UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF

Judith Eckle-Kohler[‡], Iryna Gurevych[†‡], Silvana Hartmann[‡], Michael Matuschek[‡] and Christian M. Meyer[‡]

† Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information, Schloßstr. 29, 60486 Frankfurt, Germany
‡Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

We present UBY-LMF, an LMF-based model for large-scale, heterogeneous multilingual lexical-semantic resources (LSRs). UBY-LMF allows the standardization of LSRs down to a fine-grained level of lexical information by employing a large number of Data Categories from ISOCat. We evaluate UBY-LMF by converting nine LSRs in two languages to the corresponding format: the English WordNet, Wiktionary, Wikipedia, OmegaWiki, FrameNet and VerbNet and the German Wikipedia, Wiktionary and GermaNet. The resulting LSR, UBY (Gurevych et al., 2012), holds interoperable versions of all nine resources which can be queried by an easy to use public Java API. UBY-LMF covers a wide range of information types from expert-constructed and collaboratively constructed resources for English and German, also including links between different resources at the word sense level. It is designed to accommodate further resources and languages as well as automatically mined lexical-semantic knowledge.

Keywords: Lexical Markup Framework, lexicon model, interoperability

1. Introduction

In recent years, the big demand for lexical-semantic resources (LSRs) in NLP has manifested itself in growing efforts to combine and integrate large-scale LSRs in order to enhance the performance of major NLP tasks such as word sense disambiguation and semantic role labeling. Previous integration projects covered expert-constructed resources (ECRs), such as WordNet, FrameNet (e.g., Johansson and Nugues (2007)), and collaboratively constructed resources (CCRs), such as Wikipedia and Wiktionary (e.g., Navigli and Ponzetto (2010), Meyer and Gurevych (2011)) as well as large-scale lexical acquisition and automatic integration of ECRs (Padró et al., 2011).

However, previous work on combining LSRs has rarely used standards to harmonize the representation of LSRs and thus to make them interoperable. Interoperability is a prerequisite of a smooth integration of LSRs. For this purpose, the ISO standard Lexical Markup Framework (LMF: ISO 24613:2008), a standard with a particular focus on lexical resources for NLP (Francopoulo et al., 2006), would have been an obvious choice. Yet, LMF has only been used in few integration efforts which were restricted to a specific domain or covered only a small range of LSR types (Quochi et al., 2008; Attia et al., 2010; Hayashi, 2011).

Since LMF defines an abstract meta-model of LSRs, applying LMF requires some effort to develop an LMF lexicon model, an instantiation of LMF. Thus, no attempts have been made so far to apply LMF on a large scale. Directly related to the previous neglect of LMF is the lack of APIs that facilitate easy access to integrated LSRs that are linked at the word sense level (this kind of linking is referred to as *sense alignment* hereafter).

To address these points, we introduce a comprehensive instantiation of LMF for uniformly representing a wide range of LSRs and sense alignments between them. Our lexicon model covers two types of LSRs: ECRs and CCRs.

Standardizing these two divergent types of LSRs with a single lexicon model implies that two requirements are met:

(i) *Comprehensiveness*: The model should be able to represent all the lexical information present in the ECRs, because it has been compiled by linguistic experts. This requirement leads to a fully descriptive representation of LSRs, allowing for the integration of incomplete and possibly contradictory lexical information (e.g., from CCRs).

(ii) *Extensibility*: The model should be extensible by further information types, because CCRs are subject to ongoing mining.

One of the main challenges of our work is to flesh out a single lexicon model that is standard-compliant, yet able to express the large variety of information types contained in a large selection of very heterogeneous LSRs.

The contributions of this paper are twofold:

(1) UBY-LMF, an LMF-based lexicon model for large-scale, heterogeneous multilingual ECRs and CCRs to be used in NLP. We model the lexical information down to a fine-grained level of information types (e.g., linking of syntactic and semantic arguments) and offer an LMF-compliant representation of sense alignments between LSRs.

(2) Evaluating UBY-LMF by using it for the standardization of nine resources in two languages: English WordNet (WN, Fellbaum (1998)), Wiktionary (WKT-en)[1], Wikipedia (WP-en)[2], multilingual OmegaWiki (OW-en and

[1] http://www.wiktionary.org/
[2] http://www.wikipedia.org/

OW-de[3]), FrameNet (FN, Baker et al. (1998)), and VerbNet (VN, Kipper et al. (2008)); German Wiktionary (WKT-de), Wikipedia (WP-de), and GermaNet (GN, Kunze and Lemnitzer (2002)). We also populated UBY-LMF with pairwise sense alignments for a subset of resources. The resulting large LSR, called UBY, can be accessed by a Java API. This API offers unified access to the information contained in UBY, an immediate result of the standardized format.

