Spanish Verbal Synonyms in the SynSemClass Ontology

Cristina Fernández-Alcaina, Eva Fučíková, Jan Hajič and Zdeňka Urešová
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
{alcaina,fucikova,hajic,uresova}@ufal.mff.cuni.cz

Abstract

This paper presents ongoing work in the expansion of the multilingual semantic event-type ontology SynSemClass (Czech-English-German) to include Spanish. As in previous versions of the lexicon, Spanish verbal synonyms have been collected from a sentence-aligned parallel corpus and classified into classes based on their syntactic-semantic properties. Each class member is linked to a number of syntactic and/or semantic resources specific to each language, thus enriching the annotation and enabling interoperability. This paper describes the procedure for the data extraction and annotation of Spanish verbal synonyms in the lexicon.

1 Introduction

The work presented in this paper is part of a larger project aiming at building an event-type multilingual ontology. SynSemClass (SSC) (Urešová et al., 2020) is a multilingual verbal lexicon where contextually-based synonymous verbs are classified into classes based on the semantic and syntactic properties they display. Synonymy here is understood in terms of contextually-based synonymy: a verb considered a member of a class must reflect the same (or similar) meaning expressed by the class in the same context (i.e., it has a similar "semantic behavior" as the other verbs) both monolingually and cross-lingually. Apart from providing fine-grained syntactic-semantic multilingual annotation, SynSemClass also contributes to research by building a database that links several resources in different languages (Czech, English and German). The information gathered in the lexicon allows for a comparison across languages relevant for linguistic research at the same it provides curated data for Natural Language Processing tasks, such cross-lingual synonyms or synonymy discovery.

This paper presents ongoing work on the extension of SynSemClass to include a fourth language, Spanish. Specifically, section 2 introduces the SynSemClass lexicon and section 3 describes the set of Spanish resources linked to SynSemClass. The method for data extraction and annotation of Spanish verbal synonyms is presented in section 4. The results obtained and the limitations encountered during the process are summarized in section 5. The paper closes (section 6) with a summary of the main findings and with some hints for future work.

2 SynSemClass

SynSemClass (Urešová et al., 2020) attempts to create specifications and definitions of a hierarchical event-type ontology while focusing on contextually-based synonymy (both monolingually and cross-lingually) and verb valency in a multilingual setting. The notion of synonymy is regarded in a broad sense based on definitions such as "near-synonyms", "partial synonyms" or "plesionyms" (Lyons, 1968; Jackson, 1988; Lyons, 1995; Cruse, 2000, 1986). The approach to valency used in SynSemClass is based on the linguistic descriptive framework Functional Generative Description (FGD) (Sgall et al., 1986) and its application in the Prague Dependency Treebanks (Hajič et al., 2006; Hajič et al., 2012, 2020; Hajič et al., 2020).

Entries in the lexicon are grouped into individual multilingual (for now, English, Czech, and German) *verbal synonym classes*. Each class is considered similar to an ontological unit and it is assigned a specific set of roles (i.e. *Roleset*), which expresses the prototypical meaning of the class (Urešová et al., 2022). The most important criterion for inclusion of a particular verb (sense) into a given class is the mapping of each of the semantic roles specified in the class Roleset to the verb valency slots (represented by a syntactic-semantic functor[1]) captured in the valency frame (*Role* ↔

[1]The syntactic-semantic (tectogrammatical) functor is used in the FGD valency theory as a label for valency frame members.

Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 11 - 20

Argument mapping). While sharing the same set of roles is a requirement for a verb to be included in a class, roles can be expressed by different morphosyntactic realizations and be subject to additional restrictions (Urešová et al., 2018a). Another criterion for verb sense inclusion is a functionally adequate relationship (i.e., in terms of translation, the verb senses are considered synonymous in the given context(s) if the translated verb in the target language adequately expresses the functional intent of the original language) between the meanings of all class members in one synonym class.

The SynSemClass class members (individual multilingual synonym verb senses, CMs) are linked to entries in similar language-specific syntactic and/or semantic databases (Urešová et al., 2020). The English entries are linked to FrameNet (Baker et al., 1998), Princeton WordNet (Fellbaum, 1998), VerbNet (Schuler and Palmer, 2005) and PropBank (Palmer et al., 2005), the German entries to FrameNet des Deutschen (FdD)[2], the Universal Proposition Bank (UPB) (Akbik et al., 2016)[3], the Elektronisches Valenzlexikon des Deutschen[4] (Electronic German Valency Lexicon, in short E-VALBU) (Kubczak, 2014; Schumacher et al., 2018), and Woxikon[5], and the Czech entries to PDT-Vallex (Hajič et al., 2003), which 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., 2017; Lopatková et al., 2020).

