Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1332–1343
Marseille, 20-25 June 2022

Making a Semantic Event-type Ontology Multilingual

Zdeňka Urešová[1], Karolina Zaczynska[2], Peter Bourgonje, Eva Fučíková[1], Georg Rehm[2], Jan Hajič[1]
[1]Institute of Formal and Applied Linguistics, Charles University Prague, Czech Republic
[2]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
{uresova,fucikova,hajic}@ufal.mff.cuni.cz, {karolina.zaczynska,georg.rehm}@dfki.de, bourgonje@gmail.com

Abstract

We present an extension of the SynSemClass event-type ontology, originally conceived as a bilingual Czech-English resource. We added German entries to the classes representing the concepts of the ontology. Having a different starting point than the original work (unannotated parallel corpus without links to a valency lexicon and, of course, different existing lexical resources), it was a challenge to adapt the annotation guidelines, the data model and the tools used for the original version. We describe the process and results of working in such a setup. We also show the next steps to adapt the annotation process, data structures and formats and tools necessary to make the addition of a new language in the future more smooth and efficient, and possibly to allow for various teams to work on SynSemClass extensions to many languages concurrently. We also present the latest release which contains the results of adding German, freely available for download as well as for online access.

Keywords: lexical resource, ontology, event, verb, valency, syntax, semantics, multilinguality, Czech, English, German

1 Introduction

In Natural Language Processing (NLP), lexical resources play an important role for supporting the computer's understanding of human language by providing gold standard data for NLP experiments and a wide range of NLP tasks and applications, such as information extraction, event extraction, sentence similarity, etc. This paper presents results of building a multilingual semantic event-type ontology called SynSemClass, a linguistic resource that can be used to compare semantic and syntactic properties across languages, which provides curated data for NLP experiments with cross-lingual synonyms, such as synonym discovery, feature mapping, etc.

The main purpose of our lexicon is to represent the cross-lingual meaning of the state or event expressed by the set of verbs assigned to the individual synonym classes. We believe that multilingual synonyms support deeper understanding and comparability of verb usage in different languages. Our long-term aim is to create an event-type ontology that can be referenced and used as a human-readable and human-understandable database for all types of events, processes and states. Having the applicability for different tasks in mind, the granularity of the SynSemClass synonym classes are intermediate in granularity, between the relatively broad nature of FrameNet frames, and the often very fine-grained distinctions in WordNet verb senses.

The contributions of this paper are as follows. We present a new workflow for extending the existing bilingual SynSemClass lexicon with new languages by example of German. The re-design of the annotation workflow was necessary as the original annotations for English-Czech were based on a parallel corpus with deep syntactical annotations, the Prague Czech-English Dependency Treebank, which is not available for all languages. The new workflow was designed so that the majority of steps can be applied for other languages, too, with minor modifications. We show our first results with the latest release of the lexicon which contains 153 newly added German class members based on the new annotation workflow. We describe the latest release of SynSemClass v3.5 which contains the interim results of the work described here and which is freely available.[1]

2 Related Work

Work related to our project concerns research on acquiring synonyms from language sources (Section 2.1) and research integrating examined information from multiple sources (Section 2.2).

2.1 Synonym Extraction

Synonyms are explored from many perspectives especially in the context of creation of non-English WordNets and lexical ontologies. The research on synonymy is based on both the monolingual resources within one language and on the multilingual resources with cross-linguistic extraction of synonyms.

We shall mention at least a few related projects, such as (Lam et al., 2014; Gonçalo Oliveira and Gomes, 2014) based on monolingual resources, (Helou et al., 2014; Helou et al., 2016) based on bilingual dictionaries and cross-language ontology matching, or (Ercan and Haziyev, 2019) based on multilingual translation graph from multiple Wiktionaries. It is also worth mentioning (Wu and Zhou, 2003; Jarrar et al., 2020) exploring lexical resources and corpora together. Also other combinations of sources, such as Princeton WordNet (Fellbaum, 1998), and word embeddings (Khodak et al., 2017; Al Tarouti and Kalita, 2016), exist.

[1]https://hdl.handle.net/11234/1-3750

In addition to WordNet-related synonym research, synonyms are studied also in the context of searching for translation pairs between multiple languages (Villegas et al., 2016; Torregrosa et al., 2019) and their extraction based on graph analysis as well as in the context of neural machine translation without using parallel data, e. g., (Flati and Navigli, 2012; Gracia et al., 2019).
Recently, a wide range of automatic synonym detection or extraction studies emerged, e. g., (Wang et al., 2010; Wang and Hirst, 2011; Xiang et al., 2020; Sholikah et al., 2020; Al-Matham and Al-Khalifa, 2021).
However, to the best of our knowledge, only our previous work(Urešová et al., 2018c; Urešová et al., 2018b) has addressed the research of cross-lingual verbal synonyms in connection of using syntactic and semantic features for building a cross-lingual semantic event-type ontology. This is the novelty that SynSemClass lexicon brings to research synonyms. In addition, SynSemClass workflow assumes a fully manual pass (even if the pre-annotation or pre-extraction draws on any automatic tool) to check and/or add information.