2. Background

This section gives an overview of the LMF standard and discusses previous work on using LMF.

2.1. LMF

LMF defines a meta-model of LSRs in the Unified Modeling Language (UML) by providing UML diagrams (consisting of UML classes connected by relationships) that are organized in packages. Any specific LMF lexicon model, i.e., any instantiation of LMF, has to use the core package which models the traditional headword-based organization of LSRs where a lexical entry (`LexicalEntry` class) is conceived of as a pairing of meaning (`Sense` class) and form (`Lemma` class). Depending on the particular type of LSR to be standardized, LMF offers a number of extension packages, such as the Syntax extension for subcategorization lexicons or the Semantics extension for wordnets.
The development of an LMF lexicon model involves two steps: first, establishing the structure of the lexicon model and second, specifying the linguistic vocabulary used in the lexicon model.

Structural Interoperability. The structure of a lexicon model is established by selecting a combination of suitable LMF classes. This step contributes to *structural interoperability* of lexicons represented according to the model, as it fixes the high-level organization of lexical knowledge in an LSR, e.g., whether synonymy is encoded by grouping senses into synsets (using the `Synset` class) or by specifying sense relations (using the `SenseRelation` class).

Semantic Interoperability. The linguistic vocabulary of a lexicon model is specified by defining attributes for the LMF classes and, where possible, also their values. For instance, the `LexicalEntry` class could be enriched by an attribute `PartOfSpeech` (POS) with values such as `noun`, `verb`, `adjective`. According to the standard, attributes and values can freely be defined, but they have to refer to so-called Data Categories (DCs) from ISOCat[4], the implementation of the ISO 12620 Data Category Registry (DCR), see Broeder et al. (2010). This step contributes to *semantic interoperability* with respect to the meaning of the linguistic vocabulary, because linguistic terms used in an LSR are linked to their meaning defined externally within a DCR. Accordingly, any two LSRs that share the same set of DCs are semantically interoperable (Ide and Pustejovsky, 2010).

LMF – a meta model. It has been argued before that LMF provides an abstract model of lexical resources that is not immediately usable for encoding a specific LSR (Tokunaga et al., 2009). Instead, LMF has to be developed into a full-fledged lexicon model by defining suitable attributes for the classes given in LMF. Thus, any two instantiations of LMF are likely to be different, e.g., employing different classes from the LMF extensions or different attributes.
This calls for the development of a single, but comprehensive LMF model which is able to accommodate information from various types of LSRs. We have taken up the challenge to build such a model, which we call UBY-LMF.

2.2. Previous Work

Previous work on putting the abstract LMF standard into action yielded a number of different instantiations of the LMF meta-model.

Wordnets. Much previous work on LMF focused on developing LMF models for the standardization of wordnets. Soria et al. (2009) introduced Wordnet-LMF, a lexicon model for standardizing wordnets in various diverse languages, including Asian languages (e.g., Lee et al. (2009)). Later on, specific adaptations of WordNet-LMF to GermaNet (Henrich and Hinrichs, 2010) and to the Italian wordnet (Toral et al., 2010) have been presented. None of these wordnet-centered instantiations of LMF are able to represent fine-grained lexical-syntactic information types, e.g., related to argument structure, such as the linking of syntactic arguments and semantic arguments.

Machine Readable Dictionaries. Apart from wordnets, specific LMF models were developed for standardizing monolingual machine-readable dictionaries (MRDs) (Khemakhem et al., 2009) and bilingual MRDs (Maks et al., 2008; Savas et al., 2010). MRDs often suffer from an insufficiently formalized representation structure that prevents their immediate use in NLP applications. Hence, lexicon models of such MRDs are of limited value for NLP purposes.

Combined LSRs. The potential of LMF as a tool to support the combination of LSRs has only been exploited in few integration projects with limited scope. Quochi et al. (2008) describe an LMF model for the BioLexicon, a domain-specific lexicon integrating information from different biomedical sources as well as lexical data extracted from texts or domain ontologies. While the BioLexicon LMF model offers a fine-grained representation of morphological, syntactic and lexical-semantic information types, it is restricted to the biomedical domain.
Hayashi (2011) discusses some requirements for an LMF model to be used for integrating several wordnets and two bilingual dictionaries, but he does not go into detail about the actual implementation. Also related to this line of research is work in lexical acquisition, e.g., Attia et al. (2010) who report on the envisaged use of LMF for representing an LSR of Modern Standard Arabic (Attia et al., 2010) that has automatically been extracted from corpora and Arabic Wikipedia.

Linking LSRs and ontologies. A different line of research has been pursued by Buitelaar et al. (2009) and McCrae et al. (2011) who describe full-fledged LMF-based lexicon models, LEXINFO and LEMON, for representing lexical information relative to ontologies. These lexicon models focus on representing linguistic knowledge

[3] http://www.omegawiki.org/
[4] http://www.isocat.org

of lexemes, while the meaning of lexemes is defined externally in an ontology. LEXINFO, in particular, models fine-grained morphosyntactic information and subcategorization frames. As neither of the two models have been applied at a large scale for standardizing several different LSR types, their integration capability with respect to a wide range of information types remains unclear.

Summary. To summarize, previous work on LMF was constrained along several dimensions: First, only a few types of LSRs have been considered and second, more comprehensive lexicon models have not been populated on a large scale.

We build upon previous work, but extend it significantly: UBY-LMF goes beyond modeling a single type of LSR and covers a large number of both ECRs and CCRs with very heterogeneous content. At the same time, UBY-LMF features fine-grained modeling of lexical information types, ranging from morphology and lexical semantics to lexical syntax and the mapping between syntactic and semantic arguments. Moreover, UBY-LMF enables a standard-compliant representation of sense alignments between LSRs. In contrast to previous work, we perform an evaluation of UBY-LMF by automatically populating it with nine large-scale LSRs.

3. UBY-LMF

This section introduces our lexicon model, UBY-LMF.

3.1. Scope and Architecture

Types of ECRs and CCRs. We designed UBY-LMF to cover a range of different LSR types that play an important role in NLP applications. The ECR types considered specify lexical information at the sense level and include wordnet-type LSRs, LSRs based on frame semantics and subcategorization lexicons such as VN. These ECR types are differently organized and provide largely complementary information, e.g., (Baker and Fellbaum, 2009). While wordnets primarily contain information on lexical-semantic relations, such as synonymy, LSRs modeled according to frame semantics focus on predicate-like lexemes that evoke prototypical situations (Semantic Frames) involving semantic roles (Frame Elements). Subcategorization lexicons, on the other hand, may be organized in alternation classes (e.g., VN).

Likewise, we consider prototypical exemplars of different types of CCRs that have turned out to be particularly useful for NLP, i.e., WP, WKT and OW. WP primarily provides encyclopedic information on nouns and is organized in article pages. WKT is in many ways similar to traditional dictionaries, i.e., it enumerates senses under a given headword on an entry page. OW is a multilingual resource covering multiple languages. In contrast to WKT and WP, there are no separate editions for each language. Instead, OW is based on multilingual synsets, i.e., language-independent concepts to which lexicalizations of the concepts are attached (Matuschek and Gurevych, 2011).

Architecture. The architecture of UBY-LMF is defined by the top level classes of the mandatory core package: a `LexicalResource` instance consisting of one to many `Lexicon` instances. Each LSR is modeled as a separate `Lexicon` instance. This yields interoperable LSRs due to the uniform `Lexicon` specification and moreover, offers full transparency regarding the source of each information type. Note further that LMF requires each `Lexicon` instance to belong to exactly one language, a requirement that reflects the diversity of different languages at the morphosyntactic and lexical-syntactic level. Therefore, multilingual LSRs such as OW have to be split in separate `Lexicon` instances for each language.

3.2. Classes and Attributes

We use the `Lexicon` class and most of the classes from the LMF extension packages as a basis for a uniform lexicon model, see Figure 1. In LMF, the actual lexical information present in an LSR is modeled by means of class attributes and their values. In order to cover the various information types provided by different LSR types, we enriched the classes by a large number of attributes and values. There are many ways to define these attributes and attach them to classes. Our approach to do this was mainly driven by the requirement of extensibility described above. Figure 1 shows all attributes used in our model. We will briefly comment on some selected classes and attributes.

Core Package and MRD Extension. For open word classes, the values of the `LexicalEntry` attribute `partOfSpeech` encode a small hierarchy of POS by employing a common prefix notation that allows for convenient querying of lemmas filtered by POS, e.g., `noun`, `nounCommon`, `nounProper`.

UBY-LMF employs fine-grained attributes that specify various types of sense definitions, sense examples and various kinds of lexicographic notes that are encoded in the LMF `Statement` class (e.g., usage notes, encyclopedic information, external references). For sense examples which can be attributed to a specific source, i.e., a citation or evidence from a particular corpus, UBY-LMF offers the `Context` class from the MRD extension.

Morphology Extension. The `Component` class provides attributes for a detailed modeling of multiword expressions (MWEs) as present, e.g., in FN. This includes information on the head of an MWE (`isHead`) and on the possibility to break an MWE before the multiword component considered in order to insert additional constituents. For instance, the MWE *take on* consists of two components: *take* and *on*. The component *on* has the value `true` for `breakBefore`, e.g., *to take the job on*.

We introduced an attribute `targetSense` in the `RelatedForm` class in order to link morphological relations to a sense target, because morphologically related forms might be specified at the sense level, e.g., in WN. For instance, the verb *buy* (purchase) is derivationally related to the noun *buy*, while on the other hand *buy* (accept as true, e.g., *I can't buy this story*) is not derivationally related to the noun *buy*. Discarding such a sense-level specification would lead to information loss in some cases, which would not be in agreement with our requirement of comprehensiveness.

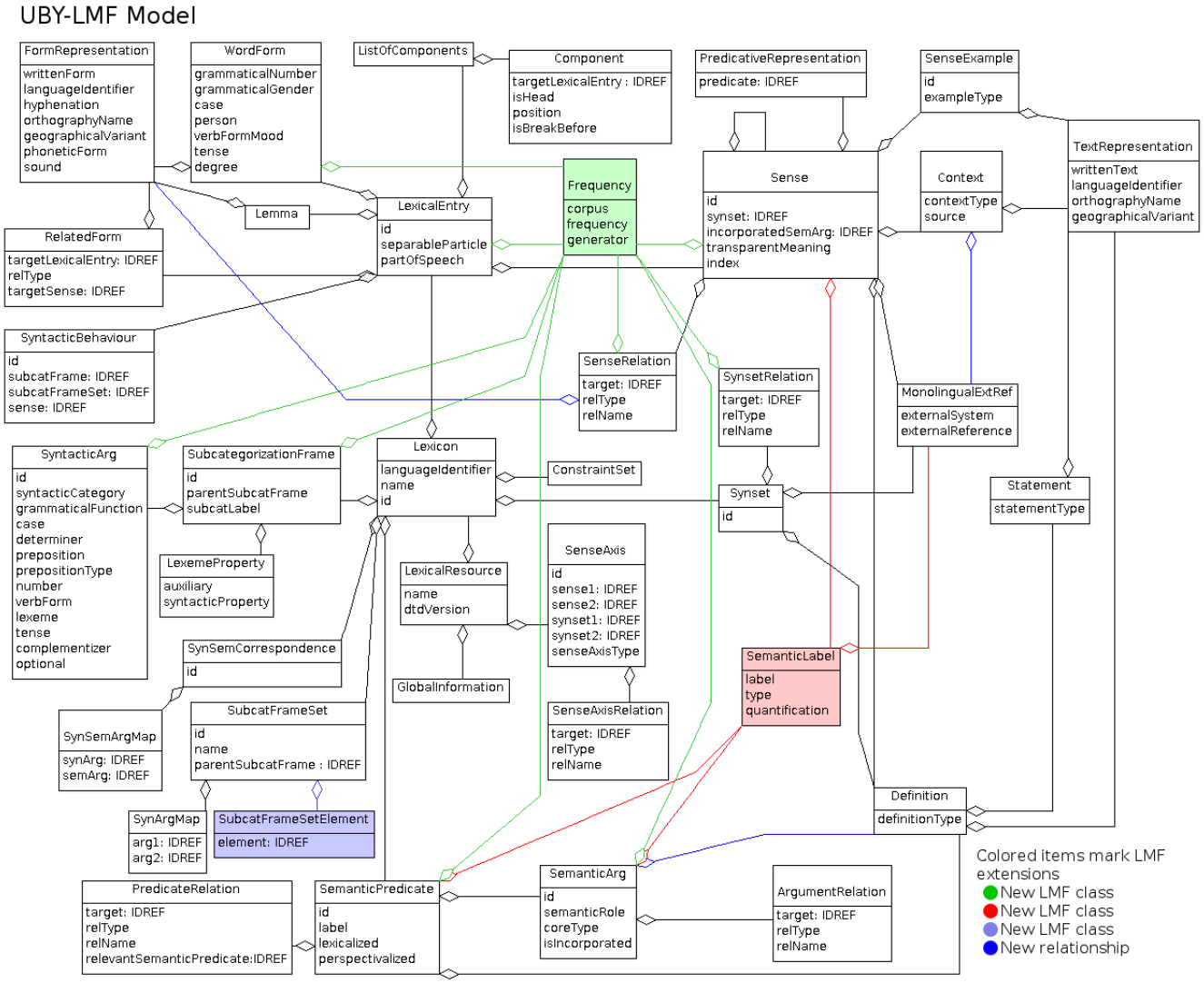

Figure 1: Classes and attributes in UBY-LMF (new classes and relationships are highlighted in red, green and blue).

Syntax Extension. Subcategorization frames are encoded in a compositional way by specifying syntactic arguments. The `SyntacticArgument` class has a large number of modular attributes. As subcategorization is highly language-specific, some of these attributes have language-specific values, e.g., `case`. A detailed description of how subcategorization frames are represented in UBY-LMF can be found in (Eckle-Kohler and Gurevych, 2012).

Semantics Extension. UBY-LMF makes use of the `MonolingualExternalRef` class, a subclass of `Sense`, to store sense IDs from the original LSRs. Access to original sense IDs is crucial for converting existing sense alignments between LSRs to UBY-LMF (section 4.2.).

The `semanticRole` attribute of `SemanticArgument` has a string value, as there is no standard set of semantic roles yet. We also encountered a lack of standard DCs for sense relations when taking both ECRs and CCRs into account. In particular, the CCRs provide a wide range of sense relation types which could not be mapped onto a standardized set of DCs without information loss. For instance, OW encodes sense relations such as *located in, partners with, is an allotrope of*, to name only a few. Thus, insufficient standardization of semantic roles and sense relations prevents our model from achieving full semantic interoperability at the lexical-semantic level. Of course, we will adopt results of ongoing and future standardization efforts in the lexical-semantic domain as they become available.

Multilingual Extension. UBY-LMF introduces a novel interpretation of the `SenseAxis` class from the LMF Multilingual Extension (Francopoulo et al., 2009). `SenseAxis` is used for the representation of sense alignments between LSRs. While this fully complies with LMF for cross-lingual alignments, it corresponds to a transferred use of `SenseAxis` for monolingual sense alignments.

Data Categories. The attributes and values defined in UBY-LMF refer to 175 ISOCat DCs.[5] We had to create

[5]The corresponding DCs can be found on the Uby website http://www.ukp.tu-darmstadt.de/data/uby and in ISOCat, see

38 new DCs in ISOCat, as particular definitions were missing. Thematic domains in ISOCat where we filled some gaps include lexical syntax (related to subcategorization), derivational morphology, and frame semantic information.

3.3. Extensions of ISO-LMF

The task of building a homogeneous lexicon model for heterogeneous LSR types and our requirement of extensibility encompassed a few extensions of LMF. We added three new classes and three new relationships between existing classes. These extensions are described below.

New Classes. First, we introduced a new `SemanticLabel` class, which is an optional subclass of `Sense`, `SemanticPredicate`, and `SemanticArgument`, to cope with a large variety of lexical-semantic labels for many different dimensions of semantic classification. Examples of such dimensions include ontological types (e.g., selectional preferences), domains, styles and registers, or sentiment (available, e.g., in FN). The `SemanticLabel` class has three attributes, encoding the name of the label, its type (e.g., ontological, register, sentiment), and a numeric quantification (e.g., sentiment strength). In this way, further dimensions of semantic classification can easily by represented in UBY-LMF.

Second, we attached the subclass `Frequency` to most of the classes in UBY-LMF, in order to encode frequency information which is of particular importance when using the resource in machine learning applications. The `Frequency` class has three attributes to encode the frequency, the corpus and the extraction tool used.[6]

Finally, we employed a class `SubcatFrameSetElement` as a subclass of `SubcategorizationFrameSet` for linking subcategorization frames belonging to a particular alternation set. This class replaces the attribute `SubcatFrameSet` suggested in the standard which has a set as value domain.[7] We found a class more appropriate than an attribute, because the cardinality of such an alternation set is not a priori given.

New Relationships. We added two new relationships between LMF classes to account for information types provided by FN: first, a link between `SemanticArgument` and `Definition`, in order to represent definitions of semantic arguments (Frame Elements in FN). Second, a relationship between the `Context` class and the `MonolingualExternalRef` class, to represent links to annotated corpus sentences.

Modeling WKT required adding another relationship between existing LMF classes. WKT contains a special kind of ambiguity in the semantic relations and translation links listed for senses: The targets of both relations and translation links are ambiguous, as they refer to lemmas (word forms), rather than to senses. These ambiguous relation targets could not directly be represented in LMF, as sense and translation relations are defined between senses. To resolve this, we linked `SenseRelation` and `FormRepresentation`, in order to encode the ambiguous WKT relation target as a word form. The disambiguation of these sense relation targets is left to future work.

4. Evaluating UBY-LMF

The actual population of the lexicon model on a large-scale can be considered as an evaluation of our model regarding its capability to represent a wide range of information types from different LSR types.

We automatically populated our model by converting the nine LSRs listed in section 1 (UBY-Lexicons hereafter) and sense alignments between them to a UBY-LMF-compliant format, yielding a large LSR called UBY, see http://www.ukp.tu-darmstadt.de/data/uby.

4.1. Converting UBY-Lexicons

The current representation of UBY-LMF is XML-based, that is, UBY-LMF is specified by a DTD. To convert the source LSRs to UBY-LMF, we developed Java-based conversion tools.[8] These tools extract information from the LSRs using their native APIs and convert it into Java objects, which are then imported into an SQL database or converted to XML. The corresponding Java Object Model directly mirrors the UBY-LMF model.

The conversion tools are based on manually defined mappings of the linguistic units and terms used in the UBY-Lexicons. Consider as an example the `Sense` class, which together with the `Lemma` class forms a `LexicalEntry`. The following units have been mapped to `Sense` instances: for WN and GN, pairs of lemma and synset, for FN, groups of lemma, POS and FN frame, for VN, groups of lemma, VN frame and semantic predicate, for WP, article pages for a lemma, for WKT, senses listed under an entry page, and for OW, pairs of lemma and multilingual synset.

To extract information from the CCRs, we used JWPL and JWKTL for WP and WKT (Zesch et al., 2008) and implemented an API to OW which has not been made publicly available yet. The mapping of these CCRs to UBY-LMF reflects the current functionality of these APIs and covers the following information types: for OW, all information is covered, for WP, the title and first paragraph, disambiguation pages, redirects and categories are mapped, and for WKT, POS, pronounciations, sense definitions, sense examples, sense relations, translations and a large variety of semantic labels are mapped.

In order to prove the correctness of the automatic conversion, we have compared the original resource statistics of classes and information types in the source LSRs to the corresponding classes in their LMF-compliant counterparts. For instance, the number of lexical relations in WN has been compared to the number of `SenseRelations` in the UBY WN lexicon.[9]

4.2. Converting Sense Alignments

For the conversion of existing sense alignments, the original sense IDs in the `MonolingualExternalRef` instances

http://www.isocat.org.

[6] This extension of the standard has already been made in WordNet-LMF.

[7] This attribute is used in the DTD example given in the Annex F of the ISO-LMF specification.

[8] All conversion tools are publicly available as open source.

[9] Detailed analysis results can be found on the Uby website http://www.ukp.tu-darmstadt.de/data/uby.

are used to identify corresponding senses in the LMF format and in the alignment data, which is typically given as sets of aligned sense IDs. These sense alignments are mapped to UBY-LMF by creating instances of `SenseAxis` for *pairs* of aligned senses.
We converted the following expert-quality sense alignments to UBY-LMF: VN–FN and VN–WN (Palmer, 2009), as well as the community-constructed sense alignments present in OW (alignments of OW entries and corresponding WP pages) and WP (inter-language links between articles in WP-en and WP-de).[10] In addition, we converted automatically created sense alignments: WN–WP-en (Niemann and Gurevych, 2011), WN–WKT-en (Meyer and Gurevych, 2011), and WN–OW-de (Gurevych et al., 2012). We plan to use the alignment framework described by Gurevych et al. (2012) to establish further alignments between the resources in UBY.

4.3. Summary

UBY currently contains more than 4.2 million lexical entries, 4.6 million senses, 5.3 million relations between senses and more than 700,000 alignments between senses. There are more than 860,000 unique German and 3.08 million unique English lemma-POS combinations. Based on the Hibernate framework, we implemented a Java API (the UBY-API) that provides easy to use access functions to some of the major LMF instances, e.g., `LexicalEntry`, `DefinitionText`.[11] The API supports both access of single LSRs (mirroring the behaviour of the legacy APIs) and cross-resource access of all LSRs combined. A tutorial showing the use of the UBY-API can be found under http://www.ukp.tu-darmstadt.de/data/uby.
The actual population of the lexicon model on a large-scale confirms the capability of UBY-LMF to represent a wide range of information types from differently organized LSRs. Hence, this step provides an evaluation of our model on real lexicon data.

5. Discussion and Outlook

The main contribution of this paper is the comprehensive instantiation of LMF – UBY-LMF – which can be used for the standardization of other English and German LSRs. While UBY-LMF particularly covers sense-disambiguated resources for NLP use, also modeling alignments between resources at the sense level, it is equally suitable for NLP lexicons that are organized at the lemma level. Our model might also be applicable to MRDs, since they have much in common with WKT. In the following paragraphs, we discuss in more detail a few important aspects related to a wider use of our model.

