
  - SW conferences
  - etc.
- Expressivity
  - Reuse of vocabularies for axioms, metadata etc.
- Conceptual Interoperability
  - Names defined by links
- Dynamicity
  - Errors can be corrected after release
  - Not always a good thing!

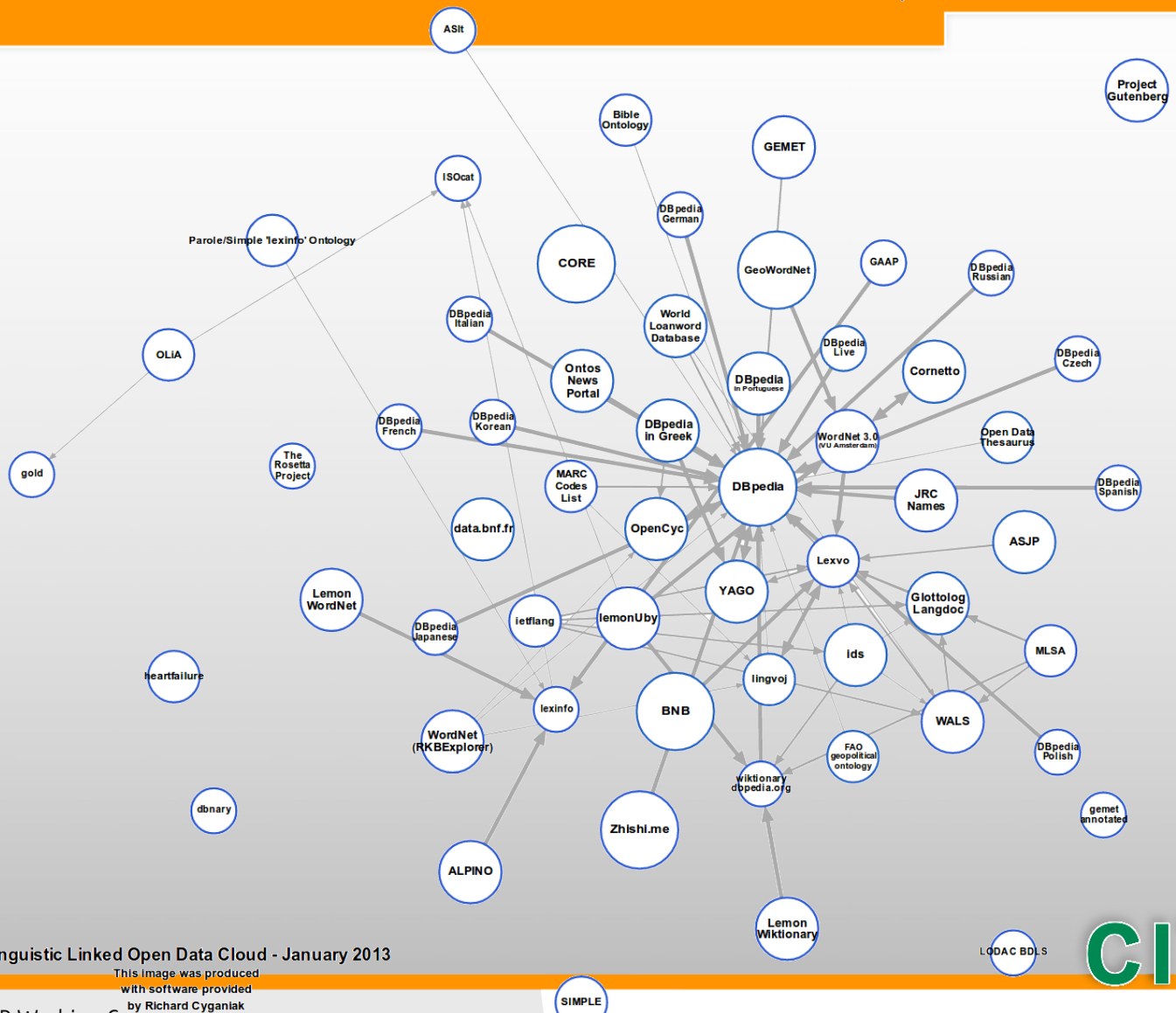
From Chiarcos et al., 2013, “Towards Open Data for Linguistics: Linguistic Linked Data”.