2.2 Linked Data

Similarly, efforts to link data have been traceable for years, such as the *Multilingual Central Repository (MCR)*[2], a cross-lingual framework for developing Wordnets (Guinovart et al., 2021) integrating information from Wordnets and ontologies (Atserias et al., 2004). The *Event and Situation Ontology (ESO)* reuses and maps across existing resources and also extracts information from text (Segers et al., 2016). Another lexical-semantic resource called *Uby* gathers information from English and German resources (Gurevych et al., 2012; Eckle-Kohler et al., 2015). In terms of Linked Data, there have been previous projects aiming to integrate knowledge kept in the individual resources, especially projects for FrameNet (Ide, 2014) and WordNet (McCrae et al., 2014).
We are aware of several projects very close to SynSemClass in that they are also integrating various existing resources for verbs in a common framework, allowing interoperability among all these sources. One particularly inspiring design we follow and cooperate with is the *SemLink* project (Palmer, 2009; Bonial et al., 2013; Bonial et al., 2012), partially mapping between FrameNet (Baker et al., 1998), VerbNet (Schuler, 2006), PropBank (Palmer et al., 2005b), and WordNet (Miller, 1995).[3]
An integration of models for verbs and predicates similar to SynSemClass is captured in the *Predicate Matrix* (Lopez de Lacalle et al., 2014; de Lacalle et al., 2016) that builds upon Burchardt et al. (2005) and Fellbaum and Baker (2013) and applies knowledge–based word-sense disambiguation algorithms for automatic mappings between WordNet, VerbNet and FrameNet.
Another closely related project is *VerbAtlas* (Di Fabio et al., 2019) which contains semantically-coherent frames with a common, prototypical argument structure while at the same time providing new concept-specific information. *VerbAtlas* is based on *BabelNet*[4], a multilingual semantic ontology (Navigli and Ponzetto, 2010; Navigli and Ponzetto, 2012) which integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia. BabelNet synsets represent a given meaning and contain synonyms which express that meaning in a range of different languages.

3 SynSemClass Lexicon

The multilingual lexicon SynSemClass groups synonymous meanings and structural properties of verbs into classes, each of which represents a single concept (e. g., of *eating*).[5] So far, we have decided to focus on verbal synonyms since they carry the key syntactic-semantic information for language understanding. As described in detail in (Urešová et al., 2019; Urešová et al., 2018a) there is no specific model or lexicographic theory behind building our database. However, the notion of synonymy used is based on the "loose" definition of synonymy by Lyons and Jackson (Lyons, 1968; Jackson, 1988), or alternatively and very closely, on both "near-synonyms" and "partial synonyms" as defined by Lyons (Lyons, 1995; Cruse, 2000) or "plesionyms" as defined by Cruse (Cruse, 1986). The valency theory used in the lexicon is based on the Functional Generative Description theory (Sgall et al., 1986).
A functionally adequate relationship (regardless of language) must exist between the meanings of all verbs (called "class members," or CMs) within one synonym class in SynSemClass, i. e., the English, Czech, and German verbs must be synonymous (at least) in some context(s), as evidenced by the corpora used. This is in line with the general approach to synonymy as described in the previous paragraph. Besides semantic synonymy between the individual CMs, the restrictions are also of semantic-syntactic nature: All synonym classes have a fixed set of *semantic roles*, and for each new CM (German or otherwise) to be added to a class, it is necessary to "fit" in this given roleset, by means of mapping the verb's syntactic arguments to the roleset. Each role describes a possible concrete participant in a concrete event that is of the type defined by the class. For example, the class `vec00476` ***absorb / pohltit***, as concept of "absorbing", has two semantic roles, *Absorber* and *Absorbed*, and for each new CM to be added there, it must be possible, in the prototypical case, to create a mapping between its syntactic arguments and the roles in that class' roleset; see the example in our web-based lexicon (Fig. 1), or in the editor

[2] https://adimen.si.ehu.es/web/MCR

[3] SynSemClass is in fact also listed as part of the SemLink project, see https://uvi.colorado.edu.

[4] https://babelnet.org

[5] This is different from the commonly used term of "semantic classes of verbs" as represented, for example, in VerbNet, where the class is defined much more broadly – such as for all verbs of movement.

used for annotation (Fig. 2).[6]

To have empirical evidence for such decisions, the SynSemClass lexicon is being developed in a "bottom-up" fashion; the first synonym classes with the EN-CZ CMs were taken from actual examples from a parallel English-Czech corpus, the Prague Czech-English Dependency Treebank (PCEDT) (Hajič et al., 2012).[7] For any extension by new entries or to a new language, we also require that corpora and existing lexicons are used to support the automatic pre-extraction of candidate words as well as the annotation decisions taken during the manual pass.

3.1 Lexicon Structure and Resources Used

For building SynSemClass entries, we decided to use a combined method both source-wise and means-wise. Both corpora and lexical resources served as data sources. In terms of the way we work, we chose a combination of manual and automatic means.

The individual "conceptual" entries of SynSemClass are simply called synonym classes; each class is assigned a common set of semantic roles, called a "roleset", indicating the prototypical meaning of the given class. A roleset contains the core (types of) "situational participants", called "semantic roles", which are common for all the class members (the individual verb senses) in one class. Each class in SynSemClass is viewed as a substitute for an ontology unit representing a single concept, similar to the treatment of WordNet synsets in (McCrae et al., 2014). While the the semantic roles resemble FrameNet "Frame Elements" (and sometimes borrow their names from there), it should be pointed out that there is one fundamental difference: the semantic roles used in SynSemClass aim at being defined across the ontology, and not per class (as they would be if we follow the "per frame" approach used in FrameNet).

Each member (a verb *sense*) of one class is denoted by a verb lemma and the valency frame ID which, roughly speaking, represents the particular verb sense. Those class members are further linked to the original resources used and also to several external resources to support, e. g., comparative studies, or any other possible research in the community. The Czech CMs refer and are mapped to the following Czech lexical resources: to the PDT-Vallex (Urešová et al., 2014) that was used for building the Czech part of the PCEDT, to the lexicon of Czech and English translation equivalents called CzEngVallex (Urešová et al., 2015) and to VALLEX (Lopatková et al., 2020), containing annotated Czech verbs with over 6,000 valency frames (Lopatková et al., 2016).[8] The English CMs are mapped to the English valency lexicon used for building the English part of the PCEDT, called EngVallex (Cinková et al., 2014), to CzEngVallex (Urešová et al., 2015), FrameNet (Baker et al., 1998; Fillmore et al., 2003), VerbNet (Schuler, 2006), PropBank (Palmer et al., 2005b), senses from OntoNotes Groups (Pradhan and Xue, 2009), and WordNet (Miller, 1995; Fellbaum, 1998).

