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**[‡]

† Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information

‡ Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt

http://www.ukp.tu-darmstadt.de

Abstract

We present UBY, a large-scale lexical-semantic resource combining a wide range of information from expert-constructed and collaboratively constructed resources for English and German. It currently contains nine resources in two languages: English WordNet, Wiktionary, Wikipedia, FrameNet and VerbNet, German Wikipedia, Wiktionary and GermaNet, and multilingual OmegaWiki modeled according to the LMF standard. For FrameNet, VerbNet and all collaboratively constructed resources, this is done for the first time. Our LMF model captures lexical information at a fine-grained level by employing a large number of Data Categories from ISOCat and is designed to be directly extensible by new languages and resources. All resources in UBY can be accessed with an easy to use publicly available API.

1 Introduction

Lexical-semantic resources (LSRs) are the foundation of many NLP tasks such as word sense disambiguation, semantic role labeling, question answering and information extraction. They are needed on a large scale in different languages. The growing demand for resources is met neither by the largest single expert-constructed resources (ECRs), such as WordNet and FrameNet, whose coverage is limited, nor by collaboratively constructed resources (CCRs), such as Wikipedia and Wiktionary, which encode lexical-semantic knowledge in a less systematic form than ECRs, because they are lacking expert supervision.

Previously, there have been several independent efforts of combining existing LSRs to enhance their coverage w.r.t. their breadth and depth, i.e. (i) the number of lexical items, and (ii) the types of lexical-semantic information contained (Shi and Mihalcea, 2005; Johansson and Nugues, 2007; Navigli and Ponzetto, 2010b; Meyer and Gurevych, 2011). As these efforts often targeted particular applications, they focused on aligning selected, specialized information types. To our knowledge, no single work focused on modeling a wide range of ECRs and CCRs in multiple languages and a large variety of information types in a standardized format. Frequently, the presented model is not easily scalable to accommodate an open set of LSRs in multiple languages and the information mined automatically from corpora. The previous work also lacked the aspects of lexicon format standardization and API access. We believe that easy access to information in LSRs is crucial in terms of their acceptance and broad applicability in NLP.

In this paper, we propose a solution to this. We define a standardized format for modeling LSRs. This is a prerequisite for resource interoperability and the smooth integration of resources. We employ the ISO standard Lexical Markup Framework (LMF: ISO 24613:2008), a metamodel for LSRs (Francopoulo et al., 2006), and Data Categories (DCs) selected from ISOCat.[1] One of the main challenges of our work is to develop a model that is standard-compliant, yet able to express the information contained in diverse LSRs, and that in the long term supports the integration of the various resources.

The main contributions of this paper can be

[1] http://www.isocat.org/

summarized as follows: (1) We present an LMF-based model for large-scale multilingual LSRs called UBY-LMF. We model the lexical-semantic information down to a fine-grained level of information (e.g. syntactic frames) and employ standardized definitions of linguistic information types from ISOCat. (2) We present UBY, a large-scale LSR implementing the UBY-LMF model. UBY currently contains nine resources in two languages: English WordNet (WN, Fellbaum (1998), Wiktionary[2] (WKT-en), Wikipedia[3] (WP-en), 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)), and the English and German entries of OmegaWiki[4] (OW), referred to as OW-en and OW-de. OW, a novel CCR, is inherently multilingual – its basic structure are multilingual synsets, which are a valuable addition to our multilingual UBY. Essential to UBY are the nine pairwise sense alignments between resources, which we provide to enable resource interoperability on the sense level, e.g. by providing access to the often complementary information for a sense in different resources. (3) We present a Java-API which offers unified access to the information contained in UBY.

We will make the UBY-LMF model, the resource UBY and the API freely available to the research community.[5] This will make it easy for the NLP community to utilize UBY in a variety of tasks in the future.

2 Related Work

The work presented in this paper concerns standardization of LSRs, large-scale integration thereof at the representational level, and the unified access to lexical-semantic information in the integrated resources.

Standardization of resources. Previous work includes models for representing lexical information relative to ontologies (Buitelaar et al., 2009; McCrae et al., 2011), and standardized single wordnets (English, German and Italian wordnets) in the ISO standard LMF (Soria et al., 2009; Henrich and Hinrichs, 2010; Toral et al., 2010).