The latest release, SynSemClass4.0 (Figure 1), is dated June 2022 and contains 883 classes with approx. 6,000 CMs.[6] As shown in Figure 1, each class in the lexicon is named using the most prototypical verb in each language (*allow*, *dovolit*, *erlauben*). For each class, the online version of the lexicon displays the information related to the Roleset assigned to the class (*Authority*, *Permitted*, *Affected*), the list of class members in each language, their valency frame (e.g., for English *allow*, the valency frame is ACT, EFF, PAT) and the related senses in the external resources used for each language (e.g., for English, English VerbNet (EV), FrameNet (FN) or OntoNotes (ON), among others). It is also possible to display the corpus examples selected to illustrate the class members.

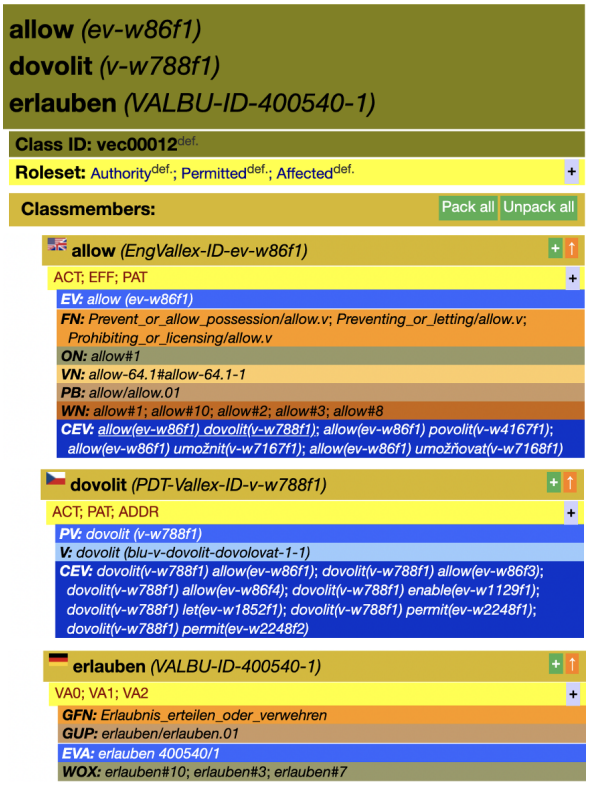

Figure 1: Simplified version of an entry in SynSemClass 4.0 (class *allow*/*dovolit*/*erlauben*).

The work on German has thus been moved to this version (from the previous version 3.5); the fourth version of the lexicon is more complete and it has additional corrections (such as some classes having been merged, etc.) Also, it adds the integration of class and roles definitions. The next release of SynSemClass (presumably version 5, planned for early spring 2023) will be enriched by Spanish synonymous verbs as described here.

3 Resources

Following (Urešová et al., 2022), the minimal set of resources required to add a language to SynSemClass is: (i) a parallel corpus and (ii) (at least) one lexical resource containing syntactic and semantic information. This section describes the corpus and the lexical resources used for Spanish.

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

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

[4] https://grammis.ids-mannheim.de/verbvalenz

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

[6] Available online at https://lindat.cz/services/SynSemClass and for download at http://hdl.handle.net/11234/1-4746. The lexicon can be also now accessed through the Unified Verb Index developed by the University of Colorado Boulder (https://uvi.colorado.edu/).

3.1 Corpus

Verbal synonyms in SynSemClass have been collected from two different parallel corpora, the Prague Czech-English Dependency Treebank Corpus (PCEDT) (Hajič et al., 2012) for Czech-English and the ParaCrawl (Chen et al., 2020) corpus for German-English. For Spanish, for which no corpus richly annotated for syntactic-semantic information is available, the corpus selected for the extraction of Spanish verbal synonyms was the X-SRL dataset (Daza and Frank, 2020). The choice of a parallel corpus is justified based on the assumption that if two words are semantically similar in a given language, their translations would also be similar in another language, both in meaning and in the translation context they share (Urešová et al., 2018b).