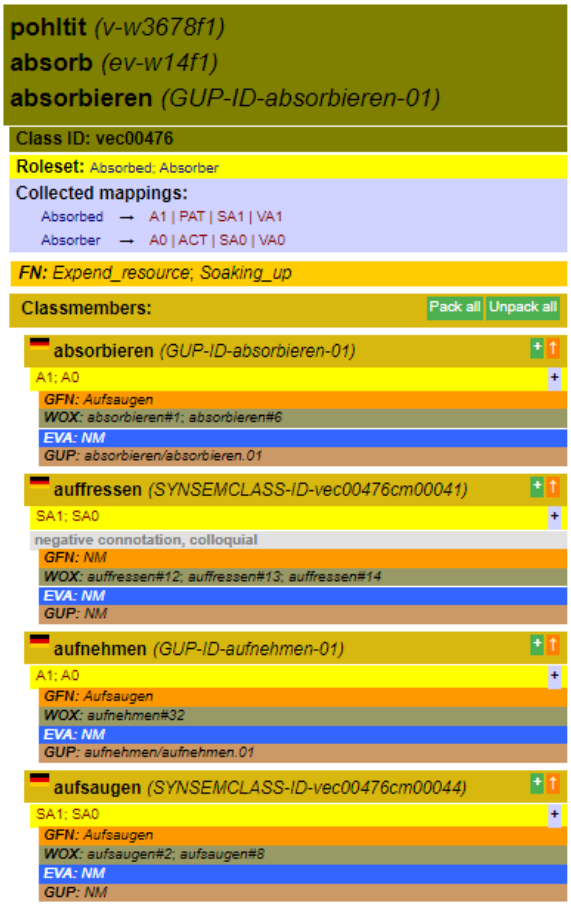

Figure 1: The "absorbieren" class with German entries only (simplified)

3.2 Annotation Process

The annotation process does not start from scratch, but (as explained above and also in previous publications, e. g., in (Uresova et al., 2020)) in steps in which first an alignment based on a parallel corpus is used to extract candidate class members in the other language. The annotators work in two phases as described below. Based on what the annotators decide to keep included in the resource, the next step works in the opposite language direction on the same parallel corpus, and again this automatically extended candidate set goes to the annotators. For Czech-English, there are four such steps, but for German, there was only one: from English as source language to German as target language (see Sect. 4). This was in the interest of time; the additional steps can be added later, or replaced by any other new means of pre-extracting additional CM candidates.

[6]The "live" web version (with cs, en, de CMs) is available at https://lindat.mff.cuni.cz/services/SynSemClass35.

[7]The predecessor resource was called "CzEngClass" since it started as a bilingual-only resource.

[8]https://ufal.mff.cuni.cz/vallex/4.0

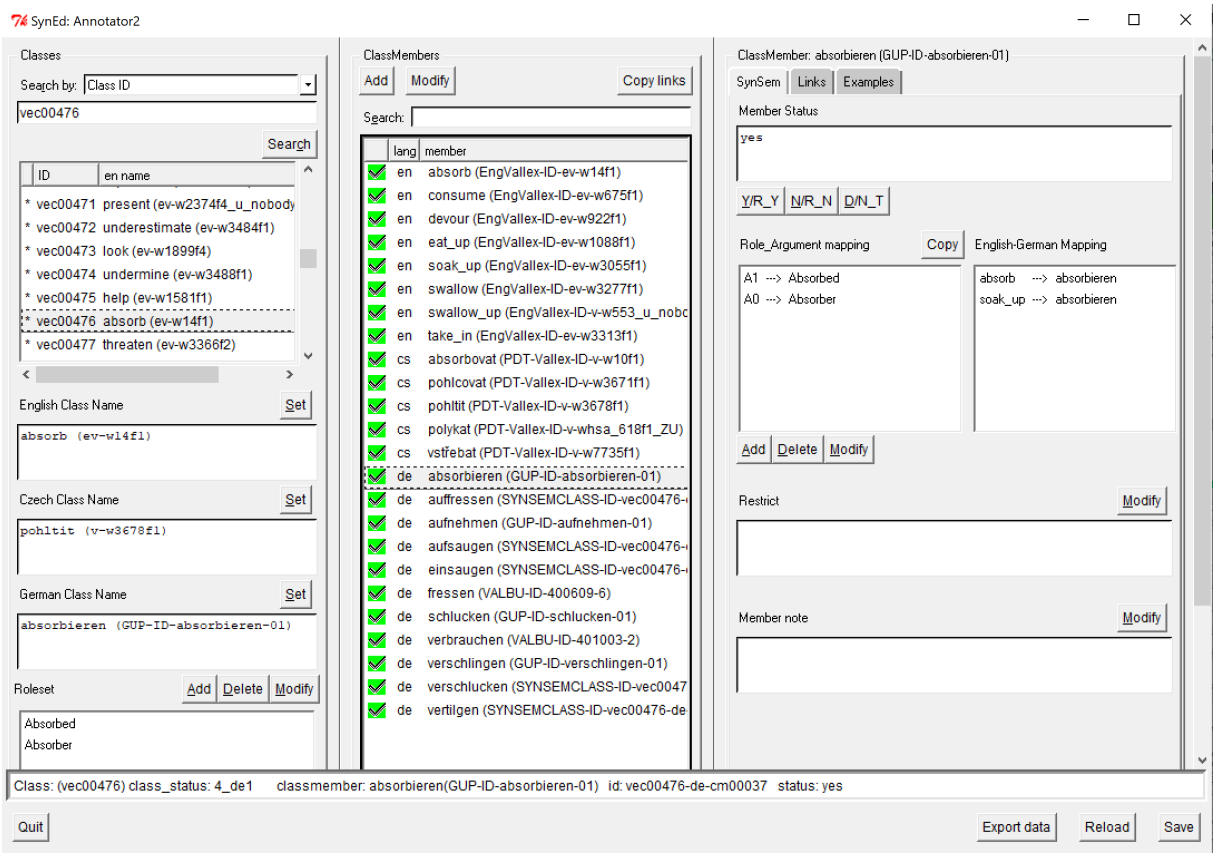

Figure 2: Our annotation tool SynEd and the class "Pohltit/Absorb/Absorbieren" with cs/en/de entries; the CM "absorbieren" (German) is highlighted to demonstrate the mapping to roles on one example.

