Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability

Judith Eckle-Kohler‡ and **Iryna Gurevych**†‡
† 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

This paper describes Subcat-LMF, an ISO-LMF compliant lexicon representation format featuring a uniform representation of subcategorization frames (SCFs) for the two languages English and German. Subcat-LMF is able to represent SCFs at a very fine-grained level. We utilized Subcat-LMF to standardize lexicons with large-scale SCF information: the English VerbNet and two German lexicons, i.e., a subset of IMSlex and GermaNet verbs. To evaluate our LMF-model, we performed a cross-lingual comparison of SCF coverage and overlap for the standardized versions of the English and German lexicons. The Subcat-LMF DTD, the conversion tools and the standardized versions of VerbNet and IMSlex subset are publicly available.[1]

1 Introduction

Computational lexicons providing accurate *lexical-syntactic* information, such as subcategorization frames (SCFs) are vital for many NLP applications involving parsing and word sense disambiguation. In parsing, SCFs have been successfully used to improve the output of statistical parsers (Klenner (2007), Deoskar (2008), Sigogne et al. (2011)) which is particularly significant in high-precision domain-independent parsing. In word sense disambiguation, SCFs have been identified as important features for verb sense disambiguation (Brown et al., 2011), which is due to the correlation of verb senses and SCFs (Andrew et al., 2004).

SCFs specify syntactic arguments of verbs and other predicate-like lexemes, e.g. the verb *say* takes two arguments that can be realized, for instance, as noun phrase and *that*-clause as in *He says that the window is open.*

Although a number of freely available, large-scale and accurate SCF lexicons exist, e.g. COMLEX (Grishman et al., 1994), VerbNet (Kipper et al., 2008) for English, availability and limitations in size and coverage remain an inherent issue. This applies even more to languages other than English.

One particular approach to address this issue is the combination and integration of existing manually built SCF lexicons. Lexicon integration has widely been adopted for increasing the coverage of lexicons regarding *lexical-semantic information types*, such as semantic roles, selectional restrictions, and word senses (e.g., Shi and Mihalcea (2005), the Semlink project[2], Navigli and Ponzetto (2010), Niemann and Gurevych (2011), Meyer and Gurevych (2011)).

Currently, SCFs are represented idiosyncratically in existing SCF lexicons. However, integration of SCFs requires a common, interoperable representation format. Monolingual SCF integration based on a common representation format has already been addressed by King and Crouch (2005) and just recently by Necsulescu et al. (2011) and Padró et al. (2011). However, neither King and Crouch (2005) nor Necsulescu et al. (2011) or Padró et al. (2011) make use of existing standards in order to create a uniform SCF representation for lexicon merging. The definition of an interoperable representation format according to an existing standard, such as the ISO standard Lexical Markup Framework (LMF, ISO 24613:2008, see Francopoulo et al. (2006)), is the

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

[2] http://verbs.colorado.edu/semlink/

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 550–560,
Avignon, France, April 23 - 27 2012.

prerequisite for re-using this format in different contexts, thus contributing to the standardization and interoperability of language resources.

While LMF models exist that cover the representation of SCFs (see Quochi et al. (2008), Buitelaar et al. (2009)), their suitability for representing SCFs at a large scale remains unclear: neither of these LMF-models has been used for standardizing lexicons with a large number of SCFs, such as VerbNet. Furthermore, the question of their applicability to different languages has not been investigated yet, a situation that is complicated by the fact that SCFs are highly language-specific.

The goal of this paper is to address these gaps for the two languages English and German by presenting a uniform LMF representation of SCFs for English and German which is utilized for the standardization of large-scale English and German SCF lexicons. The contributions of this paper are threefold: (1) We present the LMF model Subcat-LMF, an LMF-compliant lexicon representation format featuring a uniform and very fine-grained representation of SCFs for English and German. Subcat-LMF is a subset of Uby-LMF (Eckle-Kohler et al., 2012), the LMF model of the large integrated lexical resource Uby (Gurevych et al., 2012). (2) We convert lexicons with large-scale SCF information to Subcat-LMF: the English VerbNet and two German lexicons, i.e., GermaNet (Kunze and Lemnitzer, 2002) and a subset of IMSlex[3] (Eckle-Kohler, 1999). (3) We perform a comparison of these three lexicons regarding SCF coverage and SCF overlap, based on the standardized representation.