The X-SRL dataset is a sentence-aligned parallel corpus containing approx. three million words for the English-Spanish part. The texts are tokenized, lemmatized and POS-tagged.[7] Despite the existence of larger-sized corpora (such as the ParaCrawl corpus), the X-SRL dataset proved to provide enough data for the Spanish part, at least for its initial steps. Furthermore, the X-SRL dataset has the advantage of being composed by English texts extracted from the Wall Street Journal section of the Penn Treebank, on which the PCEDT is based, thus given consistency and cross-coverage of the annotation. In fact, it is possible to find some verbal synonyms for which the examples selected are the same for the Czech-English and Spanish-English parts, as illustrated by verbs *vybuchnout/erupt/hacer erupción*:

Na slavném bulváru Strip ***vybuchne*** *příští měsíc sopka: 60 stop vysoká hora chrlící každých pět minut kouř a oheň.*

A volcano ***will erupt*** *next month on the fabled Strip: a 60-foot mountain spewing smoke and flame every five minutes.*

Un volcán ***hará erupción*** *el próximo mes en la legendaria Franja: una montaña de 60 pies que arroja humo y llamas cada cinco minutos.*

3.2 Lexical resources

Spanish verbal synonyms in SynSemClass are linked to five resources. The resources are of two types: (i) a monolingual valency lexicon which serves as the sense identification source (AnCora) and (ii) four resources that provide extra information; specifically, three monolingual lexicons (SenSem, ADDESE and Spanish FrameNet) and a multilingual resource (Spanish WordNet). What follows is a description of the main features of each of the lexical resources used:

- AnCora[8] is a lexicon based on the corpus AnCora-ES, which is built on texts from Spanish newspapers. The corpus contains 500,000 words and it is annotated at different levels, including syntactic and semantic properties. The resulting lexicon consists of 2,820 lemmas (amounting to 3,938 senses and 5,117 frames). For each verb sense, AnCora provides the argument structure and the thematic roles defined. Each sense in AnCora is also linked (if available) to its English counterpart in VerbNet, PropBank, FrameNet, WordNet 3.0 and OntoNotes, to which the English class members in SynSemClass are also linked. Having links to the same resources in AnCora and SynSemClass is an advantage as it allows for the extraction of only those AnCora senses that are linked to the same English sense contained in a particular class in the lexicon, thus restricting the list of candidate verbs and facilitating their annotation (see section 4.1).

- Spanish SenSem[9] (Alonso et al., 2007) is a monolingual verbal lexicon containing the most frequent 250 verbs. The lexicon is based on the SenSem corpus (Fernández-Montraveta and Vázquez, 2014), which contains approx. 700,000 words from the Spanish newspaper 'El Periódico' and, to a lesser degree, from literary sources. For each sense, SenSem provides a definition, the argument structure and the set of semantic roles. Each sense is also linked to its equivalent in WordNet.

- ADESSE[10] (García-Miguel et al., 2005) is a monolingual verbal lexicon containing 3,400 lemmas and 4,000 verbal entries based on the ARTHUS corpus (1.5 million words), built using texts from European Spanish (78.77%)

[7] The corpus contains information regarding argument labels projected from the original English corpus, but we decided not to use this information as the annotation of arguments other than A0 and A1 is not as fine-grained as our purposes require.

[8] http://clic.ub.edu/corpus/en/ancoraverb_es

[9] http://grial.edu.es/sensem/lexico?idioma=en

[10] http://adesse.uvigo.es

and American Spanish (21.23%)[11]. The lexicon provides information regarding argument structure and semantic roles. Arguments are ordered according to their frequency in the corpus. ADESSE also provides information regarding the argument structure of alternations and examples for each alternation.

- Spanish WordNet 3.0 is integrated within the Multilingual Central Repository (MCR)[12] (Gonzalez-Agirre et al., 2012). The MCR contains wordnets for six languages: English, Basque, Galicia, Catalan, Portuguese and Spanish (including senses from varieties other than European Spanish although without specification). Cross-linguistic synonyms are connected through the Inter-Lingual-Index (ILI). The MCR is also enriched with semantically tagged glosses and contains ontology information from WordNet Domains, Top Ontology and AdimenSUMO.

- Spanish FrameNet[13] (Subirats, 2009) is the Spanish version of the FrameNet project and it is built on a corpus under construction that includes both 'New World and European Spanish'.[14] The online lexical resource is based on frame semantics and supported by corpus data. The resource contains more than 1,000 lexical items (including verbs, but also other parts of speech) from a variety of semantic domains. It provides syntactic and semantic information for each sense automatically annotated and validated by human annotators.