To facilitate the annotators' work (within each step), the annotation workflow has been split into two separate phases where the first one is a quick pre-filtering (pruning) of candidate synonyms based mainly on the semantic-only synonymy requirements (see the introductory remarks in Sect. 3). The second phase then continues with fine-grained annotations, including annotating the roleset mappings for every given class member, adding links from external lexical resources and selecting example sentences (Fig. 2). During the second phase, any CM can still be excluded if the additional requirements are not fully met (Sect. 3).

4 Adding German

Adding German synonyms to the already existing bilingual (English-Czech) synonym classes was a challenge since it is the first language added to the existing lexicon that is not processed within the original internal team only. Also, the input set of resources was not as rich as it was for Czech: We could not build upon a parallel corpus with deep syntactic annotation like the PCEDT for English-Czech. Several new technical challenges had to be tackled and a new workflow had to be designed. We also faced some new research questions; among the most important were: Is SynSemClass ready for (more) multilinguality?, Can it be extended (semi-)automatically, efficiently, with what accuracy? Are any changes in its structure, semantic roles, sense definitions, class hierarchy, etc. necessary to allow for multilinguality (to more concepts, more verbs to existing concepts, more languages)?

From our first experiments with German in a pilot study of approximately 100 German verb roots (Bourgonje et al., 2021), we learned that expanding SynSemClass to include German means the completion of two tasks:

- to find correspondences between German synonyms and their English and Czech counterparts, and
- to include semantic and syntactic information for German synonyms, including the links to external lexical resources.

To accomplish the first task and to extract German synonyms, we use an automatic aligner on a parallel corpus to extract candidate synonyms. The second task is performed by manual annotations using a tool called Synonyms Editor we designed for this purpose (SynEd) (Urešová et al., 2018d). SynEd allows to edit CMs and the required syntactic-semantic information (mapping between semantic roles and valency arguments), add and edit links to external resources and select examples.

4.1 Automatic Extraction of Candidate CMs

For the first phase of extracting a list of candidate synonyms we utilize a sentence-aligned parallel corpus called ParaCrawl[9], a web-crawled corpus with over 82 million parallel sentences for the English-German part. The goal is to go from a Czech-English synonym lexicon based on the PCEDT corpus towards a generally applicable workflow for creating a multilingual synonym lexicon. ParaCrawl seems to be sufficient for this task as the corpus contains parallel datasets for at least 23 European languages paired with English and is continuously expanded with new language pairs that we can use for the next iterations of the expansion of SynSemClass (Chen et al., 2020). Similar to our former case study (Bourgonje et al., 2021), we automate the process of candidate extraction by extracting the most common word alignments in the corpus, using already existing English CMs as input and gaining the most common German translations as candidate CMs. For each English reference verb, we extracted the most frequent alignments with a cut-off of 0.2%, meaning that if the particular English verb was aligned to a particular German word or phrase in more than 0.2% of cases (in English) it was selected, and discarded otherwise. We use the statistical word alignment tool MGIZA++ (Gao and Vogel, 2008) (a multi-threaded version of GIZA++ (Och and Ney, 2003)) to conduct the German-English word alignments.[10] Due to memory limitations for building the alignment model, we reduce the dataset to 20 million sentences and apply basic preprocessing steps, like filtering lines inside the corpus not ending with a dot (titles, chopped sentences, other parts from websites) and lines including more than one sentence, which left us with 10,400,358 sentences for each language. The repository for preprocessing and extracting the word alignments is available on GitHub.[11] The next steps include the annotation of candidate verbs (Sect. 3.2) to populate each SynSem class with German CMs, using the extracted example sentences from ParaCrawl and enriching the new CMs with semantic and syntactic information, as described in Sect. 4.4. The annotation workflow for extending our lexicon for more languages is shown in Figure 3.

Parallel to this work it was necessary to create an annotation manual (Urešová et al., 2021) for the new task, train the annotators, test the quality of their work and adapt the workflow (Sect. 4.2) as well as refine our annotation tool SynEd for the the German annotations.

4.2 Annotation of German CMs

The annotation process was conducted with two or three annotators for each annotation phase. Two of the annotators were fluent in English and German, one annotator was fluent in all three languages, therefore the German annotations could be based on the English synonyms and double-checked with the Czech ones. The annotations were monitored and, in case of disagreement, the final decisions were made by one or two authors of this paper.

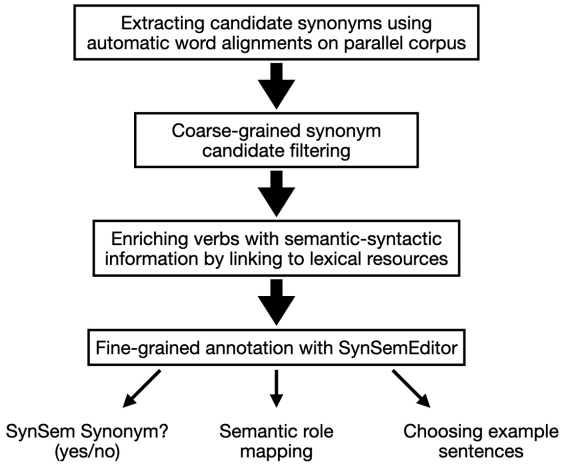

Figure 3: Annotation workflow

4.3 First Annotation Phase: Prefiltering

As first step, the output list of the most common word alignments of German verbs based on English (Sect. 4.1) was presented to the annotators, who filtered the entries mainly according to synonymous meaning and regarding formal requirements. One common reason for excluding an entry was the fact that some English words do not differentiate between their noun- and verb-form (e. g., *to rally* versus *rally*), which causes nominal output for the German alignments (we only used word alignments, no part-of-speech-tag information). Another reason was verb alignments additionally including pronouns (*erlauben es*, *allow it*) or particles (*zu genehmigen*, *to approve*), as well as some obviously nonsensical alignments which still made it past the 0.2% threshold. Furthermore, we wanted to include verbs used in an idiomatic construction, if suitable. For these cases, the annotators not only had to label the main verb but also separately label the condition for being a synonym. One example is the phrase *jmd. die Schuld geben* (*to lay the blame on someone*) as synonym for the English CM *accuse*, where the annotators labeled the lemma *geben* and the restriction *Schuld*. After this first filtering step the remaining German CM candidates were used for subsequent fine-grained annotation steps described in the next section.

