sciendo
DOI 10.2478/jazcas-2023-0033

SPANISH SYNONYMS AS PART OF A MULTILINGUAL EVENT-TYPE ONTOLOGY

CRISTINA FERNÁNDEZ-ALCAINA – EVA FUČÍKOVÁ
– JAN HAJIČ – ZDEŇKA UREŠOVÁ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic

FERNÁNDEZ-ALCAINA, Cristina – FUČÍKOVÁ, Eva – HAJIČ, Jan – UREŠOVÁ, Zdeňka: Spanish Synonyms as Part of a Multilingual Event-Type Ontology. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 153 – 162.

Abstract: This paper presents an ongoing work on the multilingual event-type ontology SynSemClass, where multilingual verbal synonymy is formalized in terms of syntactic and semantic properties. In the ontology, verbs are grouped into synonym classes, both monolingually and cross-lingually. Specifically, verbs are considered to belong to the same class if they both express the same meaning in a specific context, and their valency frame can be mapped to the set of roles defined for a particular class. SynSemClass is built following a bottom-up approach where translational equivalents are automatically extracted from parallel corpora and annotated by human annotators. The task of the annotators consists in mapping the valency frame of a particular verb with the set of roles defined for the class where the verb is included as a potential class member, establishing links to external resources, and selecting relevant examples. The Spanish part of the ontology currently contains 257 classes enriched with Spanish synonyms. The resulting resource provides fine-grained syntactic and semantic information on multilingual verbal synonyms and links to other existing monolingual and multilingual resources.

Keywords: multilingual, ontology, semantics, valency, verbs

1 INTRODUCTION

This paper presents an ongoing work on the construction of a multilingual ontology for events. SynSemClass (SSC) is a multilingual lexicon of verbs organized into classes based on their semantic and syntactic properties. For our purposes, synonymy is defined in terms of contextual synonymy, i.e., a verb is considered a member of a class if it conveys the same or similar meaning expressed by other verbs within the same class, both monolingually and across languages. The construction of SSC involves fine-grained multilingual syntactic-semantic annotation as well as linking to several external resources in different languages (Czech, English, German, and Spanish so far). The information gathered in this lexicon facilitates cross-linguistic comparison, making it a valuable resource for linguistic research. Additionally, SSC also provides curated data useful for Natural Language

Processing tasks, such as cross-lingual synonym discovery, and is used for annotation of UMR (Unified Meaning Representation) (Bonn et al. 2023; Xue et al. 2023).

This paper is structured as follows: Section 2 briefly presents the ontology. Section 3 describes the method used for the extraction and filtering of Spanish candidates. The annotation process and an assessment of its quality are presented in Section 4, and the results obtained so far are described in Section 5. The conclusion and some plans for future work are summarized in Section 6.

2 THE ONTOLOGY

The organization of the ontology into classes revolves around the definition of synonymy as 'contextual synonymy' (Palmer 1981). That is, two or more verbs are considered synonyms (and thus members of the same class) if they express the same or similar meaning in the same context. Some aspects may need clarification: by 'verbs', we refer to verb senses, as we are dealing with cases of partial synonymy; by 'synonyms', we refer to both monolingual and cross-lingual synonyms, since the ontology is multilingual; and by 'context', we refer to the set of semantic roles expressed by the arguments and adjuncts of a verb, either explicitly or implicitly and with possible restrictions.

Therefore, a verb sense in any language is included in a specific class provided that each of the roles defined for the given class can be mapped to the verb valency slots captured in the valency frame. While total mapping between roles and arguments is a requirement, roles can be expressed by different morphosyntactic realizations and additional restrictions may apply, for example, regarding register or domain (Urešová et al. 2018b). Furthermore, each class member (CM) is linked to related entries in a set of preselected language-specific lexical resources, such as VerbNet (Schuler – Palmer 2005) for English, VALLEX (Lopatková et al. 2017; Lopatková et al. 2020) for Czech, and FrameNet des Deutschen (FdD) for German, among others.

The latest version of the ontology, SynSemClass4.0 (June 2022), contains 883 classes (approx. 6,000 CMs) in English and Czech, 61 of which are enriched with German synonyms.[1]

3 DATA PREPARATION

3.1 Main resources

Previous work on the extension of the ontology described a minimal set of resources required for the addition of new languages (Urešová et al. 2022). In particular, the necessary resources required are two: a parallel corpus and (at least) one lexical resource with information on verbal valency.