The remainder of this paper is structured as follows: Section 2 gives a detailed description of Subcat-LMF and section 3 demonstrates its usefulness for representing and cross-lingually comparing large-scale English and German lexicons. Section 4 provides a discussion including related work and section 5 concludes.

2 Subcat-LMF

2.1 ISO-LMF: a meta-model

LMF defines a *meta-model* of lexical resources, covering NLP lexicons and Machine Readable Dictionaries. This meta-model is based on the Unified Modeling Language (UML) and specifies a core package and a number of extensions for modeling different types of lexicons, including subcategorization lexicons.

The development of an LMF-compliant lexicon model requires 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). While the LMF core package models a lexicon in terms of lexical entries, each of which is defined as the pairing of one to many forms and zero to many senses, the LMF extensions provide UML classes for different types of lexicon organization, e.g., covering the synset-based organization of WordNet and the class-based organization of VerbNet. The first step results in a set of UML classes that are associated according to the UML diagrams given in ISO LMF.

In the second step, these UML classes may be enriched by attributes. While neither attributes nor their values are given by the standard, the standard states that both are to be linked to Data Categories (DCs) defined in a Data Category Registry (DCR) such as ISOCat.[4] DCs that are not available in ISOCat may be defined and submitted for standardization. The second step results in a so-called Data Category Selection (DCS).

DCs specify the linguistic vocabulary used in an LMF model. Consider as an example the linguistic term *direct object* that often occurs in SCFs of verbs taking an accusative NP as argument. In ISOCat, there are two different specifications of this term, one explicitly referring to the capability of becoming the clause subject in passivization[5], the other not mentioning passivization at all.[6] Consequently, the use of a DCR plays a major role regarding the *semantic interoperability* of lexicons (Ide and Pustejovsky, 2010). Different resources that share a common definition of their linguistic vocabulary are said to be *semantically interoperable*.

2.2 Fleshing out ISO-LMF

Approach: We started our development of Subcat-LMF with a thorough inspection of large-scale English and German resources providing SCFs for verbs, nouns, and adjectives. For

[3] http://www.ims.uni-stuttgart.de/projekte/IMSLex/

[4] http://www.isocat.org/, the implementation of the ISO 12620 DCR (Broeder et al., 2010).

[5] http://www.isocat.org/datcat/DC-1274

[6] http://www.isocat.org/datcat/DC-2263

English, our analysis included VerbNet[7] and FrameNet syntactically annotated example sentences from Ruppenhofer et al. (2010). For German, we inspected GermaNet, SALSA annotation guidelines (Burchardt et al., 2006) and IMSlex documentation (Eckle-Kohler, 1999). In addition, the EAGLES synopsis on morphosyntactic phenomena[8] (Calzolari and Monachini, 1996), as well as the EAGLES recommendations on subcategorization[9] have been used to identify DCs relevant for SCFs.

We specified Subcat-LMF by a DTD yielding an XML serialization of ISO-LMF. Thus, existing lexicons can be standardized, i.e. converted into Subcat-LMF format, based on the DTD.[10]

Lexicon structure: Next, we defined the lexicon structure of Subcat-LMF. In addition to the core package, Subcat-LMF primarily makes use of the LMF Syntax and Semantics extension. Figure 1 shows the most important classes of Subcat-LMF including `SynsemCorrespondence` where the linking of syntactic and semantic arguments is encoded. It might by worth noting that both synsets from GermaNet and verb classes from VerbNet can be represented in Subcat-LMF by using the `Synset` and `SubcategorizationFrameSet` class.

Diverging linguistic properties of SCFs in English and German: For verbs (and also for predicate-like nouns and adjectives), SCFs specify the syntactic and morphosyntactic properties of their arguments that have to be present in concrete realizations of these arguments within a sentence. While some properties of syntactic arguments in English and German correspond (both English and German are Germanic languages and hence closely related), there are other properties, mainly morphosyntactic ones that diverge. By way of examples, we illustrate some of these divergences in the following (we contrast English examples with their German equivalents):

- overt case marking in German:
  He helps him. vs. *Er hilft ihm.* (dative)
- specific verb form in verb phrase arguments:
  He suggested cleaning the house. (ing-form)
  vs.
  Er schlug vor, das Haus zu putzen. (to-infinitive)
- morphosyntactic marking of verb phrase arguments in the main clause: *He managed to win.* (no marking) vs.
  Er hat es geschafft zu gewinnen. (obligatory *es*)
- morphosyntactic marking of clausal arguments in the main clause: *That depends on who did it.* (preposition) vs.
  Das hängt davon ab, wer es getan hat. (pronominal adverb)