McCrae et al. (2011) propose LEMON, a conceptual model for lexicalizing ontologies as an extension of the LexInfo model (Buitelaar et al., 2009). LEMON provides an LMF-implementation in the Web Ontology Language (OWL), which is similar to UBY-LMF, as it also uses DCs from ISOCat, but diverges further from the standard (e.g. by removing structural elements such as the predicative representation class). While we focus on modeling lexical-semantic information comprehensively and at a fine-grained level, the goal of LEMON is to support the linking between ontologies and lexicons. This goal entails a task-targeted application: domain-specific lexicons are extracted from ontology specifications and merged with existing LSRs on demand. As a consequence, there is no available large-scale instance of the LEMON model.

Soria et al. (2009) define WordNet-LMF, an LMF model for representing wordnets used in the KYOTO project, and Henrich and Hinrichs (2010) do this for GN, the German wordnet. These models are similar, but they still present different implementations of the LMF meta-model, which hampers interoperability between the resources. We build upon this work, but extend it significantly: UBY goes beyond modeling a single ECR and represents a large number of both ECRs and CCRs with very heterogeneous content in the same format. Also, UBY-LMF features deeper modeling of lexical-semantic information. Henrich and Hinrichs (2010), for instance, do not explicitly model the argument structure of subcategorization frames, since each frame is represented as a string. In UBY-LMF, we represent them at a fine-grained level necessary for the transparent modeling of the syntax-semantics interface.

Large-scale integration of resources. Most previous research efforts on the integration of resources targeted at world knowledge rather than lexical-semantic knowledge. Well known examples are YAGO (Suchanek et al., 2007), or DBPedia (Bizer et al., 2009).

Atserias et al. (2004) present the Meaning Multilingual Central Repository (MCR). MCR integrates five local wordnets based on the Interlingual Index of EuroWordNet (Vossen, 1998). The overall goal of the work is to improve word sense disambiguation. This work is similar to ours, as it

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

[3] http://www.wikipedia.org/

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

[5] http://www.ukp.tu-darmstadt.de/data/uby

aims at a large-scale multilingual resource and includes several resources. It is however restricted to a single type of resource (wordnets) and features a single type of lexical information (semantic relations) specified upon synsets. Similarly, de Melo and Weikum (2009) create a multilingual wordnet by integrating wordnets, bilingual dictionaries and information from parallel corpora. None of these resources integrate lexical-semantic information, such as syntactic subcategorization or semantic roles.

McFate and Forbus (2011) present NULEX, a syntactic lexicon automatically compiled from WN, WKT-en and VN. As their goal is to create an open-license resource to enhance syntactic parsing, they enrich verbs and nouns in WN with inflection information from WKT-en and syntactic frames from VN. Thus, they only use a small part of the lexical information present in WKT-en.

Padró et al. (2011) present their work on lexicon merging within the Panacea Project. One goal of Panacea is to create a lexical resource development platform that supports large-scale lexical acquisition and can be used to combine existing lexicons with automatically acquired ones. To this end, Padró et al. (2011) explore the automatic integration of subcategorization lexicons. Their current work only covers Spanish, and though they mention the LMF standard as a potential data model, they do not make use of it.

Shi and Mihalcea (2005) integrate FN, VN and WN, and Palmer (2009) presents a combination of Propbank, VN and FN in a resource called SEMLINK in order to enhance semantic role labeling. Similar to our work, multiple resources are integrated, but their work is restricted to a single language and does not cover CCRs, whose popularity and importance has grown tremendously over the past years. In fact, with the exception of NULEX, CCRs have only been considered in the sense alignment of individual resource pairs (Navigli and Ponzetto, 2010a; Meyer and Gurevych, 2011).