Based on a sample of four SynSem classes with approx. 6,75 English verbs and 295 German candidates per SynSem class, it took the annotators about one hour and 18 minutes to annotate (pre-filter) one class. Most candidate class members have been filtered out: out of the 7,442 class member candidates in 32 classes (on which we did multiple annotation for measuring inter-

[9] Corpus release v8.0, April 2021, https://paracrawl.eu/v8

[10] The small 0.2% threshold was set in this initial experiment not to lose recall in the manual pass. Since many candidates have been discarded, it might be increased in the future.

[11] https://github.com/linatal/SynSemClass_Ger

annotator agreement, IAA) extracted from the automatic alignments, only 3,61% was kept in their class, as counted on the gold data obtained by the usual adjudication process.[12] Since such result represents a very skewed distribution, we are thus presenting three different IAA figures for the reader to get perhaps a less "biased" picture (Table 1). The Cohen's κ value is macroaveraged over three batches of 11, 11 and 10 classes, accuracy measures the agreement between the gold data and the Yes/No annotations, and F_1 is a standard measure that takes into account recall and precision (computed from true positives, false positives and false negatives only), abstracting from the many true negatives which bias the other two measures heavily.

Cohen's κ	Accuracy	F_1-measure
0.35	96.89%	0.61

Table 1: IAA for German class membership, using three different metrics

We consider the F_1-measure to be the most realistic assessment of how well the annotators can judge the "synonymity" of the candidate verbs to retain them in the suggested class. As seen in Table 1, Cohen's κ is low (for a binary task, even though Cohen himself (Cohen, 1960) calls kappas between 0.21–0.40 as "fair") and the plain Accuracy unnaturally high, due the extremely skewed prior distribution. The F_1-measure focuses on the gold retained values; it might seem low, but we still have to consider that this was (a) the first 32 classes done for German; the measure was going up across the batches (0.56→0.73) which reflects the growing experience of the annotators, and that (b) the many true negatives still influenced the decisions while trying to draw the line among so many candidate verbs.

4.4 Second Phase: Adding Information to the Retained German Synonym Class Members

For the retained CMs (Sect. 4.3), the annotators curate the links to external lexical resources, map the given roleset of the synonym class to the CM and choose example sentences (from the ParaCrawl corpus). We now describe these steps in more detail.

The annotation of German synonyms builds upon established bilingual classes of English-Czech verb synonyms, which means that the roleset for each synonym class is already defined, and the German verbal candidates are restricted to it. Since the English and Czech synonyms in SynSemClass are linked to many lexical resources (as described in Sect. 3), we wanted to establish similar links for German. Furthermore, the lexical resources provided syntactic information about the valency and/or semantic roles for the German verbs, which were used by the annotators in the annotation editor for establishing the semantic-syntactic mapping (see also Fig. 2). Linking SynSemClass to other relevant lexical sources supports interoperability with other work, and thus gives the resource a higher value for use in computational linguistics. At the same time, links to other semantic databases enable providing richer and comparative information about the meaning, characteristics and use of the CMs as described in other lexical entries.

The external lexical resources for German we link to are the following:

(1) *FrameNet des Deutschen* (FdD)(Lönneker and Ziem, 2018)[13] is a frame-based lexicon for words and fixed multi-word units of German based on the Berkeley FrameNet (Baker et al., 1998).[14] It currently comprises about 1,220 frames with a definition, example sentence and a link to the original FrameNet entry, but it does not include lexical entries for German as in the original FrameNet. For SynSemClass, the annotators were asked to link each German CM to a FdD entry when possible. It is therefore possible to see SynSemClass as a potential extension for FdD providing verbal lexical entries for each class, which still need to be evaluated. From the point of view of the annotation of semantic roles, FdD is the most important source and inspiration for SynSemClass since it also works with semantic roles (frame elements) within semantic frames.

(2) *Universal Proposition Banks* (UPB) (Akbik et al., 2016)[15], provides a list of German verbs annotated with frame and role labels linked to the English Proposition Bank (PropBank) (Palmer et al., 2005a). As it contains mappings between syntactic and semantic information for each verb entry, the resource helped the annotators in creating a link between the argument structure of a CM and the roleset of a class.

(3) *Elektronisches Valenzlexikon des Deutschen* (Electronic German Valency Lexicon, in short E-VALBU), is a valency lexicon for German verbs. It proved to be very helpful for our purpose, as E-VALBU provides entries for different senses of one verb (so-called *Lesarten*). One lexical form of a verb can therefore have several entries, each for one meaning. The entries include syntactic information (i. e., valency frames), a description of the meaning and example sentences. By matching the CMs to the E-VALBU entries we could therefore add valuable information regarding the usages and meanings of verbs into our lexicon.

(4) *Woxikon*[16] is a German synonym lexicon organizing different lemmas or fixed word groups into semantically related synonym groups but without providing

[12] For IAA, we mapped the annotators' decisions to a simple binary one: to keep the candidate verb or not.

[13] https://gsw.phil.hhu.de/framenet/

[14] For more details see several papers on FdD at https:// https://gsw.phil.hhu.de/wp?id=179. FdD is currently being updated and we will relink to the actual version once it is fully online.

[15] https://github.com/System-T/UniversalPropositions

[16] https://synonyme.woxikon.de

any further (syntactic or semantic) information.
We used the entries of UPB and E-VALBU not only as the additional semantic and syntactic information for the roleset mapping, but also to have a URL for each CM. The URLs ensure linked data compatibility for the German part of SynSemClass. The annotators chose the best fitting external entry from which the valency frames as well as the IDs were derived for the CMs. In cases where both the UPB and the E-VALBU entry could be linked to the same CM, we gave preference to the latter. If none of the existing resources could be linked, we created an individual SynSemClass-ID for the CM.