Uniform Data Categories for English and German: Thus, the main challenge in developing Subcat-LMF has been the specification of DCs (attributes and attribute values) *in such a way*, that a uniform specification of SCFs in the two languages English and German can be achieved. The specification of DCs for Subcat-LMF involved fleshing out ISO-LMF, because it is a meta-standard in the sense that it provides only few linguistic terms, i.e. DCs, and these DCs are not linked to any DCR: in the Syntax Extension, the standard only provides 7 class names, see Figure 1), complemented by 17 example attributes given in an informative, non-binding Annex F. These are by far not sufficient to represent the fine-grained SCFs available in such large-scale lexicons as VerbNet.

In contrast, the Syntax part of Subcat-LMF comprises 58 DCs that are properly linked to ISOCat DCs; a number of DCs were missing in ISOCat, so we entered them ourselves.[11] The majority of the attributes in Subcat-LMF are attached to the `SyntacticArgument` class. The corresponding DCs can be divided into two main groups:

Cross-lingually valid DCs for the specification of grammatical functions (e.g. `subject`, `prepositionalComplement`) and syntactic categories (e.g. `nounPhrase`, `prepositionalPhrase`), see Table 1.

Partly language-specific morphosyntactic DCs that further specify the syntactic arguments (e.g. attribute `case`, attribute `verbForm` and

[7] SCFs in VerbNet also cover SCFs in VALEX, a lexicon automatically extracted from corpora.

[8] http://www.ilc.cnr.it/EAGLES96/morphsyn/

[9] http://www.ilc.cnr.it/EAGLES96/synlex/

[10] Available at http://www.ukp.tu-darmstadt.de/data/uby

[11] The Subcat-LMF DCS is publicly available on the ISOCat website.

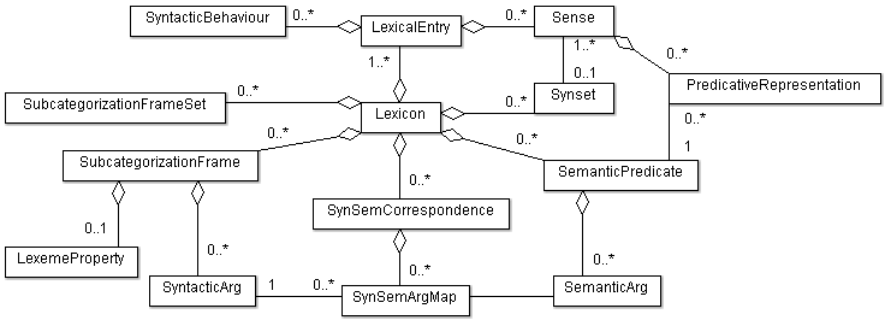

Figure 1: Selected classes of Subcat-LMF.

Values of `grammaticalFunction`	Example
subject	*They arrived in time.*
subjectComplement	*He becomes a teacher.*
directObject	*He saw a rainbow.*
objectComplement	*They elected him governor.*
complement	*He told him a story.*
prepositionalComplement	*It depends on several factors.*
adverbialComplement	*They moved far away.*
Values of `syntacticCategory`	**Example**
nounPhrase	*The train stopped.*
reflexive	*He drank himself sick.*
expletive	*It is raining.*
prepositionalPhrase	*It depends on several factors.*
adverbPhrase	*They moved far away.*
adjectivePhrase	*The light turned red.*
verbPhrase	*She tried to exercise.*
declarativeClause	*He says he agrees.*
subordinateClause	*He believes that it works.*

Table 1: Cross-lingually valid (English-German) attributes and values of the `SyntacticArgument` class.

values `toInfinitive`, `bareInfinitive`, `ingForm`, `participle`), see Table 2.