The representation of Spanish varieties in the lexical resources listed above is uneven since most resources are built exclusively (AnCora, SenSem) or mainly on European Spanish (ADESSE). However, in these resources, there is no specific information on this aspect and some characteristics of the senses of a variety other than European Spanish appear without any explicit reference, e.g., *manejar* is included in ADESSE as 'drive' (more frequently used in American Spanish) without further specification regarding variety. To overcome this drawback and whenever possible, annotators have used the information provided by the *Diccionario de la lengua española*[15] and the *Diccionario de americanismos*.[16] For now, the information regarding variety is specified as a 'Member note', as specified in the guidelines provided to annotators (Fernández-Alcaina et al., 2022).

[11] http://adesse.uvigo.es/data/corpus.php

[12] https://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl

[13] http://sfn.spanishfn.org/SFNreports.php

[14] http://spanishfn.org/corpus

4 Extending SynSemClass by Spanish

This section describes two phases of the annotation of Spanish verbal synonyms: (i) automatic data extraction (section 4.1) and (ii) manual data annotation (section 4.2).

4.1 Data extraction

The data extraction and preparation process consists of two phases: (i) automatic extraction of English-Spanish pairs from the corpus and (ii) data filtering.

In the first phase, candidate synonyms were extracted from the sentence-aligned corpus X-SRL (section 3.1). Pairs of Spanish-English were automatically extracted using the existing English Class Members as input. That is, for each English verb contained in SynSemClass, the Spanish counterpart attested in the corpus was extracted.

The dataset contained 39,279 sentences for each language. As an initial step, the extraction of synonym pairs was restricted to sentences containing the same number of verbs in both languages. The final dataset amounted to 21,551 sentences, i.e., 40,408 verbs. The number of different verbal types (after discarding wrongly-tagged elements) amounted to 1,715. For each sentence in each language, verbs were extracted as a list and paired to their translation counterparts according to index (e.g., $\text{Verb1}_{\text{en}} \rightarrow \text{Verb1}_{\text{spa}}$, $\text{Verb2}_{\text{en}} \rightarrow \text{Verb2}_{\text{spa}}$).

The second phase consists of two steps: (i) manual filtering of verbs and (ii) automatic filtering of the argument structures imported from AnCora (used as the source of valency frames).

Step 1: Manual filtering

The list of automatically paired verbs for each class was presented to annotators, who were asked to discard the verbs that did not belong to the class where they were automatically included. Discarded verbs were of two types:

- Wrongly-paired verbs during the automatic

[15] https://dle.rae.es/

[16] https://www.asale.org/damer/

extraction process (e.g., *seek-declarar*) due to mismatches in POS tagging or in word order.

- Verbs that were translation counterparts of a certain verb but that did not reflect the same meaning in that particular class (e.g., *solicitar* can be translated as *seek* but this is not the sense represented by class *hledat/search*, defined as 'A Seeker looks for a Sought entity'[17]).

Apart from labelling entries as belonging (or not) to a particular class, annotators were also asked to specify any restrictions applying to the inclusion of a verb in that particular class, e.g., if the verb is part of an idiomatic construction (e.g., *hacer erupción* 'erupt'). In this phase, annotators could also add any comments relevant for the annotation.

Based on a sample of 59 classes (51 verbs per class on average), it took approx. 30 minutes (on average) to filter one class. Out of the 3,016 lemmas initially included in the 59 classes, only 990 lemmas were kept (32% of the initial list). Inclusion of a verb by one annotator was enough to consider a verb a potential CM of a particular class. After annotating the first ten classes of the set, even if the initial list was considerably reduced, it was clear that the list obtained still contained a large number of verbs that did not belong to the class in which they had been included, thus slowing down the process of annotation.

Step 2: Automatic filtering of AnCora senses
As described in section 3.2, the AnCora lexicon links senses to several English resources, such as PropBank and VerbNet, two resources that are used in the English part of our lexicon. Using the links provided by AnCora, the list of potential CMs manually filtered in the previous step was filtered again to retrieve Spanish potential CMs for which:

- AnCora senses were linked to the same PropBank and VerbNet entries that the English CMs already contained in our lexicon, and
- no links were available in AnCora because the sense represented is not available in any of the English resources used.

The data annotation workflow is illustrated in Figure 2.