API access for resources. An important factor to the success of a large, integrated resource is a single public API, which facilitates the access to the information contained in the resource. The most important LSRs so far can be accessed using various APIs, for instance the Java WordNet API,[6] or the Java-based Wikipedia API.[7]

With a stronger focus of the NLP community on sharing data and reproducing experimental results these tools are becoming important as never before. Therefore, a major design objective of UBY is a single API. This is similar in spirit to the motivation of Pradhan et al. (2007), who present integrated access to corpus annotations as a main goal of their work on standardizing and integrating corpus annotations in the OntoNotes project.

To summarize, related work focuses either on the standardization of single resources (or a single type of resource), which leads to several slightly different formats constrained to these resources, or on the integration of several resources in an idiosyncratic format. CCRs have not been considered at all in previous work on resource standardization, and the level of detail of the modeling is insufficient to fully accommodate different types of lexical-semantic information. API access is rarely provided. This makes it hard for the community to exploit their results on a large scale. Thus, it diminishes the impact that these projects might achieve upon NLP beyond their original specific purpose, if their results were represented in a unified resource and could easily be accessed by the community through a single public API.

3 UBY – Data model

LMF defines a metamodel of LSRs in the Unified Modeling Language (UML). It provides a number of UML packages and classes for modeling many different types of resources, e.g. wordnets and multilingual lexicons. The design of a standard-compliant lexicon model in LMF involves two steps: in the first step, the structure of the lexicon model has to be defined by choosing a combination of the LMF core package and zero to many extensions (i.e. UML packages). In the second step, these UML classes are enriched by attributes. To contribute to semantic interoperability, it is essential for the lexicon model that the attributes and their values refer to Data Categories (DCs) taken from a reference repository. DCs are standardized specifications of the terms that are used for attributes and their values, or in other words, the linguistic vocabulary occurring

[6] http://sourceforge.net/projects/jwordnet/

[7] http://code.google.com/p/jwpl/

in a lexicon model. Consider, for instance, the term *lexeme* that is defined differently in WN and FN: in FN, a lexeme refers to a word form, not including the sense aspect. In WN, on the contrary, a lexeme is an abstract pairing of meaning and form. According to LMF, the DCs are to be selected from ISOCat, the implementation of the ISO 12620 Data Category Registry (DCR, Broeder et al. (2010)), resulting in a Data Category Selection (DCS).

Design of UBY-LMF. We have designed UBY-LMF[8] as a model of the union of various heterogeneous resources, namely WN, GN, FN, and VN on the one hand and CCRs on the other hand.

Two design principles guided our development of UBY-LMF: first, to preserve the information available in the original resources and to uniformly represent it in UBY-LMF. Second, to be able to extend UBY in the future by further languages, resources, and types of linguistic information, in particular, alignments between different LSRs.

Wordnets, FN and VN are largely complementary regarding the information types they provide, see, e.g. Baker and Fellbaum (2009). Accordingly, they use different organizational units to represent this information. Wordnets, such as WN and GN, primarily contain information on lexical-semantic relations, such as synonymy, and use synsets (groups of lexemes that are synonymous) as organizational units. FN focuses on groups of lexemes that evoke the same prototypical situation (so-called *semantic frames*, Fillmore (1982)) involving semantic roles (so-called *frame elements*). VN, a large-scale verb lexicon, is organized in Levin-style verb classes (Levin, 1993) (groups of verbs that share the same syntactic alternations and semantic roles) and provides rich subcategorization frames including semantic roles and a specification of semantic predicates.

UBY-LMF employs several direct subclasses of `Lexicon` in order to account for the various organization types found in the different LSRs considered. While the `LexicalEntry` class reflects the traditional headword-based lexicon organization, `Synset` represents synsets from wordnets, `SemanticPredicate` models FN semantic frames, and `SubcategorizationFrameSet` corresponds to VN alternation classes. `SubcategorizationFrame` is composed of syntactic arguments, while `SemanticPredicate` is composed of semantic arguments. The linking between syntactic and semantic arguments is represented by the `SynSemCorrespondence` class.

The `SenseAxis` class is very important in UBY-LMF, as it connects the different source LSRs. Its role is twofold: first, it links the corresponding word senses from different languages, e.g. English and German. Second, it represents monolingual sense alignments, i.e. sense alignments between different lexicons in the *same* language. The latter is a novel interpretation of `SenseAxis` introduced by UBY-LMF.

The organization of lexical-semantic knowledge found in WP, WKT, and OW can be modeled with the classes in UBY-LMF as well. WP primarily provides encyclopedic information on nouns. It mainly consists of article pages which are modeled as `Senses` in UBY-LMF.

WKT is in many ways similar to traditional dictionaries, because it enumerates senses under a given headword on an entry page. Thus, WKT entry pages can be represented by `LexicalEntries` and WKT senses by `Senses`.

OW is different from WKT and WP, as it is organized in multilingual synsets. To model OW in UBY-LMF, we split the synsets per language and included them as monolingual `Synsets` in the corresponding `Lexicon` (e.g., OW-en or OW-de). The original multilingual information is preserved by adding a `SenseAxis` between corresponding synsets in OW-en and OW-de.

The LMF standard itself contains only few linguistic terms and does neither specify attributes nor their values. Therefore, an important task in developing UBY-LMF has been the specification of attributes and their values along with the proper attachment of attributes to LMF classes. In particular, this task involved selecting DCs from ISOCat and, if necessary, adding new DCs to ISOCat.

Extensions in UBY-LMF. Although UBY-LMF is largely compliant with LMF, the task of building a homogeneous lexicon model for many highly heterogeneous LSRs led us to extend LMF in several ways: we added two new classes and several new relationships between classes.

First, we were facing a huge variety of lexical-semantic labels for many different dimensions of

[8] See www.ukp.tu-darmstadt.de/data/uby

semantic classification. Examples of such dimensions include ontological type (e.g. selectional restrictions in VN and FN), domain (e.g. Biology in WN), style and register (e.g. labels in WKT, OW), or sentiment (e.g. sentiment of lexical units in FN). Since we aim at an extensible LMF-model, capable of representing further dimensions of semantic classification, we did not squeeze the information on semantic classes present in the considered LSRs into existing LMF classes. Instead, we addressed this issue by introducing a more general class, `SemanticLabel`, which is an optional subclass of `Sense`, `SemanticPredicate`, and `SemanticArgument`. This new 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).

Second, we attached the subclass `Frequency` to most of the classes in UBY-LMF, in order to encode frequency information. This is of particular importance when using the resource in machine learning applications. This extension of the standard has already been made in WordNet-LMF (Soria et al., 2009). Currently, the `Frequency` class is used to keep corpus frequencies for lexical units in FN, but we plan to use it for enriching many other classes with frequency information in future work, such as `Senses` or `SubcategorizationFrames`.

Third, the representation of FN in LMF required adding two new relationships between LMF classes: we added a relationship between `SemanticArgument` and `Definition`, in order to represent the definitions available for frame elements in FN. In addition, we added a relationship between the `Context` class and the `MonoLingualExternalRef`, to represent the links to annotated corpus sentences in FN.

Finally, WKT turned out to be hard to tackle, because it 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 (Meyer and Gurevych, 2010). These ambiguous relation targets could not directly be represented in LMF, since sense and translation relations are defined between senses. To resolve this, we added a relationship between `SenseRelation` and `FormRepresentation`, in order to encode the ambiguous WKT relation target as a word form. Disambiguating the WKT relation targets to infer the target sense is left to future work.

A related issue occurred, when we mapped WN to LMF. WN encodes morphologically related forms as sense relations. UBY-LMF represents these related forms not only as sense relations (as in WordNet-LMF), but also at the morphological level using the `RelatedForm` class from the LMF Morphology extension. In LMF, however, the `RelatedForm` class for morphologically related lexemes is not associated with the corresponding sense in any way. Discarding the WN information on the senses involved in a particular morphological relation would lead to information loss in some cases. Consider as an example the WN verb *buy* (purchase) which 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*. We addressed this issue by adding a sense attribute to the `RelatedForm` class. Thus, in extension of LMF, UBY-LMF allows sense relations to refer to a form relation target and morphological relations to refer to a sense relation target.