Outline

LLOD Working Group

- Open Knowledge Foundation's Working Group on Open Data in Linguistics.
- Purpose:
  - Promote open data in Linguistics
  - Act as a central point of reference and support
  - Facilitate communication between researchers
  - Mediate between providers and users
  - Build and maintain an index of open linguistic data sources and tools
  - Assemble best-practice guidelines and use cases
  - Gather information on legal issues
- Founded by Christian Chiarcos (Frankfurt), Sebastian Hellmann (Leipzig), Sebasitan Nordhoff (MPI-EVA, Leipzig)

LLOD Cloud


ASIt
Project Gutenberg
Bible Ontology
GEMET
ISOcat
DBpedia German
Parole/Simple 'lexinfo' Ontology
CORE
GeoWordNet
GAAP
DBpedia Russian
DBpedia Italian
World Loanword Database
DBpedia Live
OLiA
Ontos News Portal
DBpedia in Portuguese
Cornetto
DBpedia Czech
DBpedia French
DBpedia Korean
DBpedia in Greek
WordNet 3.0 (VU Amsterdam)
Open Data Thesaurus
gold
The Rosetta Project
MARC Codes List
DBpedia
JRC Names
DBpedia Spanish
data.bnf.fr
OpenCyc
ASJP
Lexvo
Lemon WordNet
YAGO
Glottolog Langdoc
DBpedia Japanese
ietflang
lemonUby
lds
MLSA
heartfailure
lingvoj
lexinfo
BNB
WALS
WordNet (RKBExplorer)
FAO geopolitical ontology
DBpedia Polish
wiktionary dbpedia.org
dbnary
Zhishi.me
gemet annotated
ALPINO
Lemon Wiktionary
LODAC BDLS
SIMPLE
Linguistic Linked Open Data Cloud - January 2013
This image was produced with software provided by Richard Cyganiak

Resources using lemon

- Uby (Darmstadt; Gurevych et al., 2012)
  - Existing resources, standardized to LMF and interlinked
  - WordNet, FrameNet, VerbNet, OmegaWiki, Wiktionary
  - English, German
- wiktionary.dbepdia.org (Leipzig)
  - Conversion of Wiktionary to lemon
  - Many languages
- PAROLE/SIMPLE Lexicon (Pompeu Fabra, Barcelona; Villegas and Bel, Under review)
  - English, Spanish, Catalan
- DBNary (Grenoble; Sérasset, 2012)
  - Conversion of Wiktionary to LMF and lemon
  - French, English and German.

Outline

Applications of lemon

- Ontology-based information extraction
- Ontology localization
- Natural language generation
- Integration into NLP pipelines (Davis et al., 2011)
- Question answering

Outline

Question Answering with Pythia

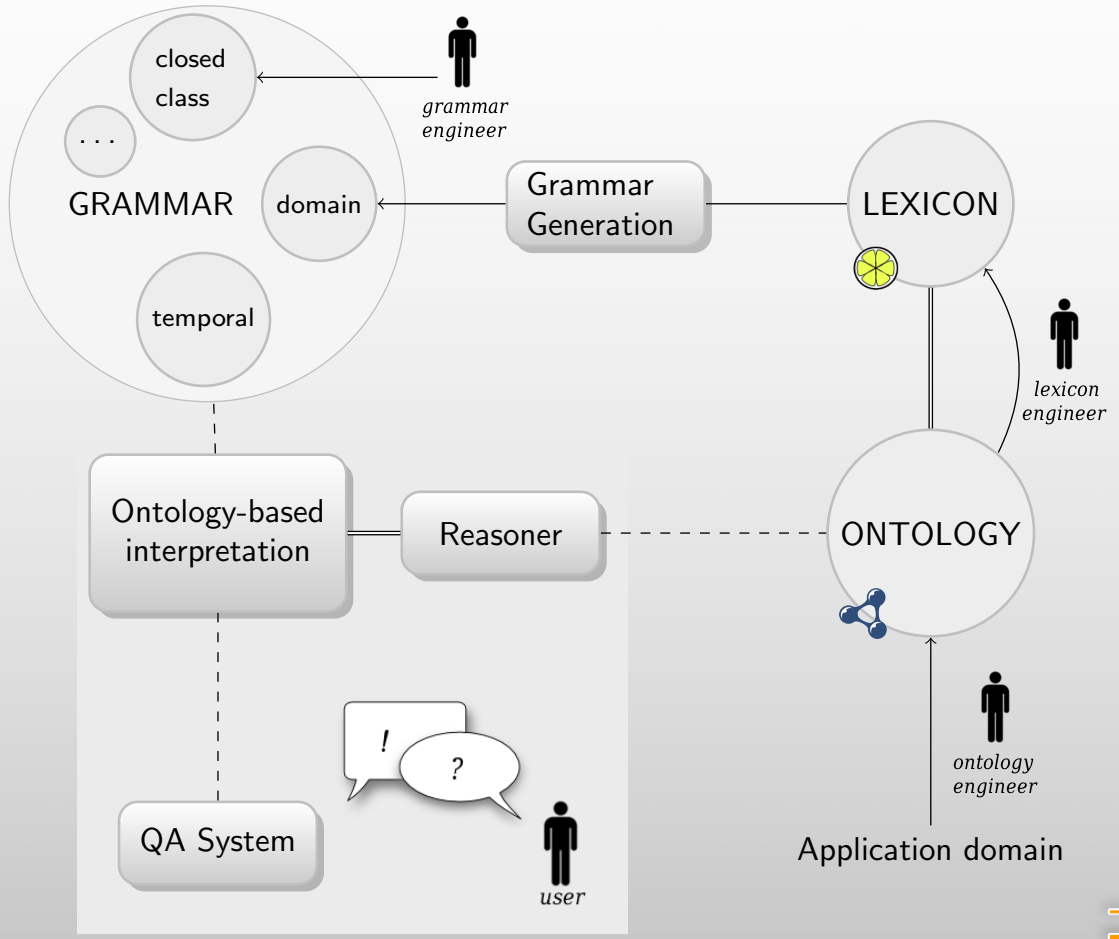
Pythia is an ontology-based question answering system that