4.5 Towards Multilinguality

The extension of SynSemClass by German entries helped us to better understand the task of adding another language to the ontology. While it did not reveal any problems with the "theoretical" approach, namely, the overall design of the ontology and its principles, it did reveal certain issues that originally did not come up when working on the first pair of languages. These issues can be categorized as follows:

- Technical issues: while having a certain appeal from the users' perspective, such as simplicity of use, it is no longer possible to keep all data in one file: perhaps it would still be possible when the number of languages is in the single digits, but if it grows (as we hope), then moving a file of several gigabytes for both editing and using in applications (especially if only some languages are in focus) is not feasible.

- Organizational issues: it is assumed that in the future, if a substantial number of languages is to be added, work has to be done concurrently by many teams across many places, not just one team in one institution. This also requires that the datasets (files) are as independent as possible, even if stored centrally in one repository, such as GitHub. Resolving conflicts in one huge file is far from optimal.

- Resource issues: every new language has a different set of "input" resources available, some more and some less rich. For example, the original pair has had richly annotated bilingual parallel corpus and valency dictionaries available, while for German, only a parallel corpus and relatively sparse coverage valency lexicons have been at our disposal. This has led to the working definition of a minimally required set of resources for a new language to be added: a parallel corpus and at least one external resource that contains at least some syntactic and/or sense information for verbs.

- Needs for tools: the tools to be used (and offered to other teams to work on) must reflect the above changes, as must the guidelines. The tools have to be configurable to work on a (sub)set of languages, the team working on a given language must be able to add language-specific external lexicons and define certain sets of labels to work with (such as for syntactic properties). The central maintainers must have tools to validate individual languages, as well as to identify changes that affect the main dataset(s) and the labels and external resources defined centrally, and which affect only individual languages, and vice versa.

The property of any multilingual ontology (and SynSemClass is no exception) is that the core set of "concepts" (or by whichever name this set is known) to which the language-specific "words" (or terms, or MWEs, etc.) are linked to in order to enable human readability, is very hard to keep in sync with these language-specific "descriptions". Every change in the core set of concepts (like a split of a concept into two more specific ones, merge with another concept, change in properties or features of that concept, change in the hierarchy of concepts, or addition of a new concept) affects possibly all the language-specific parts. This will have to be reflected in the workflows, versioning, etc., in order to keep the whole resource consistent, but also to allow for its expansion both in terms of languages and contents (coverage).
These considerations all affect SynSemClass, too. It would be naive to think that only new classes will be added to SynSemClass – a change that is the least disrupting (since language content to the new class can be added gradually, as it will work from the start with all languages); based on feedback from the various language-specific annotation efforts, some classes will have to be split, or merged, or semantic roles changed. A suitable mechanism for handling such changes, i. e., changes above adding language content in parallel, adding a class, or adding an external resource for a particular language, will still have to be worked out. We are inspired by the success of the Universal Dependencies (UD) project,[17], but not all of the setup and environment can be taken over from it – exactly because of the existence of the central dataset (the core list of concepts), which is absent from text annotation projects (like the UD, where the only central things are the guidelines, technical dataset format, and small, fixed sets of labels, such as dependency relations or morphological features which only change once in several years with a major version change).
So far, we have created a new data schema, where language-dependent data are separate from the core (main) definitions. The main editing tool has been adapted[18] to work with the new data schema. We have defined a set of requirements which has to be configured by the new language team to start with a new lan-

[17] https://universaldependencies.org

[18] The SynSemClass-related tools will be described in detail in a different upcoming paper.

guage, namely the set of annotators and their authorizations to edit the data, the external lexicons by means of URLs or APIs, the way of acquiring IDs for the entries in terms of individual entries (in case they are polysemous), and the list of labels for predicate arguments (or valency slot labels, depending which resource is used for that language).
This new setup will be used for adding Spanish, to be done by our own group, but in such a way that it will "simulate" the case when some group/team will want to work on their own language (almost) independently (meaning just with the usual central support, as it is done within the UD project, for example).

5 Results

The latest release, SynSemClass3.5[19], published in July 2021, contains 600 classes. Of the original 56,022 CM candidates, approx. 9,000 CMs remained in this lexicon version, i. e., approx. 4,630 English, 4,249 Czech – and the 153 newly added German CMs (see Sect. 4). We consider this edition to be groundbreaking in that it contains a new language (German) which comes from another external corpus. It marks the start of making SynSemClass multilingual, even if several issues still remain (cf. Sect. 4.5).
The resource is also available for browsing online.[20] As a step towards full multilinguality, it now allows users to select the language(s) whose CMs to show when displaying an entry. Similarly, it allows to display (only) user-selected external resources (such as FrameNet for English, Woxikon for German, VALLEX for Czech, etc.), in order to unclutter the display. More features will be added (such as search across languages, search by roles, etc.) as the number of languages grows.
Since July 2021, the lexicon has grown considerably, doubling the number of classes. From the original 70,839 CM candidates, 6,105 English, 6,037 Czech, and 533 German CMs (in 61 classes) are kept. The release of SynSemClass version 4 is planned for early spring 2022. Version 4 is also undergoing both a detailed and intensive annotation check and contains new features, such as semantic role definitions, semantic role hierarchy, (Czech) aspectual verb counterparts, etc.; these additional features are however mostly unrelated to the work on German as described here (Sect. 4).

6 Conclusions and Future Work

We have described the process and results of the extension of the SynSemClass event-type ontology to cover German. This extension did not only serve to add German verbs to the ontology, but also has helped us to understand the necessary conditions for adding another language to such a resource, and to formulate the "division of work" (and of resources and tools) between the shared, central (core or main) part of the ontology and the language-dependent parts. At the same time, we have been able to generalize the description of the resources needed for the preparation phase of adding a new language, which includes the identification of the necessary resources, including existing syntactic and semantic lexicons and sufficiently-sized parallel corpora. On the technical side, the data structure has been changed to separate the language-dependent data in a stand-off fashion, and the editor has been updated and enhanced to reflect this change. The dataset (version 4 of SynSemClass) will be published soon, as will the editor and related specifications.

For the future, we plan to expand on the work described here in several directions. First, we have already started adding Spanish, and we hope that after that, we will be able to assemble a larger group of interested teams to work on more languages in parallel, either from scratch or by converting other similar resources and updating them according to the SynSemClass specifications. We chose ParaCrawl[21] because it is the largest resource, and it is constantly being developed and more languages are added (in version 9 from Sept. 2021, Paracrawl provides corpora for over 40 language-pairs). Therefore, we (or anyone interested) can easily expand SynSemClass to other languages by adopting our data preparation steps as developed for the extension to German.

At the same time, we plan to enrich the SynSemClass specification in several respects – across all languages, or more precisely, in the language-independent core part: add definitions/descriptions to all semantic roles used and later to all classes and add hierarchy (also to both roles and classes). In the language-specific part, we are working on providing (automated) tools for expanding the coverage on a large scale (using large contextual language models), adding nouns and adjectives expressing the concept in syntactically different ways, and enhancing the web browsing and searching functionality for the resulting lexicon. Also, we plan to publish separately the more theoretical lexical-semantic considerations, findings and comparisons that arose during the so far rather bottom-up approach that has not been driven, in the core questions, by any existing theoretical framework.

It should also be mentioned that the project is part of a larger early-stage project for multilingual knowledge representation, where the SynSemClass entries (classes) will serve as a grounding (of sorts) for all events and states in that representation, connecting (relating) all other entities in the resulting representation which will also be grounded (by other means). While some verb annotation experiments have been done for the previous versions of SynSemClass, the full specification is still to be developed.

[19] https://hdl.handle.net/11234/1-3750
[20] https://lindat.cz/services/SynSemClass35
[21] https://paracrawl.eu

7 Acknowledgements

The idea of adding German and the related work to make it happen has been conceived and subsequently funded in part as a collaborative microproject by the EU project HumaneAI-Net,[22] (GA no. 952026). The SynSemClass project, the dataset creation and curation, and its online access has been supported by the LINDAT/CLARIAH-CZ Research Infrastructure (project no. LM2018101), funded by the Ministry of Education, Youth and Sports of the Czech Republic and by the project LUSyD, funded by the Grant Agency of the Czech Republic under the EXPRO programme as project no. GX20-16819X, and by funding from the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (Wachstumskern no. 03WKDA1A). We would also like to thank the annotators and their valuable input to the work presented here: Charlotte Friedrich, Kateřina Rysová and Danny Srp.

8 Bibliographical References

Akbik, A., Guan, X., and Li, Y. (2016). Multilingual aliasing for auto-generating proposition Banks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3466–3474, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Al-Matham, R. N. and Al-Khalifa, H. S. (2021). Synoextractor: a novel pipeline for arabic synonym extraction using word2vec word embeddings. *Complexity*, 2021.

Al Tarouti, F. and Kalita, J. (2016). Enhancing automatic Wordnet construction using word embeddings. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 30–34, San Diego, California, June. Association for Computational Linguistics.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The meaning multilingual central repository. In *Proceedings of the Second International WordNet Conference*, pages 80–210.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (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 – Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bonial, C., Feely, W., Hwang, J. D., and Palmer, M. (2012). Empirically Validating VerbNet Using SemLink. In *Seventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, Istanbul, Turkey, May.

[22]https://www.humane-ai.eu

Bonial, C., Stowe, K., and Palmer, M. (2013). Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy, September. Association for Computational Linguistics.

Bourgonje, P., Zaczynska, K., Schneider, J. M., Rehm, G., Uresova, Z., and Hajic, J. (2021). SynSemClass for German: Extending a Multilingual Verb Lexicon. In Adrian Paschke, et al., editors, *Proceedings of QURATOR 2021 – 2nd International Conference on Digital Curation Technologies*, volume 2836, Berlin, Germany, 02. CEUR-WS, CEUR-WS. CEUR Workshop Proceedings, Volume 2836. 11/12 February 2021.

Burchardt, A., Erk, K., and Frank, A. (2005). A wordnet detour to framenet. *Sprachtechnologie, Mobile Kommunikation und Linguistische Resourcen*, 01.

Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M., Kamran, A., Kirefu, F., Koehn, P., Ortiz, S., Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of ACL'2020*, pages 4555–4567, 01.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cruse, D. (1986). *Lexical Semantics*. Cambridge University Press, UK.

Cruse, A. (2000). *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford University Press. Oxford, UK.

de Lacalle, M. L., Laparra, E., Aldabe, I., and Rigau, G. (2016). A multilingual predicate matrix. In *LREC*.

Di Fabio, A., Conia, S., and Navigli, R. (2019). VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China, November. Association for Computational Linguistics.

Eckle-Kohler, J., McCrae, J. P., and Chiarcos, C. (2015). LemonUby – A large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, 6:371–378.

Ercan, G. and Haziyev, F. (2019). Synset expansion on translation graph for automatic wordnet construction. *Inf. Process. Manag.*, 56:130–150.

Fellbaum, C. and Baker, C. (2013). Comparing and harmonizing different verb classifications in light of a semantic annotation task. *Linguistics*, 51, 01.

Christiane Fellbaum, editor. (1998). *WordNet: An*

Electronic Lexical Database. MIT Press, Cambridge, MA and London.

Fillmore, C. J., Johnson, C. R., and L.Petruck, M. R. (2003). Background to FrameNet: FrameNet and Frame Semantics. *International Journal of Lexicography*, 16(3):235–250.

Flati, T. and Navigli, R. (2012). The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary. *Journal of Artificial Intelligence Research*, 43:135–171, Feb.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Gonçalo Oliveira, H. and Gomes, P. (2014). Eco and onto.pt: A flexible approach for creating a portuguese wordnet automatically. *Language Resources and Evaluation*, 42:373–, 06.

Jorge Gracia, et al., editors. (2019). *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019), Leipzig, Germany, May 20, 2019*, volume 2493 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Guinovart, X. G., Gonzalez-Dios, I., Oliver, A., and Rigau, G. (2021). Multilingual central repository: a cross-lingual framework for developing wordnets. *ArXiv*, abs/2107.00333.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby – a large-scale unified lexical-semantic resource based on lmf. In *EACL*.