Data Categories in UBY-LMF. We encountered large differences in the availability of DCs in ISOCat for the morpho-syntactic, lexical-syntactic, and lexical-semantic parts of UBY-LMF. Many DCs were missing in ISOCat and we had to enter them ourselves. While this was feasible at the morpho-syntactic and lexical-syntactic level, due to a large body of standardization results available, it was much harder at the lexical-semantic level where standardization is still ongoing. At the lexical-semantic level, UBY-LMF currently allows string values for a number of attribute values, e.g. for semantic roles. We can easily integrate the results of the ongoing standardization efforts into UBY-LMF in the future.

4 UBY – Population with information

4.1 Representing LSRs in UBY-LMF

UBY-LMF is represented by a DTD (as suggested by the standard) which can be used to automatically convert any given resource into the corresponding XML format.[9] This conversion requires a detailed analysis of the resource to be converted, followed by the definition of a mapping of the

[9] Therefore, UBY-LMF can be considered as a serialization of LMF.

concepts and terms used in the original resource to the UBY-LMF model. There are two major tasks involved in the development of an automatic conversion routine: first, the basic organizational unit in the source LSR has to be identified and mapped, e.g. synset in WN or semantic frame in FN, and second, it has to be determined, how a (LMF) sense is defined in the source LSR.

A notable aspect of converting resources into UBY-LMF is the harmonization of linguistic terminology used in the LSRs. For instance, a WN *Word* and a GN *Lexical Unit* are mapped to `Sense` in UBY-LMF.

We developed reusable conversion routines for the future import of updated versions of the source LSRs into UBY, provided the structure of the source LSR remains stable. These conversion routines extract lexical data from the source LSRs by calling their native APIs (rather than processing the underlying XML data). Thus, all lexical information which can be accessed via the APIs is converted into UBY-LMF.

Converting the LSRs introduced in the previous section yielded an instantiation of UBY-LMF named UBY. The `LexicalResource` instance UBY currently comprises 10 `Lexicon` instances, one each for OW-de and OW-en, and one lexicon each for the remaining eight LSRs.

4.2 Adding Sense Alignments

Besides the uniform and standardized representation of the single LSRs, one major asset of UBY is the semantic interoperability of resources at the sense level. In the following, we (i) describe how we converted already existing sense alignments of resources into LMF, and (ii) present a framework to infer alignments automatically for any pair of resources.

Existing Alignments. Previous work on sense alignment yielded several alignments, such as WN–WP-en (Niemann and Gurevych, 2011), WN–WKT-en (Meyer and Gurevych, 2011) and VN–FN (Palmer, 2009).

We converted these alignments into UBY-LMF by creating a `SenseAxis` instance for each pair of aligned senses. This involved mapping the sense IDs from the proprietary alignment files to the corresponding sense IDs in UBY.

In addition, we integrated the sense alignments already present in OW and WP. Some OW entries provide links to the corresponding WP page. Also, the German and English language editions of WP and OW are connected by inter-language links between articles (`Senses` in UBY). We can expect that these links have high quality, as they were entered manually by users and are subject to community control. Therefore, we straightforwardly imported them into UBY.

Alignment Framework. Automatically creating new alignments is difficult because of word ambiguities, different granularities of senses, or language specific conceptualizations (Navigli, 2006). To support this task for a large number of resources across languages, we have designed a flexible alignment framework based on the state-of-the-art method of Niemann and Gurevych (2011). The framework is generic in order to allow alignments between different kinds of entities as found in different resources, e.g. WN synsets, FN frames or WP articles. The only requirement is that the individual entities are distinguishable by a unique identifier in each resource.

The alignment consists of the following steps: First, we extract the alignment candidates for a given resource pair, e.g. WN sense candidates for a WKT-en entry. Second, we create a gold standard by manually annotating a subset of candidate pairs as “valid“ or “non-valid“. Then, we extract the sense representations (e.g. lemmatized bag-of-words based on glosses) to compute the similarity of word senses (e.g. by cosine similarity). The gold standard with corresponding similarity values is fed into Weka (Hall et al., 2009) to train a machine learning classifier, and in the final step this classifier is used to automatically classify the candidate sense pairs as (non-)valid alignment. Our framework also allows us to train on a combination of different similarity measures.