- translates natural language into formal queries
- is developed in Bielefeld, by Christina Unger and Philipp Cimiano
- can successfully handle small domains (Geobase, MusicBrainz)
- is now ready to scale to larger domains on the Semantic Web (such as DBpedia)

Main workflow

- Specify verbalizations of ontology concepts and use the resulting lexicon-ontology interface for automatic grammar generation
  - this way, the grammar uses vocabulary aligned to the ontology and thereby ensures a precise and correct mapping from natural language expressions to ontology concepts
- based on grammar: compositional construction of principled linguistic representations
  - allows the interpretation of linguistically complex questions (e.g. involving quantification, comparatives and superlatives, negation, etc.)
  - allows easy integration of linguistic insights and tools (e.g. ambiguity resolution)

Pythia Architecture



Pythia: Example (lemon)

```
   MusicBrainzLexicon:collaboratesWith lemon:sense [ lemon:reference
<http://purl.org/vocab/relationship/collaboratesWith> ;
                                                    lemon:subjOfProp :arg1collab ;
                                                    lemon:objOfProp :arg2collab ] ;
   lemon:synBehavior [ rdf:type lexinfo:IntransitivePPFrame ;
          lexinfo:subject :arg1collab ;
                        lexinfo:prepositionalObject :arg2collab ] ;
   lemon:canonicalForm [ lemon:writtenRep "collaborates"@en ;
                         lexinfo:tense lexinfo:present ;
                         lexinfo:person lexinfo:thirdPerson ;
                         lexinfo:number lexinfo:singular ] ;
   lemon:otherForm [ lemon:writtenRep "collaborate"@en ;
                         lexinfo:tense lexinfo:present ;
                         lexinfo:person lexinfo:thirdPerson ;
                         lexinfo:number lexinfo:plural ] ;
   lemon:otherForm [ lemon:writtenRep "collaborating"@en ;
                         lexinfo:verbFormMood lexinfo:gerundive ] ;
   lemon:otherForm [ lemon:writtenRep "collaborated"@en ;
                         lexinfo:verbFormMood lexinfo:participle ;
                         lexinfo:aspect lexinfo:perfective ] ;
   lexinfo:partOfSpeech lexinfo:verb .

   :arg2collab lemon:marker :with .
```

Pythia: Example (LTAG)

```
 collaborates with ||
(S DP[domain] (VP V:'collaborates' P:'with' DP[range])) ||
<e, l1, t, [ l1:[ | ], l2:[ | rel:collaboratesWith(x,y) ] ],
           [ (l3,x,domain,<<e,t>,t>), (l4,y,range,<<e,t>,t>) ],
           l3<l1, l4<l1, l2<scope(l3), l2<scope(l4) ], []>
```

- (Show automatic generation)
- Graphical version on p. 221 of automatic generation

Universität Bielefeld

Outline

CITEC

Grammatical Framework

- "A special purpose language for grammars"
- Developed by Aarne Ranta (Gothenburg)
- Applications
  - Translation
  - NL interfaces
  - Dialog systems
  - Natural language generation

Mapping lemon to GF

1. Take a lemon resource
2. Extract information using SPARQL Queries
3. Use templating language to generate GF
4. Get GF Grammars

- `http://prezi.com/fxy36jugmiep/lemon-sparql-mustache-gf/`
- Original version developed in Monnet
- Christina Unger's Python implementation: `https://github.com/cunger/lemon2gf`

Why map lemon to GF

- Increasing number of lexica in lemon available on the web
- lemon is an interchange format for other systems
- Similarity in modelling
  - Abstract Grammar ⇔ Ontology
  - Concrete Grammar ⇔ Lexicon
- Extension of GF to enable ontology reasoning?