Helou, M. A., Palmonari, M., Jarrar, M., and Fellbaum, C. (2014). Towards building lexical ontology via cross-language matching. In *Proceedings of the Seventh Global Wordnet Conference*, pages 346–354, Tartu, Estonia, January. University of Tartu Press.

Helou, M. A., Palmonari, M., and Jarrar, M. (2016). Effectiveness of automatic translations for cross-lingual ontology mapping. *J. Artif. Int. Res.*, 55(1):165–208, jan.

Ide, N. (2014). FrameNet and linked data. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 18–21, Baltimore, MD, USA, June. Association for Computational Linguistics.

Jackson, H. (1988). *Words and Their Meaning*. Routledge.

Jarrar, M., Karajah, E., Khalifa, M., and Shaalan, K. (2020). Extracting synonyms from bilingual dictionaries.

Khodak, M., Risteski, A., Fellbaum, C., and Arora, S. (2017). Automated WordNet construction using word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23, Valencia, Spain, April. Association for Computational Linguistics.

Lam, K. N., Tarouti, F. A., and Kalita, J. (2014). Automatically constructing wordnet synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 106–111. The Association for Computer Linguistics.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Nakladatelství Karolinum, Praha, Czechia.

Lopez de Lacalle, M., Laparra, E., and Rigau, G. (2014). Predicate matrix: extending SemLink through WordNet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 903–909, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press.

Lyons, J. (1995). *Linguistic Semantics*. Cambridge University Press.

Lönneker, B. and Ziem, A., (2018). *Wissensdarstellung mit Frames in modernen Terminologiesystemen*, pages 289–330. dup.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics, colocated with LREC 2014*, Reykjavik, Iceland.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November.

Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *ACL*.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Palmer, M., Gildea, D., and Kingsbury, P. (2005a). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, M., Gildea, D., and Kingsbury, P. (2005b). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, March.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9–15.

Pradhan, S. S. and Xue, N. (2009). OntoNotes: The 90% solution. In *Proceedings of Human Language*

Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts, pages 11–12, Boulder, Colorado, May. Association for Computational Linguistics.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Segers, R., Laparra, E., Rospocher, M., Vossen, P., Rigau, G., and Ilievski, F. (2016). The predicate matrix and the event and implied situation ontology: Making more of events. In *GWC*.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht.

Sholikah, R. W., Arifin, A. Z., Fatichah, C., and Purwarianti, A. (2020). Semantic relation detection based on multi-task learning and cross-lingual-view embedding. *INTERNATIONAL JOURNAL*, 13(3):33–45.

Torregrosa, D., Arcan, M., Ahmadi, S., and McCrae, J. P. (2019). TIAD 2019 shared task: Leveraging knowledge graphs with neural machine translation for automatic multilingual dictionary generation. In Jorge Gracia, et al., editors, *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019), Leipzig, Germany, May 20, 2019*, volume 2493 of *CEUR Workshop Proceedings*, pages 24–31. CEUR-WS.org.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018a). Creating a Verb Synonym Lexicon Based on a Parallel Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018b). A Cross-lingual synonym classes lexicon. *Prace Filologiczne*, LXXII:405–418.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018c). Defining verbal synonyms: between syntax and semantics. In Dag Haug, et al., editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), Vol. 155*, Linköping Electronic Conference Proceedings, pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018d). Tools for Building an Interlinked Synonym Lexicon Network. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2019). Meaning and Semantic Roles in CzEngClass Lexicon. *Jazykovedný časopis / Journal of Linguistics*, 70(2):403–411.

Uresova, Z., Fucikova, E., Hajicova, E., and Hajic, J. (2020). SynSemClass linked lexicon: Mapping synonymy between languages. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 10–19, Marseille, France, May. European Language Resources Association.

Urešová, Z., Fučíková, E., Hajič, J., and Zaczynska, K. (2021). Annotation guidelines for German verbal synonyms included in SynSemClass Lexicon. Technical Report TR-2021-70, ÚFAL MFF UK.

Villegas, M., Melero, M., Bel, N., and Gracia, J. (2016). Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 868–876, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Wang, T. and Hirst, G. (2011). Refining the Notions of Depth and Density in WordNet-Based Semantic Similarity Measures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1003–1011, USA. Association for Computational Linguistics.

Wang, W., Thomas, C., Sheth, A., and Chan, V. (2010). Pattern-based synonym and antonym extraction. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, New York, NY, USA. Association for Computing Machinery.

Wu, H. and Zhou, M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.

Xiang, L., Guo, G., Yu, J., Sheng, V. S., and Yang, P. (2020). A convolutional neural network-based linguistic steganalysis for synonym substitution steganography. *Mathematical Biosciences and Engineering*, 17(2):1041–1058.

9 Language Resource References

Cinková, Silvie and Fučíková, Eva and Šindlerová, Jana and Hajič, Jan. (2014). *EngVallex - English Valency Lexicon*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0023-4337-2.

Hajič, Jan and Hajičová, Eva and Panevová, Jarmila and Sgall, Petr and Cinková, Silvie and Fučíková, Eva and Mikulová, Marie and Pajas, Petr and Popelka, Jan and Semecký, Jiří and Šindlerová, Jana and Štěpánek, Jan and Toman, Josef and Urešová, Zdeňka and Žabokrtský, Zdeněk. (2012). *Prague Czech-English Dependency Treebank 2.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, ISLRN or PID: https://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.

Lopatková, Markéta and Kettnerová, Václava and Vernerová, Anna and Bejček, Eduard and Žabokrtský, Zdeněk. (2020). *VALLEX 4.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, ISLRN or PID: https://hdl.handle.net/11234/1-3524.

Urešová, Zdeňka and Štěpánek, Jan and Hajič, Jan and Panevová, Jarmila and Mikulová, Marie. (2014). *PDT-Vallex: Czech Valency lexicon linked to treebanks*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0023-4338-F.

Zdeňka Urešová and Eva Fučíková and Jan Hajič and Jana Šindlerová. (2015). *CzEngVallex - Czech English Valency Lexicon*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1512.