Using our framework, we were able to reproduce the results reported by Niemann and Gurevych (2011) and Meyer and Gurevych (2011) based on the publicly available evaluation datasets[10] and the configuration details reported in the corresponding papers.

Cross-Lingual Alignment. In order to align word senses across languages, we extended the monolingual sense alignment described above to the cross-lingual setting. Our approach utilizes

[10] http://www.ukp.tu-darmstadt.de/data/sense-alignment/

Moses,[11] trained on the Europarl corpus. The lemma of one of the two senses to be aligned as well as its representations (e.g. the gloss) is translated into the language of the other resource, yielding a monolingual setting. E.g., the WN synset {*vessel, watercraft*} with its gloss *'a craft designed for water transportation'* is translated into {*Schiff, Wasserfahrzeug*} and *'Ein Fahrzeug für Wassertransport'*, and then the candidate extraction and all downstream steps can take place in German. An inherent problem with this approach is that incorrect translations also lead to invalid alignment candidates. However, these are most probably filtered out by the machine learning classifier as the calculated similarity between the sense representations (e.g. glosses) should be low if the candidates do not match.

We evaluated our approach by creating a cross-lingual alignment between WN and OW-de, i.e. the concepts in OW with a German lexicalization.[12] To our knowledge, this is the first study on aligning OW with another LSR. OW is especially interesting for this task due to its multilingual concepts, as described by Matuschek and Gurevych (2011). The created gold standard could, for instance, be re-used to evaluate alignments for other languages in OW.

To compute the similarity of word senses, we followed the approach by Niemann and Gurevych (2011) while covering both translation directions. We used the cosine similarity for comparing the German OW glosses with the German translations of WN glosses and cosine and personalized page rank (PPR) similarity for comparison of the German OW glosses translated into English with the original English WN glosses. Note that PPR similarity is not available for German as it is based on WN. Thereby, we filtered out the OW concepts without a German gloss which left us with 11,806 unique candidate pairs. We randomly selected 500 WN synsets for analysis yielding 703 candidate pairs. These were manually annotated as being (non-)alignments. For the subsequent machine learning task we used a simple threshold-based classifier and ten-fold cross validation.

Table 1 summarizes the results of different system configurations. We observe that translation into English works significantly better than into German. Also, the more elaborate similarity measure PPR yields better results than cosine similarity, while the best result is achieved by a combination of both. Niemann and Gurevych (2011) make a similar observation for the monolingual setting. Our F-measure of 0.717 in the best configuration lies between the results of Meyer and Gurevych (2011) (0.66) and Niemann and Gurevych (2011) (0.78), and thus verifies the validity of the machine translation approach. Therefore, the best alignment was subsequently integrated into UBY.

Translation direction	Similarity measure	P	R	F_1
EN > DE	Cosine (Cos)	0.666	0.575	0.594
DE > EN	Cos	0.674	0.658	0.665
DE > EN	PPR	0.721	**0.712**	0.716
DE > EN	PPR + Cos	**0.723**	**0.712**	**0.717**

Table 1: Cross-lingual alignment results

5 Evaluating UBY

We performed an intrinsic evaluation of UBY by computing a number of resource statistics. Our evaluation covers two aspects: first, it addresses the question if our automatic conversion routines work correctly. Second, it provides indicators for assessing UBY in terms of the gain in coverage compared to the single LSRs.

Correctness of conversion. Since we aim to preserve the maximal amount of information from the original LSRs, we should be able to replace any of the original LSRs and APIs by UBY and the UBY-API without losing information. As the conversion is largely performed automatically, systematic errors and information loss could be introduced by a faulty conversion routine. In order to detect such errors and to prove the correctness of the automatic conversion and the resulting representation, we have compared the original resource statistics of the classes and information types in the source LSRs to the corresponding classes in their UBY counterparts. For instance, the number of lexical relations in WordNet has been compared to the number of `SenseRelations` in the UBY WordNet lexicon.[13]

[11] http://www.statmt.org/moses/

[12] OmegaWiki consists of interlinked language-independent concepts to which lexicalizations in several languages are attached.

[13] For detailed analysis results see the UBY website.