Universität Bielefeld

Outline

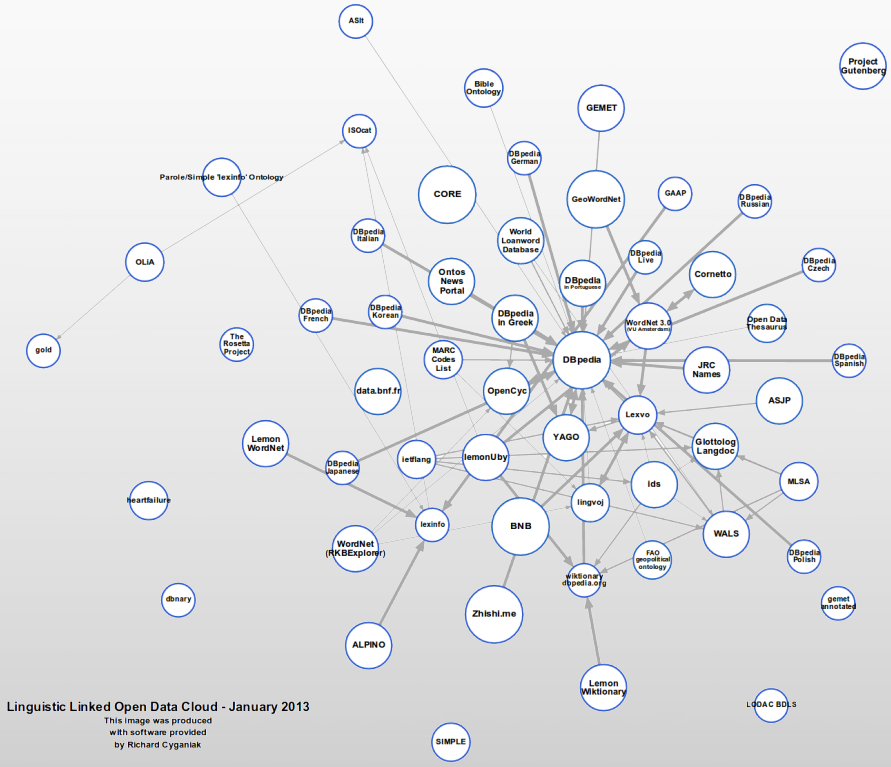
CITEC

lemon

- lemon is a model for ontology-lexica
- Sophisticated linguistic modelling
- Principled, logical ontological semantics
- Clear interface between two layers
- Concise and modular
- Further development in W3C Community Group
  - Potentially leading to recommendation/standardization

Linguistic Linked Data

- New paradigm for publishing language resources
- Better interoperability
- Improved lifecycle
- Increased visibility



Application of lemon

- Early stages (resources are now becoming available)
- Publishing/sharing data from the web has much promise
- Interaction with GF already showing results

Links

- http://lemon-model.net/
- http://www.monnet-project.eu/
- http://www.sc.cit-ec.uni-bielefeld.de/
- http://john.mccr.ae/
- http://www.w3.org/community/ontolex

References I

Berners-Lee, Tim (2006). Linked Data. URL:
http://www.w3.org/DesignIssues/LinkedData.

Buitelaar, P (2010). ``Ontology-based Semantic Lexicons: Mapping between terms and object descriptions''. In: Ontology and the Lexicon, pp. 212–223.

Chiarcos, Christian et al. (2013). ``Towards Open Data for Linguistics: Linguistic Linked Data''. In: New Trends of Research in Ontologies and Lexical Resources.

Cyganiak, Richard (2011). The Linking Open Data cloud diagram. URL:
http://lod-cloud.net/#how-to-join.

Davis, Brian et al. (2011). ``Squeezing lemon with GATE''. In: MSW 2011, p. 74.

Gurevych, Iryna et al. (2012). ``UBY--A Large-Scale Unified Lexical-Semantic Resource Based on LMF''. In: Proc. EACL. Citeseer.

Polikoff, Irene (2013). Who needs SKOS-XL? Maybe no one. URL:
http://topquadrantblog.blogspot.de/2012/07/who-needs-skos-xl-maybe-no-one.html.

References II

Sérasset, Gilles (2012). ``Dbnary: Wiktionary as a LMF based Multilingual RDF network''. In: Language Resources and Evaluation Conference.

Villegas, Marta and Nuria Bel (Under review). ``PAROLE/SIMPLE 'LexInfo' ontology and lexicons''. In: