UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann,
Michael Matuschek, Christian M. Meyer, and Christian Wirth
http://www.ukp.tu-darmstadt.de/uby/

UBY in a Nutshell

Developing a large-scale unified lexical-semantic resource (LSR) in a standardized format for NLP

Motivation:

- *Problem*: Limited resource coverage in NLP
  - *Solution*: Integration of LSRs
- *Problem*: Incompatible formats of (integrated) LSRs
  - *Solution*: Standards for modeling LSRs

Vision: *One-stop resource*

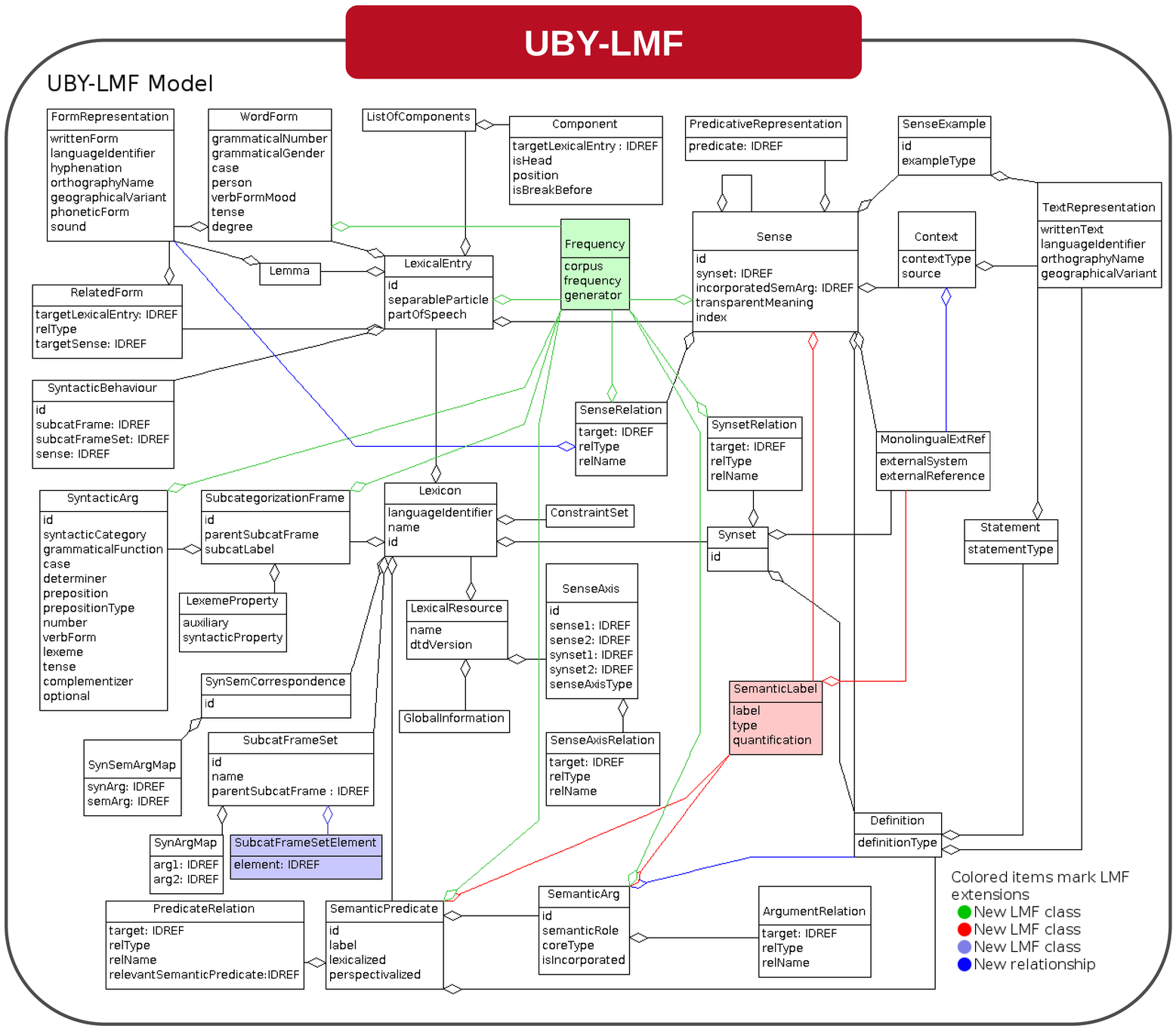
Contributions:

1. **UBY-LMF**: Lexicon Model
   - ISO Standard Lexical Markup Framework (LMF)
   - Expert-built and collaboratively created LSRs
   - Fine-grained modeling of information types
2. **UBY**
   - 9 LSRs in English and German
   - 9 pairwise sense alignments between LSRs
3. **UBY-API**:
   - **Uniform access** to LSRs in UBY via JAVA API

Applications:

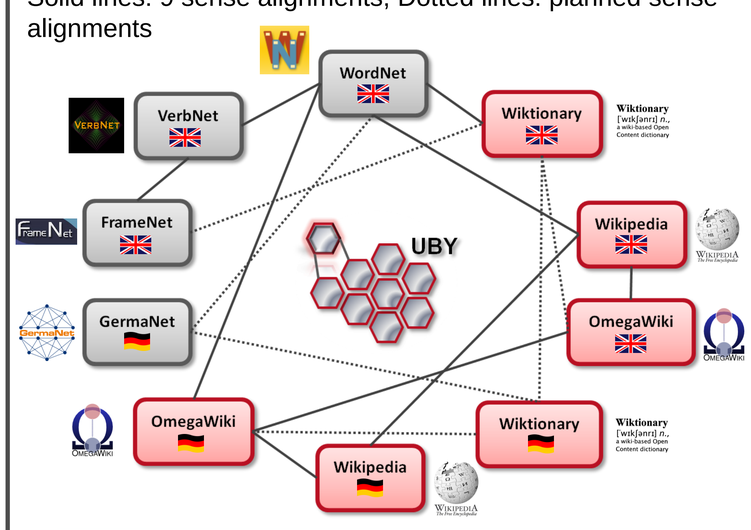
Word Sense Disambiguation,
Information Extraction,
Machine Translation,
Question Answering,…

UBY-LMF



UBY

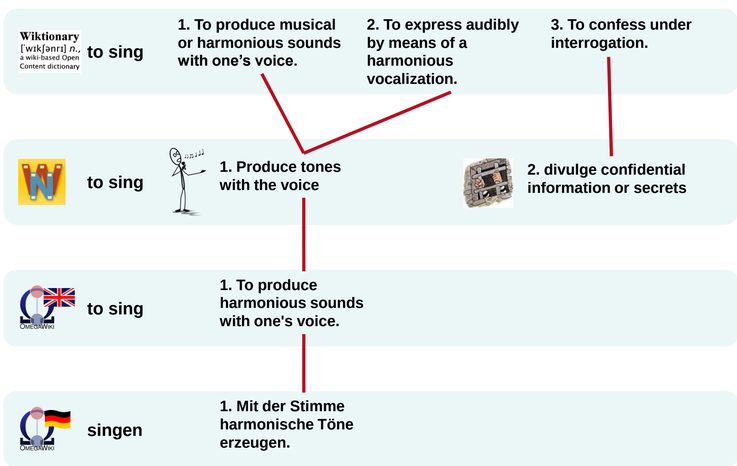
10 UBY Lexicons
Solid lines: 9 sense alignments; Dotted lines: planned sense alignments



UBY Lexicon	Lexical Entry	Sense	Sense Relation
FrameNet 1.5	9,704	11,942	-
GermaNet 6.0	83,091	93,407	329,213
OmegaWiki-DE	30,967	34,691	60,054
OmegaWiki-EN	51,715	57,921	85,952
Wikipedia-DE	790,430	838,428	571,286
Wikipedia-EN	2,712,117	2,921,455	3,364,083
Wiktionary-DE	85,575	72,752	434,358
Wiktionary-EN	335,749	421,848	716,595
WordNet 3.0	156,584	206,978	8,559
VerbNet	3,962	31,891	-
UBY	**4,259,894**	**4,691,313**	**5,570,100**

Sense Alignments

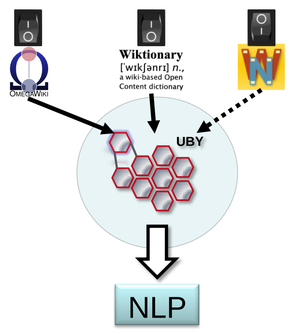
Enable **semantic operability** between resources in UBY:



- Senses linked by **SenseAxis** class
  - Currently 758 435 SenseAxis instances
- **Existing alignments**
  - Created by experts, users, or automatically
- **Automatic alignments**
  - Created by **Aligment Framework**:
    - Cross-lingual alignment (OmegaWiki-DE – WordNet)
    - Based on semantic similarity (+ machine translation)
- **Planned alignments**
  - OmegaWiki-EN – Wiktionary,…

Using UBY

UBY-API
- JAVA with Hibernate
- NLP application:
  - Easy *swapping* of resources
  - Easy *combination* of resources



Open licenses for data and software:
- UBY-API, LSR to UBY conversion routines: http://code.google.com/p/uby/
- UBY-LMF DTD, UBY release 1.0 (MySQL database dump): www.ukp.tu-darmstadt.de/uby/

See also

UBY-LMF
Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer: **UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF**, in: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), (to appear), May 2012. Istanbul, Turkey.

Subcat-LMF
Judith Eckle-Kohler and Iryna Gurevych: **Subcat-LMF – Fleshing out a standardized format for subcategorization frame interoperability**, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), (to appear), April 2012. Avignon, France.

This work has been supported by the
Emmy Noether Program of the German Research Foundation (DFG) under the grant No. 798/3-1
and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806.