[17] https://lindat.cz/services/SynSemClass40/SynSemClass40.html

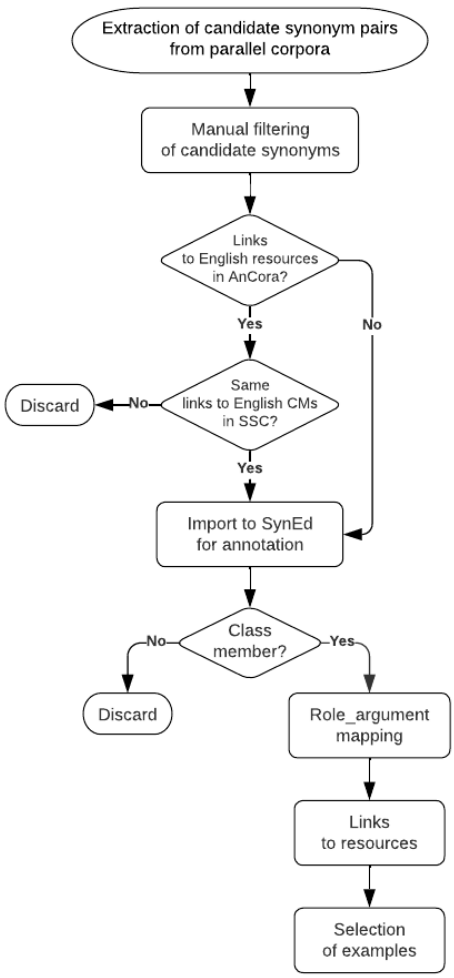

Figure 2: Data extraction and annotation workflow.

4.2 Data annotation

Data were annotated by three native Peninsular Spanish speakers fluent in English with similar backgrounds and previous experience working on a bilingual dictionary (English-Spanish). The three annotators also had a basic knowledge of German, which they can use to cross-check meanings. Annotators were given instructions on how to proceed before and during the process of annotation and they were also provided with annotation guidelines specifically designed for Spanish verbal synonyms (Fernández-Alcaina et al., 2022). The quality of annotators' work was tested on an initial set in which they were asked to annotate four classes.

After the first 40 classes, which were annotated by the three annotators, each set of classes is assigned to two annotators. Annotations are systematically monitored by one of the authors of this paper and unclear cases are discussed whenever necessary (section 4.2.2).

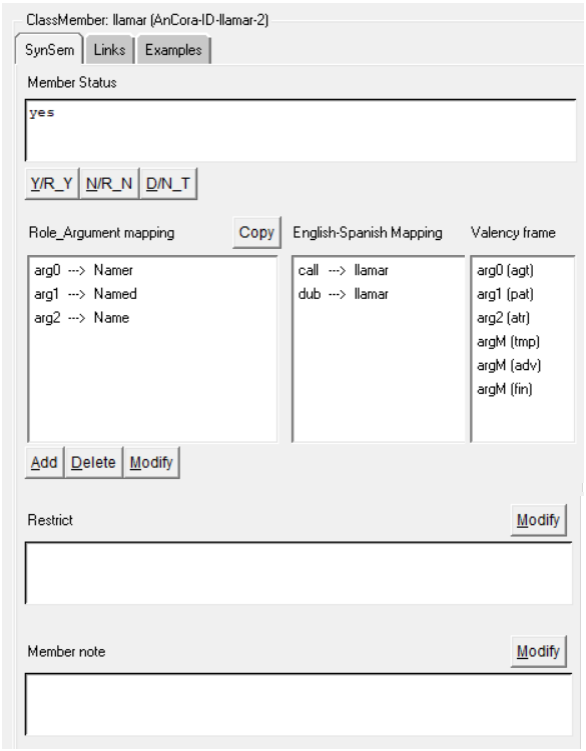

Figure 3: Role-Argument mapping for *llamar (AnCora-ID-llamar-2*) (class *call/nazvat/llamar*, ID vec00043).

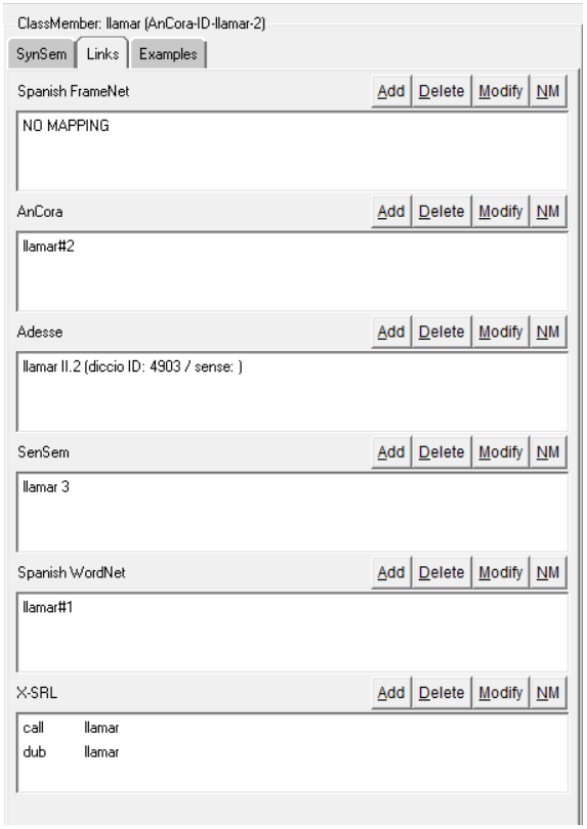

Figure 4: Selection of links to external resources for *llamar (AnCora-ID-llamar-2*) (class *call/nazvat/llamar*, ID vec00043).

The final complex annotation (role to argument mapping, external links, example selection, etc.) has been done using the SynEd editor (Urešová et al., 2018; Fučíková et al., 2023) available from the SynSemClass maintainers. However, as part of the task of adding Spanish, the data structure and the editor have been refactored to allow for more convenient and modular annotation for any number of languages. The technical details of these, however substantial, modifications are out of scope of this paper; please see (Fučíková et al., 2023) for the description of the modifications made to the SynEd editor.

For the candidate verbs retained after the filtering phase (section 4.1) and imported to SynEd, annotators were asked to provide fine-grained syntactic and semantic annotation by mapping the Roleset of the class with the valency frame of each verb, add links to external resources and select a set of representative examples from the corpus. To facilitate the work of annotators, SynEd provides both roles and class definitions in Czech, English, German and now also Spanish.

The process of annotation is divided into several interlinked steps (Figures 3, 4 and 5). In the first step, the task of annotators is to decide whether a candidate member matched the syntactic and semantic properties of the class. The annotation of Spanish synonyms–built upon existing synonym classes with Czech, English and in part German verbs–used semantic roles already defined for each class. The task of the annotators is thus to map the valency frame of the verb with each role in the existing Roleset associated with the given class (Figure 3). For example, based on the valency frame described for *llamar (AnCora-ID-llamar-2*) in AnCora and on the roles defined for class vec00043, the mapping is as follows: arg0→*Namer*, arg1→*Named* and arg2→*Name*. If a candidate verb is not included in AnCora, then it is imported to the editor using the label "SynSemClass-ID", as described in (Urešová et al., 2022) for German.

Since synonyms in SynSemClass are linked to external lexical resources (described in more detail in section 3.2), the second task in the annotation process consists in adding links to Spanish lexical resources (Figure 4). This is considered to be an essential step of the annotation process as linking the verbs in SynSemClass with other resources provides rich and comparable syntactic and semantic

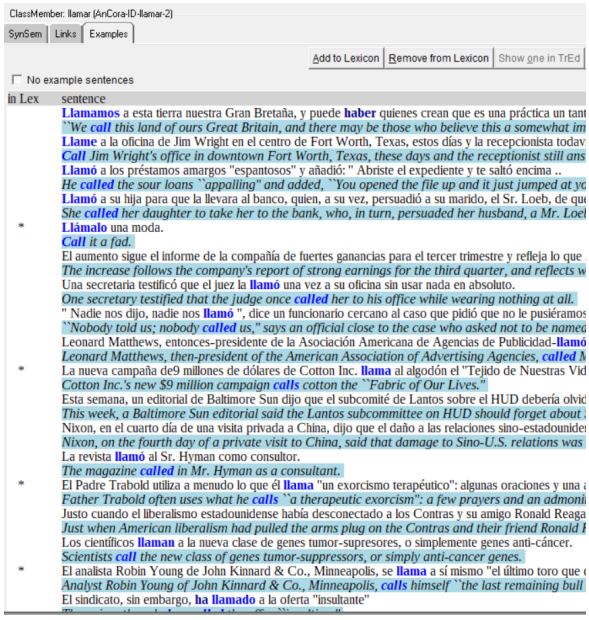

Figure 5: Selection of examples for *llamar (AnCora-ID-llamar-2*) (class *call/nazvat/llamar*, ID vec00043).

information about class members. Furthermore, linked data supports interoperability, thus adding value to the lexicon and enabling its use in Computational Linguistics.

In the last step of the annotation, annotators select a set of examples for Spanish and their English counterparts extracted from the corpus, if available (Figure 5).