Lexicon	Lexical Entry	Sense	Sense Relation
FN	9,704	11,942	–
GN	83,091	93,407	329,213
OW-de	30,967	34,691	60,054
OW-en	51,715	57,921	85,952
WP-de	790,430	838,428	571,286
WP-en	2,712,117	2,921,455	3,364,083
WKT-de	85,575	72,752	434,358
WKT-en	335,749	421,848	716,595
WN	156,584	206,978	8,559
VN	3,962	31,891	–
UBY	**4,259,894**	**4,691,313**	**5,300,941**

Table 2: UBY resource statistics (selected classes).

Lexicon pair	Languages	SenseAxis
WN–WP-en	EN–EN	50,351
WN–WKT-en	EN–EN	99,662
WN–VN	EN–EN	40,716
FN–VN	EN–EN	17,529
WP-en–OW-en	EN–EN	3,960
WP-de–OW-de	DE–DE	1,097
WN–OW-de	EN–DE	23,024
WP-en–WP-de	EN–DE	463,311
OW-en–OW-de	EN–DE	58,785
UBY	**All**	**758,435**

Table 3: UBY alignment statistics.

Gain in coverage. UBY offers an increased coverage compared to the single LSRs as reflected in the resource statistics. Tables 2 and 3 show the statistics on central classes in UBY. As UBY is organized in several Lexicons, the number of UBY lexical entries is the sum of the lexical entries in all 10 Lexicons. Thus, UBY contains more than 4.2 million lexical entries, 4.6 million senses, 5.3 million semantic relations between senses and more than 750,000 alignments. These statistics represent the total numbers of lexical entries, senses and sense relations in UBY without filtering of identical (i.e. corresponding) lexical entries, senses and relations. Listing the number of unique senses would require a full alignment between all integrated resources, which is currently not available.

We can, however, show that UBY contains over 3.08 million unique lemma-POS combinations for English and over 860,000 for German, over 3.94 million in total, see Table 4. Therefore, we assessed the coverage on lemma level. Table 4 also shows the number of lemmas with entries in one or more than one lexicon, additionally split by POS and language. Lemmas occurring only once in UBY increase the coverage at lemma level. For lemmas with parallel entries in several UBY lexicons, new information becomes available in the form of additional sense definitions and complementary information types attached to lemmas.

Finally, the increase in coverage at sense level can be estimated for senses that are aligned across at least two UBY-lexicons. We gain access to all available, partly complementary information types attached to these aligned senses, e.g. semantic relations, subcategorization frames, encyclopedic or multilingual information. The number of pairwise sense alignments provided by UBY is given in Table 3. In addition, we computed how many senses simultaneously take part in at least two pairwise sense alignments. For English, this applies to 31,786 senses, for which information from 3 UBY lexicons is available.

EN Lexicons	noun	verb	adjective
5	1	699	-
4	1,630	1,888	430
3	8,439	1,948	2,271
2	53,856	4,727	12,290
1	2,900,652	50,209	41,731
Σ (unique EN)	3,080,771		
DE Lexicons	noun	verb	adjective
4	1,546	-	-
3	10,374	372	342
2	26,813	3,174	2,643
1	803,770	6,108	7,737
Σ (unique DE)	862,879		

Table 4: Number of lemmas (split by POS and language) with entries in i UBY lexicons, $i = 1, \ldots, 5$.

6 Using UBY

UBY API. For convenient access to UBY, we implemented a Java-API which is built around the Hibernate[14] framework. Hibernate allows to easily store the XML data which results from converting resources into Uby-LMF into a corresponding SQL database.

Our main design principle was to keep the access to the resource as simple as possible, despite the rich and complex structure of UBY. Another

[14] http://www.hibernate.org/

important design aspect was to ensure that the functionality of the individual, resource-specific APIs or user interfaces is mirrored in the UBY API. This enables porting legacy applications to our new resource. To facilitate the transition to UBY, we plan to provide reference tables which list the corresponding UBY-API operations for the most important operations in the WN API, some of which are shown in Table 5.

WN function	UBY function
Dictionary	**UBY**
getIndexWord(pos, lemma)	getLexicalEntries(pos, lemma)
IndexWord	**LexicalEntry**
getLemma()	getLemmaForm()
Synset	**Synset**
getGloss()	getDefinitionText()
getWords()	getSenses()
Pointer	**SynsetRelation**
getType()	getRelName()
Word	**Sense**
getPointers()	getSenseRelations()

Table 5: Some equivalent operations in WN API and UBY API.

While it is possible to limit access to single resources by a parameter and thus mimic the behavior of the legacy APIs (e.g. only retrieve Synsets and their relations from WN), the true power of UBY API becomes visible when no such constraints are applied. In this case, all imported resources are queried to get one combined result, while retaining the source of the respective information. On top of this, the information about existing sense alignments across resources can be accessed via `SenseAxis` relations, so that the returned combined result covers not only the lexical, but also the sense level.

Community issues. One of the most important reasons for UBY is creating an easy-to-use powerful LSR to advance NLP research and development. Therefore, community building around the resource is one of our major concerns. To this end, we will offer free downloads of the lexical data and software presented in this paper under open licenses, namely: The UBY-LMF DTD, mappings and conversion tools for existing resources and sense alignments, the Java API, and, as far as licensing allows,[15] already converted resources. If resources cannot be made available for download, the conversion tools will still allow users with access to these resources to import them into UBY easily. In this way, it will be possible for users to build their "custom UBY" containing selected resources. As the underlying resources are subject to continuous change, updates of the corresponding components will be made available on a regular basis.

7 Conclusions

We presented UBY, a large-scale, standardized LSR containing nine widely used resources in two languages: English WN, WKT-en, WP-en, FN and VN, German WP-de, WKT-de, and GN, and OW in English and German. As all resources are modeled in UBY-LMF, UBY enables structural interoperability across resources and languages down to a fine-grained level of information. For FN, VN and all of the CCRs in English and German, this is done for the first time. Besides, by integrating sense alignments we also enable the lexical-semantic interoperability of resources. We presented a unified framework for aligning any LSRs pairwise and reported on experiments which align OW-de and WN. We will release the UBY-LMF model, the resource and the UBY-API at the time of publication.[16] Due to the added value and the large scale of UBY, as well as its ease of use, we believe UBY will boost the performance of NLP making use of lexical-semantic knowledge.

Acknowledgments

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under grant No. GU 798/3-1 and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Richard Eckart de Castilho, Yevgen Chebotar, Zijad Maksuti and Tri Duc Nghiem for their contributions to this project.

References

Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek

[15] Only GermaNet is subject to a restricted license and cannot be redistributed in UBY format.

[16] http://www.ukp.tu-darmstadt.de/data/uby

Vossen. 2004. The Meaning Multilingual Central Repository. In *Proceedings of the second international WordNet Conference (GWC 2004)*, pages 23–30, Brno, Czech Republic.

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 '09, 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'98*, pages 86–90, Montreal, Canada.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, (7):154–165.

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 7th 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 Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl, editors, *The Semantic Web: Research and Applications*, pages 111–125, Berlin/Heidelberg, Germany. Springer.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, CIKM '09, pages 513–522, New York, NY, USA. ACM.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Charles J. Fillmore. 1982. Frame Semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul, Korea.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

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.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, pages 27–30, Tartu, Estonia.

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.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, IL, USA.

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, Germany.

Clifton J. McFate and Kenneth D. Forbus. 2011. NULEX: an open-license broad coverage lexicon. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 363–367, Portland, OR, USA.

Christian M. Meyer and Iryna Gurevych. 2010. Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 11th International Conference*, volume 6008 of *Lecture Notes in Computer Science*, pages 38–49. Berlin/Heidelberg: Springer, Iaşi, Romania.

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. 2010a. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July.

Roberto Navigli and Simone Paolo Ponzetto. 2010b. Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 105–112, Sydney, Australia.

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 (RANLP)*, pages 296–301, Hissar, Bulgaria.

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex-09)*, pages 9–15, Pisa, Italy.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, pages 517–526, Washington, DC, USA.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 100–111, Mexico City, Mexico.

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, CA, USA.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, Banff, Canada.

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 (GWC)*, Bombay, India.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, Netherlands.