In the class `LexemeProperty`, we introduced an attribute `syntacticProperty` to encode control and raising properties of verbs taking infinitival verb phrase arguments.[12]

In Subcat-LMF, syntactic arguments can be specified by a selection of appropriate attribute-value pairs. While all syntactic arguments are uniformly specified by a grammatical function and a syntactic category, the use of the morphosyntactic attributes depends on the particular type of syntactic argument. Different phrase types are specified by different subsets of morphosyntactic attributes, see Table 2. The following examples illustrate some of these attributes:

- `number`: the number of a noun phrase argument can be lexically governed by the verb as in *These types of fish mix well together.*
- `verbForm`: the verb form of a clausal complement can be required to be a bare infinitive as in *They demanded that he be there.*
- `tense`: not only the verb form, but also the tense of a verb phrase complement can be lexically governed, e.g., to be a participle in the past tense as in *They had it removed.*

[12]Control or raising specify the co-reference between the implicit subject of the infinitival argument and syntactic arguments in the main clause, either the subject (subject control or raising) or direct object (object control or raising).

Morphosyntactic attributes and values	NP	PP	VP	C
case: nominative, genitive, dative, accusative	x	x		
determiner: possessive, indefinite	x		x	
number: singular, plural	x			
verbForm: toInfinitive, bareInfinitive, ingForm(!), Participle			x	x
tense: present, past			x	
complementizer: thatType, whType, yesNoType				x
prepositionType: external ontological type, e.g. locative		x	x	x
preposition: (string) (!)		x	x	x
lexeme: (string) (!)	x			x

Table 2: Morphosyntactic attributes of `SyntacticArgument` and phrase types for which the attributes are appropriate (NP: noun phrase, PP: prepositional phrase, VP: verb phrase, C: clause). Language-specific attributes are marked by (!).

3 Utilizing Subcat-LMF

3.1 Standardizing large-scale lexicons

Lexicon Data: We converted VerbNet (VN) and two German lexicons, i.e., GermaNet (GN) and a subset of IMSlex (ILS) to Subcat-LMF format. ILS has been developed independently from GN and the lexicon data were published in Eckle-Kohler (1999).

VN is organized in verb classes based on Levin-style syntactic alternations (Levin, 1993): verbs with common SCFs and syntactic alternation behavior that also share common semantic roles are grouped into classes. VN (version 3.1) lists 568 frames that are encoded as phrase structure rules (XML element `SYNTAX`), specifying phrase types and semantic roles of the arguments, as well as selectional, syntactic and morphosyntactic restrictions on the arguments. Additionally, a descriptive specification of each frame is given (XML element `DESCRIPTION`). The verb *learn*, for instance, has the following VN frame:

```
DESCRIPTION (primary):  NP V NP
SYNTAX: Agent V Topic
```

We extracted both the descriptive specifications and the phrase structure rules, using the API available for VN[13], resulting in 682 unique VN frames.[14]

GN provides detailed SCFs for verbs, in contrast to the Princeton WordNet: GN version 6.0 from April 2011 accessed by the GN API[15] lists 202 frames. GN SCFs are represented as a dot-separated sequence of letter pairs. Each letter pair specifies a syntactic argument: the first letter encodes the grammatical function and the second letter the syntactic category.[16] For instance, the following shows the GN code for transitive verbs: `NN.AN`.

ILS is represented in delimiter-separated values format and contains 784 verbs in total. Of these 784 verbs, 740 of them are also present in GN, and 44 are listed in ILS only. Although ILS contains only verbs that take clausal arguments and verb phrase arguments, a total number of 220 SCFs is present in ILS, also including SCFs without clausal and verb phrase arguments. ILS lists for each verb lemma a number of SCFs, thus specifying coarse-grained verb senses given by a lemma-SCF pair.[17] The SCFs are represented as parenthesized lists. For instance, the ILS SCF for transitive verbs is: `(subj(NPnom),obj(NPacc))`.

Automatic Conversion: We implemented Java tools for the conversion of VN, GN and ILS to Subcat-LMF. These tools convert the source lexicons based on a manual mapping of lexicon units and terms (e.g., VN verb class, GN synset) to Subcat-LMF. For the majority of SCFs, this mapping is defined on argument level. Lexical data is extracted from the source lexicons by using the native APIs (VN, GN) and additional Perl scripts.

[13] http://verbs.colorado.edu/verb-index/inspector/

[14] The VN API was used with the view options wrexyzsq for verb frame pairs and ctuqw for verb class information.

[15] GermaNet Java API 2.0.2

[16] See http://www.sfs.uni-tuebingen.de/GermaNet/-verb_frames.shtml

[17] In addition, ILS provides a semantic class label for each verb; however, these semantic labels are attached at lemma level, i.e. they need to be disambiguated.