4.2.1 Special cases

The annotation of certain cases demands special attention and requires a specific procedure:

- *Light Verb Constructions* (LVCs) (e.g., *hacer erupción*, 'erupt'). For LVCs, the nominal complement is considered to be part of the predicate and thus it is not mapped to any functor in the Role-Argument mapping but it is included as a restriction and a note specifying 'LVC' is added for easier identification. For example, the Roleset defined for the class *erupt/explodovat* (vec00018) contains only one Role (*Explosive*). The LVC *hacer erupción* has been included in the class by mapping A0 to the role *Explosive*, adding the noun complement ('erupción') as a restriction.

- Prepositional complements (e.g., *remontarse a*, 'date back'). Whenever necessary, the obligatory preposition introducing a prepositional complements is specified in the Role-Argument mapping. For the example above, the Role-Argument mapping is the following: arg0→*Entity*, arg1(**a**)→*Origin*.

4.2.2 Annotation monitoring

This section presents the criteria for the monitoring of the annotation in each step of the annotation. In cases of disagreement between annotators regarding the class membership, the verb is discarded from the class if:

- It expresses a general meaning that is better captured by a different class.

- There is a pattern in the inclusion of similar verbs whose meaning is deemed to be better captured by another class.

If there is a disagreement in the links to external resources, the rule of thumb is to reach a consensus between annotators so that only those links that are selected by the two annotators remain in the final annotation. Exceptions to this rule may occur, especially in the case of WordNet, where the selection of a sense is not as straightforward as in the rest of the resources and more variation is observed. It is up to the researcher then to decide if a sense must be included or not in the final annotation, for which the English equivalents provided may be of help.

In case of disagreement in the selection of examples, if no examples selected by the two annotators are available, the criteria are the following: (i) explicit argument realization (as much as possible), (ii) avoidance of highly specific or technical vocabulary, and (iii) preference for shorter sentences.

5 Results and limitations

The Spanish part of the SynSemClass lexicon is still in its early stages, but results have been encouraging so far even if the number of classes annotated represents a small part of the lexicon (72 classes, 5,200 verbs).[18]

The results obtained are especially relevant from a methodological perspective. While some changes in the tools used and/or in the annotation process can be expected as more data is processed, the results obtained so far set the basis for future work on Spanish. Since the addition of Spanish to SynSemClass was devised as an opportunity to "simulate" a scenario where a team works almost independently (that is, only with central support), the results obtained are also relevant for the future extension of the lexicon.

[18] As of January 2023.

Apart from the methodological aspects mentioned, adding a new language from a different linguistic subfamily enriches the existing lexicon by providing more linguistic evidence (including special cases, such as LVCs or prepositional complements) that led to a refinement of synonym classes. In order to accommodate new data, new classes will be added to the lexicon and it will be necessary to split or merged existing classes as more verbs are added to the lexicon.

Regarding Spanish and to the best of our knowledge, SynSemClass has become the first multilingual richly annotated resource of a general ontology type that includes Spanish. It is also the first one in linking various existing Spanish lexical resources, in line with other initiatives such as the UVI for English.

Even if theoretically feasible, including a new language in SynSemClass inevitably leads to certain issues that need to be addressed regarding technical, organizational and resource-related aspects, some of them being already tackled in (Urešová et al., 2022). In particular for Spanish, the main issues arising concern the limitations related to the resources available. While the Czech-English part of the lexicon relies on an annotated human-translated parallel dependency corpus with semantic information and on rich lexical resources, the comprehensiveness of the resources for Spanish is more limited as, to the best of our knowledge, no deeply syntactically annotated parallel corpus (similar to the PCEDT corpus) or bilingual verbal valency lexicon are available.

Another limitation is the scarce representation of dialectal varieties other than European Spanish. Although some of the resources (e.g., ADESSE or Spanish FrameNet) are not restricted to European Spanish, its coverage is uneven and entries do not contain specific information in this respect. To avoid this limitation where possible, verb senses from varieties other than European Spanish are included in SSC and specified in the lexicon based on the information provided by two dictionaries (with the limitations lexicographic resources entail).

6 Conclusions and future work

This paper has described the process of data processing and annotation and the initial results of adding Spanish to SynSemClass. Based on the method used for adding German class members to the resource, Spanish synonymous verbs have been extracted from a parallel corpus and linked to a set of lexical resources available. While part of the methodology for adding Spanish to the lexicon built on previous work on German, the specific features of the resources used for Spanish have required to make changes (some quite substantial, at least from the technological point of view) in the process of data extraction and adapt the tool used for annotation.