Alternative representations. There are many ways to implement an LMF lexicon model (Francopoulo et al., 2007). Currently, UBY-LMF is represented by a DTD. This XML-based representation of our model benefits from good tool support, but it has some drawbacks as a serialization of LMF, e.g., it is not possible to include links to ISOCat in the DTD as part of the schema. Instead, the UBY-LMF DTD specifies links to ISOCat DCs within XML comments.
For implementing LMF, there are alternative representation languages which might be more suitable in a particular context, e.g., XML Schema, RDF/OWL or RDF Schema. As part of future work, a representation of UBY-LMF in RDF/OWL along the lines of Ide et al. (2003) could be pursued in order to publish UBY as Open Data and link it to other open linguistic resources (Chiarcos et al., to appear 2012).

Extensibility. UBY-LMF is a scalable lexicon model, because it can be applied to other information types and languages with no or only minor changes:
First, the newly defined classes `SemanticLabel` and `Frequency` make our model immediately usable for automatically mined information types, such as corpus frequencies and lexical-semantic knowledge automatically extracted from corpus text.
Second, the extension of UBY-LMF with respect to other languages or information types mainly encompasses changing or adding attribute *values*, rather than adding new attributes. This is important for NLP systems using UBY, because changing only attribute values does in most cases not affect the API functions, thus keeping working NLP systems based on UBY intact.
Covering further languages requires adding values to attributes in the Syntax part of our model, e.g., to the attribute `case` of `SyntacticArgument`. Likewise, UBY-LMF can accommodate future results of standardizing semantic roles and lexical-semantic relations by replacing the currently used string value of attributes by standardized DCs.

Future Work. We plan to extend UBY-LMF to allow for the convenient representation of further information types related to the alignments of LSRs at various levels. First, we will employ a class `Meta` which can be attached to any LMF class and which encodes information on the generator of the lexical information (e.g., an extraction tool, a human user) and a confidence score. A `Meta` class has been used before in WordNet-LMF (Soria et al., 2009) for the same purpose.
Second, we will define new *Axis* classes to encode the linking of LSRs for information types other than sense. For instance, similar to the `SenseAxis` class which links senses, we will introduce a class `SemanticRoleAxis` which links semantic role sets from different LSRs.[12] In this way, we will generalize the `SenseAxis` class even more to account for alignments between LSRs at different levels.

NLP Applications and Beyond. An important outcome of our work is the resource UBY resulting from the large scale population of our lexicon model. We believe that this high-coverage LSR in combination with its Java-based API advances NLP research and applications. Most of all, the UBY-API is a uniform interface between LSRs and NLP applications. As such, it enables NLP applications to easily switch between LSRs, thus opening up the possibility to

[10] The alignments in the CCRs were entered manually by users and are subject to community control. Nevertheless, they are still less reliable than the expert-quality alignments.

[11] Alternatively, the API can be used to export lexical data from the database, using XML as export format.

[12] Linking of semantic roles from VN and FN has been part of the SemLink project, see http://verbs.colorado.edu/semlink/.

perform extensive task-based comparisons and evaluations of different LSRs.

Another point worth mentioning is the easy cross-resource access to a wide range of information types made possible by the UBY-API. In this context, UBY could be used to support standardization efforts, e.g., in the lexical-semantic domain. For instance, the large number of different semantic label types and sense relation types occurring in UBY can straightforwardly be extracted across all nine UBY-Lexicons and used as a broad basis for a comparative analysis of lexical-semantic information types.

Further Mining of CCRs. The presented conversion of CCRs to our model is constrained by (i) structural properties of the resources themselves (e.g., non-disambiguated sense relation targets in WKT) and, (ii), by the capability of their APIs to extract lexical information (e.g., the WKT API JWKTL does not yet provide access to morphosyntactic knowledge in WKT, such as inflectional properties of word forms). We plan to address these points in the future by mining and extracting further lexical knowledge from the CCRs.

6. Conclusion

We presented UBY-LMF, a uniform and comprehensive model for standardizing large-scale heterogeneous LSRs to be used in NLP. UBY-LMF enables structural and semantic interoperability across resources and languages down to a fine-grained level of semantic and syntactic information including sense alignments between resources. We performed an evaluation of our model by converting nine widely used resources in two languages to UBY-LMF yielding the large resource UBY. A Java API offers unified cross-resource access to all LSRs in UBY. The LMF model, the conversion tools, the resource UBY and the API are freely available to the research community. Due to the comprehensiveness of UBY-LMF and the availability of an API, we believe that UBY-LMF will boost standardization and evaluation of lexical resources at a large scale.

7. Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Richard Eckart de Castilho and Christian Wirth for their contributions to this work and Yevgen Chebotar, Zijad Maksuti and Tri Duc Nghiem for implementing large parts of the accompanying software.

8. References

Mohammed Attia, Lamia Tounsi, and Josef van Genabith. 2010. Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic. Technical report, The NCLT Seminar Series, DCU, Dublin, Ireland.

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as complementary resources for annotation. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP, pages 125–129, Suntec, Singapore.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pages 86–90, Montreal, Canada.

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards Linguistically Grounded Ontologies. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, pages 111–125, Berlin Heidelberg. Springer-Verlag.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff, to appear 2012. *Linking linguistic resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer, Heidelberg.

Judith Eckle-Kohler and Iryna Gurevych. 2012. 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)*, page (to appear), Avignon, France.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.

Gil Francopoulo, Nuria Bel, Monte Georg, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2007. Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. In *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, 43:57–70.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - A Large-Scale Unified Lexical-Semantic Resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page (to appear), Avignon, France.

Yoshihiko Hayashi. 2011. A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences. In *Proceedings of the International Confer-*

ence on Computational Semantics (IWCS), pages 155–164, Oxford, UK.

Verena Henrich and Erhard Hinrichs. 2010. Standardizing wordnets in the ISO standard LMF: Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 456–464, Beijing, China.

Nancy Ide and James Pustejovsky. 2010. What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong.

Nancy Ide, Alessandro Lenci, and Nicoletta Calzolari. 2003. RDF instantiation of ISLE/MILE lexical entries. In *Proceedings of the ACL 2003 workshop on Linguistic annotation: getting the model right*, pages 30–37, Sapporo, Japan.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proc. of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages at the 16th Nordic Conf. of Computational Linguistics (NODALIDA)*, pages 27–30, Tartu, Estonia.

Aida Khemakhem, Imen Elleuch, Bilel Gargouri, and Abdelmajid Ben Hamadou. 2009. Towards an Automatic Conversion Approach of Editorial Arabic Dictionaries into LMF-ISO 24613 Standardized Model. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A Large-scale Classification of English Verbs. *Language Resources and Evaluation*, 42:21–40.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet – representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1485–1491, Las Palmas, Canary Islands, Spain.

Lung-Hao Lee, Shu-Kai Hsieh, and Chu-Ren Huang. 2009. CWN-LMF: Chinese WordNet in the lexical markup framework. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 123–130, Suntec, Singapore.

Isa Maks, Carole Tiberius, and Remco van Veenendaal. 2008. Standardising Bilingual Lexical Resources According to the Lexicon Markup Framework. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*, pages 1723–1727, Marrakech, Morocco.

Michael Matuschek and Iryna Gurevych. 2011. Where the journey is headed: Collaboratively constructed multilingual wiki-based resources. In SFB 538: Mehrsprachigkeit, editor, *Hamburger Arbeiten zur Mehrsprachigkeit*. Hamburg, Germany.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *The Semantic Web: Research and Applications*, volume 6643 of *Lecture Notes in Computer Science*, pages 245–259. Springer Berlin / Heidelberg.

Christian M. Meyer and Iryna Gurevych. 2011. What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–892, Chiang Mai, Thailand.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 216–225, Uppsala, Sweden.

Elisabeth Niemann and Iryna Gurevych. 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 205–214, Oxford, UK.

Muntsa Padró, Núria Bel, and Silvia Necsulescu. 2011. Towards the Automatic Merging of Lexical Resources: Automatic Mapping. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 296–301, Hissar, Bulgaria.

Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15, Pisa, Italy.

Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. 2008. A lexicon for biology and bioinformatics: the BOOTStrep experience. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2285–2292, Marrakech, Morocco.

Bora Savas, Yoshihiko Hayashi, Monica Monachini, Claudia Soria, and Nicoletta Calzolari. 2010. An LMF-based Web Service for Accessing WordNet-type Semantic Lexicons. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, pages 507–513, Valletta, Malta.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for Wordnet interoperability. In *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pages 139–146, Palo Alto, California, USA.

Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh, and Kiyoaki Shirai. 2009. Query Expansion using LMF-Compliant Lexical Resources. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 145–152, Suntec, Singapore.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the Italian WordNet: upgrading, standarising, extending. In *Proceedings of the 5th Global WordNet Conference*, Bombay, India.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 1646–1652, Marrakech, Morocco.