	# LexicalEntry	# Sense	# Subcat.Frame	# SemanticPred.
LMF-VN orig. VN	3962 (3962 verbs)	31891 (31891 groups of verb, frame, sem.pred.)	284 (568 frames)	617 (572 sem. Pred.)
LMF-GN orig. GN	8626 (8626 verbs)	12981 (12981 verb-synset pairs)	147 (202 GN frames)	84 (no sem. Pred.)
LMF-ILS orig. ILS	784 (784 verbs)	3675 (3675 verb-frame pairs)	217 (220 SCFs)	10 (no sem. Pred.)

Table 3: Evaluation of the automatic conversion. Numbers of Subcat-LMF instances in the converted lexicons compared to numbers of corresponding units in original lexicons.

Evaluation of Automatic Conversion: Table 3 shows the mapping of the major source lexicon units (such as verb-synset pairs) to Subcat-LMF and lists the corresponding numbers of units.

For VN, groups of VN verb, frame and semantic predicate have been mapped to LMF senses. VN classes have been mapped to `SubcategorizationFrameSet`. Thus, the original VN-sense, a pairing of verb lemma and class, can be recovered by grouping LMF senses that share the same verb class. There is a significant difference between the original VN frames and their Subcat-LMF representation: the semantic information present in VN frames (semantic roles and selectional restrictions) is mapped to semantic arguments in Subcat-LMF, i.e. the mapping splits VN frames into a purely syntactic and a purely semantic part. Consequently, the number of unique SCFs in the Subcat-LMF version of VN is much smaller than the number of frames in the original VN. The conversion tool creates for each sense (specifying a unique verb, frame, semantic predicate combination) a `SynSemCorrespondence`.
On the other hand, the Subcat-LMF version of VN contains more semantic predicates than VN. This is due to selectional restrictions for semantic arguments that are specified in Subcat-LMF within semantic predicates, in contrast to VN.

For GN, verb-synset pairs (i.e., GN lexical units), have been mapped to LMF senses. Few GN frame codes also specify semantic role information, e.g. manner, location. These were mapped to the semantics part of Subcat-LMF resulting in 84 semantic predicates that encode the semantic role information in their semantic arguments.

ILS specifies similar semantic role information as GN; these few cases were mapped in the same way as for GN. Therefore, the LMF version of ILS, too, specifies less SCFs, but additional semantic predicates not present in the original.

Discussion: Grammatical functions of arguments are specified distinctly in the three lexicons. While both GN and ILS specify grammatical functions, they are not explicitly encoded in VN. They have to be inferred on the basis of the phrase structure rules given in the `SYNTAX` element. We assigned `subject` to the noun phrase which directly precedes the verb and `directObject` to the noun phrase directly following the verb *and* having the semantic role Patient. The semantic role information has to be considered at this point, because not all noun phrase arguments are able to become the subject in a corresponding passive sentence. An example is the verb *learn* which has the VN frame `NP(Agent) V NP(Topic)`; here, the Topic-NP is not able to become the subject of a corresponding passive sentence. We assigned the grammatical function `complement` to all other phrase types.

Argument order constraints in SCFs are represented in LMF by a list implementation of syntactic arguments. Most SCFs from VN require the subject to be the first argument, reflecting the basic word order in English sentences. VN lists one exception to this rule for the verb *appear*, illustrated by the example *On the horizon appears a ship.*

Argument optionality in VN is expressed at the semantic level and at the syntactic level in parallel: it is explicitly specified at the semantic level and implicitly specified at the syntactic level. At the syntactic level, two SCF versions exist in VN, one with the optional argument, the other without it. In addition, the semantic predicate attached to

these SCFs marks optional (semantic) arguments by a ?-sign. GN, on the other hand, expresses argument optionality at the level of syntactic arguments, i.e., within the frame code. In Subcat-LMF, optionality is represented at the syntactic level by an (optional) attribute `optional` for syntactic arguments, thus reflecting the explicit representation used in GN and the implicit representation present in VN.[18]

GN frames specify syntactic alternations of argument realizations, e.g. adverbial complements that can alternatively be realized as adverb phrase, prepositional phrase or noun phrase. We encoded this generalization in Subcat-LMF by introducing attribute values for these aggregated syntactic categories.

3.2 Cross-lingual comparison of lexicons

Lexicons that are standardized according to Subcat-LMF can be quantitatively compared regarding SCFs. For two lexicons, such a comparison gives answers to questions, such as: how many SCFs are present in both lexicons (overlapping SCFs), how many SCFs are only listed in one of the lexicons (complementary SCFs). Answers to these questions are important, for instance, for assessing the potential gain in SCF coverage that can be achieved by lexicon merging.

In order to validate our claim that Subcat-LMF yields a cross-lingually uniform SCF representation, we contrast the monolingual comparison of GN and ILS with the cross-lingual comparison of VN, GN and VN and ILS. Assuming that our claim is valid, the cross-lingual comparisons can be expected to yield similar results regarding overlapping and complementary SCFs as the monolingual comparison.

Comparison: The comparison of SCFs from two lexicons that are in Subcat-LMF format can be performed on the basis of the uniform DCs. As Subcat-LMF is implemented in XML, we compared string representations of SCFs. SCFs from VN, GN and ILS were converted to strings by concatenating attribute values of syntactic arguments and `lexemeProperty`. We created string representations of different granularities: First, fine-grained, language-specific string SCFs have been generated by concatenating all attribute values apart from the attribute `optional` which is specific to GN (resulting in a considerably smaller number of SCFs in GN). Second, fine-grained, but cross-lingual string SCFs were considered; these omit the attributes `case`, `lexeme`, `preposition` and the attribute value `ingForm`. Finally, coarse-grained cross-lingual string SCFs were compared. These only contain the values of the attributes `syntactic category`, `complementizer` and `verbForm` (without the attribute value `ingForm`). For instance, a coarse cross-lingual string SCF for transitive verbs is `nounPhrasenounPhrase`.

Table 4 lists the results of our quantitative comparison. For each lexicon pair, the number of overlapping SCFs and the numbers of complementary SCFs are given. Regarding VN and the German lexicons, the overlap at the language-specific level is (close to) zero, which is due to the specification of case, e.g. dative, for German arguments. However, the numbers for cross-lingual SCFs clearly validate our claim: the numbers of overlapping SCFs for the German lexicon pair and for the two German-English pairs are comparable, ranging from 12 to 18 for the fine-grained SCFs and from 20 to 21 for the coarse SCFs.

Based on the sets of cross-lingually overlapping SCFs, we made an estimation on how many high frequent verbs actually have SCFs that are in the cross-lingual SCF overlap of an English-German lexicon pair. For this, we used the lemma frequency lists of the English and German WaCky corpora (Baroni et al., 2009) and extracted verbs from VN, GN and ILS that are on 100 top ranked positions of these lists, starting from rank 100.[19] Table 5 shows the results for the cross-lingual SCF overlap between VN – GN and between VN – ILS. While only around 40% of the high frequent verbs have an SCF in the fine-grained SCF overlap, more than 70% are in the coarse overlap between VN – GN, and even more than 80% in the coarse overlap between VN – ILS.

Analysis of results: The small numbers of overlapping *cross-lingual* SCFs (relative to the total number of SCFs), at both levels of granularity, indicate that the three lexicons each encode substantially different lexical-syntactic properties of

[18] As a consequence, all semantic arguments specified in the Subcat-LMF version of VN have a corresponding syntactic argument.

[19] Since the WaCky frequency lists do not contain POS information, our lists of extracted verbs contain some noise, which we tolerated, because we aimed at an approximate estimate.

	language-specific (fine-grained)	cross-lingual (fine-grained)	cross-lingual (coarse)
GN vs. ILS	72 GN 21 both, 196 ILS	61 GN, 23 both, 69 ILS	40 GN, 24 both, 23 ILS
VN vs. GN	284 VN, 0 both, 93 GN	96 VN, **15** both, 69 GN	29 VN, **24** both, 40 GN
VN vs. ILS	283 VN, 1 both, 216 ILS	93 VN, **18** both, 74 ILS	31 VN, **22** both, 25 ILS

Table 4: Comparison of lexicon pairs regarding SCF overlap and complementary SCFs.

VN-GN overlap fine-grained (15 SCFs)	VN-GN overlap coarse (24 SCFs)	VN-ILS overlap fine-grained (18 SCFs)	VN-ILS overlap coarse (22 SCFs)
43% VN verbs	85% VN verbs	41% VN verbs	84% VN verbs
41% GN verbs	71% GN verbs	43% ILS verbs	87% ILS verbs

Table 5: Percentage of 100 high frequent verbs from VN, GN, ILS with a SCF in the cross-lingual SCF overlap (fine-grained vs. coarse) between VN – GN and VN – ILS.

verbs. This can at least partly be explained by the historic development of these lexicons in different contexts, e.g., Levin's work on verb classes (VN), Lexical Functional Grammar (ILS), as well as their use for different purposes and applications.

Another reason of the small SCF overlap is the comparison of strings derived from the XML format. A more sophisticated representation format, notably one that provides semantic typing and type hierarchies, e.g., OWL, could be employed to define hierarchies of grammatical functions (e.g. direct object would be a sub-type of complement) and other attributes. These would presumably support the identification of further overlapping SCFs.

During a subsequent qualitative analysis of the overlapping and complementary SCFs, we collected some enlightening background information. Overlapping SCFs in the cross-lingual comparison (both fine-grained and coarse) include prominent SCFs corresponding to transitive and intransitive verbs, as well as verbs with that-clause and verbs with to-infinitive.

GN and ILS are highly complementary regarding SCFs: for instance, while many SCFs with adverbial arguments are unique in GN, only ILS provides a fine-grained specification of prepositional complements including the preposition, as well as the case the preposition requires.[20] VN, too, contains a large number of SCFs with a detailed specification of possible prepositions, partly specified as language-independent preposition types. A large number of complementary SCFs in VN vs. GN and GN vs. ILS are due to a diverging linguistic analysis of extraposed subject clauses with an *es* (*it*) in the main clause (e.g., *It annoys him that the train is late.*). In GN, such clauses are not specified as subject, whereas in VN and ILS they are.

Regarding VN and ILS, only VN lists subject control for verbs, while both VN and ILS list object control and subject raising. GN, on the other hand, does not specify control or raising at all.

4 Discussion

4.1 Previous Work

Merging SCFs: Previous work on merging SCF lexicons has only been performed in a monolingual setting and lacks the use of standards. King and Crouch (2005) describe the process of unifying several large-scale verb lexicons for English, including VN and WordNet. They perform a conversion of these lexicons into a uniform, but non-standard representation format, resulting in a lexicon which is integrated at the level of verb senses, SCFs and lexical-semantics. Thus, the result of their work is not applicable to cross-lingual settings.

Necsulescu et al. (2011) and Padró et al. (2011) report on approaches to automatic merging of two Spanish SCF lexicons. As these lexicons lack sense information apart from the SCFs, their merging approach only works on a very coarse-grained sense level given by lemma-SCF pairs. The fully automatic merging approach described

[20] In German, prepositions govern the case of their noun phrase.

in (Padró et al., 2011) assumes that one of the lexicons to be integrated is already represented in the target representation format, i.e. given two lexicons, they map one lexicon to the format of the other. Moreover, their approach requires a significant overlap of SCFs and verbs in any two lexicons to be merged. The authors state that it is presently unclear, how much overlap is required to obtain sufficiently precise merging results.

Standardizing SCFs: Much previous work on standardizing NLP lexicons in LMF has focused on WordNet-like resources. Soria et al. (2009) describe WordNet-LMF, an LMF model for representing wordnets which has been used in the KYOTO project.[21] Later, WordNet-LMF has been adapted by Henrich and Hinrichs (2010) to GermaNet and by Toral et al. (2010) to the Italian WordNet. WordNet-LMF does not provide the possibility to represent subcategorization at all. The adaption of WordNet-LMF to GN (Henrich and Hinrichs, 2010) allows SCFs to be respresented as string values. However, this extension is not sufficient, because it provides no means to model the syntax-semantics interface, which specifies correspondences between syntactic and semantic arguments of verbs and other predicates. Quochi et al. (2008) report on an LMF model that covers the syntax-semantics mapping just mentioned; it has been used for standardizing an Italian domain-specific lexicon. Buitelaar et al. (2009) describe LexInfo, an LMF-model that is used for lexicalizing ontologies. LexInfo is implemented in OWL and specifies a linking of syntactic and semantic arguments. For SCFs and arguments, a type hierarchy is defined. In their paper, Buitelaar et al. (2009) show only few SCFs and do not indicate what kinds of SCFs can be represented with LexInfo in principle. On the LexInfo website[22], the current LexInfo version 2.0 can be viewed, but no further documentation is given. We inspected LexInfo version 2.0 and found that it specifies a large number of fine-grained SCFs. However, LexInfo has not been evaluated so far on large-scale SCF lexicons, such as VerbNet.