Spanish is one step more towards the creation of a collaborative multilingual event-type ontology. For the time being, plans in the near future include extending the lexicon to cover Korean. While the addition of German and Spanish in the lexicon will certainly provide the basis for the addition of more languages, it is assumed that both the lexicon and the tools will continually evolve to adapt to the intricacies of new languages.

From a more global perspective, this project is part of a larger early-stage project aimed at multilingual knowledge representation. SynSemClass classes will serve as the grounding for the events and states included in such 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 so far, a detailed specification of the process is still to be developed.

Acknowledgements

The work described herein has been supported by the Grant Agency of the Czech Republic under the EXPRO program as project "LUSyD" (project No. GX20-16819X) and uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (project No. LM2018101, supported by the Ministry of Education of the Czech Republic). The German part of the lexicon is partly supported by the grant Humane AI Network, funded by the EC by award No. 952026. We would like to thank the reviewers for their insightful comments and the annotators Cristina Lara-Clares and Alba E. Ruz for their work and invaluable input.

References

Alan Akbik, Xinyu Guan, and Yunyao Li. 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. The COLING 2016 Organizing Committee.

Laura Alonso, Joan Antoni Capilla, Irene Castellón, Ana Fernández-Montraveta, and Gloria Vázquez. 2007. The SenSem project: Syntactico-semantic annotation of sentences in Spanish. *Recent Advances in Natural Language Processing IV*, pages 89–98.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz, Leopoldo Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of ACL'2020*, pages 4555–4567.

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

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

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. *arXiv preprint arXiv:2010.01998*.

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

Cristina Fernández-Alcaina, Eva Fučíková, and Zdeňka Urešová. 2022. Annotation guidelines for Spanish verbal synonyms in the SynSemClass lexicon. Technical Report 72, UFAL MFF UK.

Ana Fernández-Montraveta and Gloria Vázquez. 2014. The SenSem corpus: An annotated corpus for Spanish and Catalan with information about aspectuality, modality, polarity and factuality. *Corpus Linguistics and Linguistic Theory*, 10(2):273–288.

Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Corpus-based multilingual event-type ontology: annotation tools and principles. Note = To be published at GURT/TLT, Wash., D.C,.

José M García-Miguel, Lourdes Costas, and Susana Martínez. 2005. Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. *Entre semántica léxica, teoría del léxico y sintaxis*, pages 373–384.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529, Istanbul, Turkey. European Language Resources Association (ELRA).

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2020. Prague dependency treebank - consolidated 1.0 (PDT-c 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková Razímová, and Zdeňka Urešová. 2006. *Prague Dependency Treebank 2.0*. LDC2006T01. LDC, Philadelphia, PA, USA.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68.

Howard Jackson. 1988. *Words and Their Meaning*. Routledge.

Jacqueline Kubczak. 2014. Valenzwörterbuch e-VALBU. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. 2017.

Valenční slovník českých sloves VALLEX. Nakladatelství Karolinum, Praha.

Markéta Lopatková, Václava Kettnerová, Anna Vernerová, Eduard Bejček, and Zdeněk Žabokrtský. 2020. VALLEX 4.0 (2021-02-12). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3524 and https://ufal.mff.cuni.cz/vallex/4.0.

John Lyons. 1968. *Introduction to Theoretical Linguistics*. Cambridge University Press.

John Lyons. 1995. *Linguistic Semantics*. Cambridge University Press.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Karin Kipper Schuler and Martha S. Palmer. 2005. *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, USA. AAI3179808.

Helmut Schumacher, Jacqueline Kubczak, Renate Schmidt, and Vera de Ruiter. 2018. *VALBU - Valenzwörterbuch deutscher Verben*. Narr, Tübingen.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht.

Carlos Subirats. 2009. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In *Multilingual FrameNets in Computational Lexicography*, pages 135–162. De Gruyter Mouton.

Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Jana Šindlerová. 2015. CzEngVallez – Czech–English Valency Lexicon.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018a. Creating a verb synonym lexicon based on a parallel corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1432–1437, Paris, France. European Language Resources Association.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. Synonymy in bilingual context: The CzEngClass Lexicon. In *Proceedings of The 27th International Conference on Computational Linguistics*, pages 2456–2469, Sheffield, GB. ICCL, ICCL.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018. Tools for building an interlinked synonym lexicon network. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020. SynSemClass Linked Lexicon: Mapping Synonymy between Languages. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography (LREC 2020)*, pages 10–19, Marseille, France. European Language Resources Association.

Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, and Jan Hajič. 2022. Making a semantic event-type ontology multilingual. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1332–1343, Marseille, France. European Language Resources Association.