4.2 Subcat-LMF

Subcat-LMF enables the uniform representation of fine-grained SCFs across the two languages English and German. By mapping large-scale SCF lexicons to Subcat-LMF, we have demonstrated its usability for uniformly representing a wide range of SCFs and other lexical-syntactic information types in English and German.

As our cross-lingual comparison of lexicons has revealed many complementary SCFs in VN, GN and ILS, mono- and cross-lingual alignments of these lexicons at sense level would lead to a major increase in SCF coverage. Moreover, the cross-lingually uniform representation of SCFs can be exploited for an additional alignment of the lexicons at the level of SCF arguments. Such a fine-grained alignment of SCFs can be used, for instance, to project VN semantic roles to GN, thus yielding a German resource for semantic role labeling (see Gildea and Jurafsky (2002), Swier and Stevenson (2005)).

Subcat-LMF could be used for standardizing further English and German lexicons. The automatic conversion of lexicons to Subcat-LMF requires the manual definition of a mapping, at least for syntactic arguments. Furthermore, the automatic merging approach by Padró et al. (2011) could be tested for English: given our standardized version of VN, other English SCF lexicons could be merged fully automatically with the Subcat-LMF version of VN.

5 Conclusion

Subcat-LMF contributes to fostering the standardization of language resources and their interoperability at the lexical-syntactic level across English and German. The Subcat-LMF DTD including links to ISOCat, all conversion tools, and the standardized versions of VN and ILS[23] are publicly available at http://www.ukp.tu-darmstadt.de/data/uby.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank the anonymous reviewers for their valuable comments. We also thank Dr. Jungi Kim and Christian M. Meyer for their contributions to this paper, and Yevgen Chebotar and Zijad Maksuti for their contributions to the conversion software.

[21] http://www.kyoto-project.eu/

[22] See http://lexinfo.net/

[23] The converted version of GN can not be made available due to licensing.

References

Galen Andrew, Trond Grenager, and Christopher D. Manning. 2004. Verb sense and subcategorization: using joint inference to improve performance on complementary tasks. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 150–157, Barcelona, Spain.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

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.

Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. 2011. VerbNet Class Assignment as a WSD Task. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 85–94, Oxford, UK.

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. Springer-Verlag.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 969–974, Genoa, Italy.

Nicoletta Calzolari and Monica Monachini. 1996. EAGLES Proposal for Morphosyntactic Standards: in view of a ready-to-use package. In G. Perissinotto, editor, *Research in Humanities Computing*, volume 5, pages 48–64. Oxford University Press, Oxford, UK.

Tejaswini Deoskar. 2008. Re-estimation of lexical parameters for treebank PCFGs. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 193–200, Manchester, United Kingdom.

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

Judith Eckle-Kohler. 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*. Logos-Verlag, Berlin, Germany. PhDThesis.

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.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288, September.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex Syntax: Building a Computational Lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 268–272, Kyoto, Japan.

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 2012)*, page (to appear), Avignon, France.

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.

Tracy Holloway King and Dick Crouch. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbruecken, Germany.

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

Manfred Klenner. 2007. Shallow dependency labeling. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Companion Volume Proceedings of the Demo and Poster Sessions*, pages 201–204, Prague, Czech Republic.

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, USA.

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.

Silvia Necsulescu, Núria Bel, Munsta Padró, Montserrat Marimon, and Eva Revilla. 2011. Towards the Automatic Merging of Language Resources. In *Proceedings of the 2011 ESSLI Workshop on Lexical Resources (WoLeR 2011)*, Ljubljana, Slovenia.

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.

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'08)*, pages 2285–2292, Marrakech, Morocco, may.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice, September.

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.

Anthony Sigogne, Matthieu Constant, and Éric Laporte. 2011. Integration of data from a syntactic lexicon into generative and discriminative probabilistic parsers. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 363–370, Hissar, Bulgaria.

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.

Robert S. Swier and Suzanne Stevenson. 2005. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, pages 883–890, Vancouver, British Columbia, 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*, Bombay, India.