[1] The ontology is available for browsing and download at: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4746.

Regarding the corpus, the data for Spanish have been extracted from the X-SRL dataset (Daza – Frank 2020), a sentence-aligned parallel corpus containing approx. three million words for the English-Spanish part. The corpus is composed of texts from the Wall Street Journal section of the Penn Treebank and their Spanish translations. Although automatically translated, the quality of the translations was evaluated by human annotators with positive results (Daza – Frank 2020, p. 3909). Texts are tokenized, lemmatized, and POS-tagged.

Regarding the lexical resource, valency information was retrieved from AnCora-ES (Taulé et al. 2008). This verbal lexicon contains 2,820 lemmas (3,938 senses) and was built based on a corpus containing texts from a Spanish newspaper. One of the advantages of using AnCora is that, although monolingual, each sense is linked to several English resources that are also used in the English part of the ontology (specifically, VerbNet, PropBank, FrameNet, WordNet 3.0, and OntoNotes). This feature facilitates annotation process in two ways: first, it makes possible the automatic selection of senses to be imported to the tool used for annotation (see Section 4), thus restricting the number of candidates; second, it simplifies the process of determining verb class membership for human annotators as it offers comparable information between English and Spanish.

3.2 Candidate extraction

Candidate extraction is done in two phases: i) automatic extraction of English-Spanish pairs from the corpus, and ii) data filtering. An illustration of the workflow for data extraction is presented in Fig. 1:

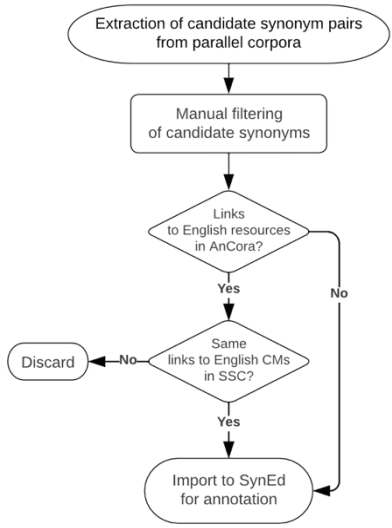

Fig. 1. Candidate extraction workflow

In the first phase, candidate pairs of synonyms are automatically extracted from the parallel corpus. The final dataset amounted to 40,408 verbs divided into 1,715 verbal types. Each Spanish verb in the dataset was paired with its possible English translation in the given context (e.g. *abrir-emerge*, *abrir-leave*, *abrir-come out*, etc.). To prevent an excessive number of irrelevant cases (i.e., incorrect pairings) from being imported into the tool for final manual complex annotation, data was prefiltered in two steps: by lemma and by sense. In the first step, Spanish-English pairs extracted from the corpus were filtered manually by annotators. For each class, annotators discarded the lemmas that did not belong to the class where they were automatically included if they were i) the result of errors during the automatic extraction processes; and ii) verbs that, despite being possible translations of the English verbs, do not express the meaning reflected by the class where they have been automatically included. Based on a sample of 59 classes, amounting to 3,016 verbs, the list of potential candidates was reduced by 68%, i.e., only 990 lemmas were retained for annotation. Although this step drastically reduced the size of the list, a second step in the prefiltering phase was added to restrict the number of candidates available for final annotation. To this aim and based on the information provided by AnCora, the candidates imported for final annotation were of two types: i) AnCora senses linked to the same links to PropBank and/or VerbNet to which English class members in SynSemClass are linked, and ii) AnCora senses with no links to other resources (as these usually represent multi-word expressions with no direct equivalent in English).

4 ANNOTATION

4.1 Annotation setup

The annotation was carried out by three native Spanish speakers who are proficient in English and trained in linguistics. Annotators were provided with annotation guidelines (Fernández-Alcaina et al. 2022) and trained on a preliminary set of classes. To ensure the quality of the annotations, each set of classes was processed by two annotators and the annotations were monitored by one of the authors of this paper. Any discrepancies that arose during the annotation process were discussed as needed. For the task of annotation, which involves mapping roles to arguments, identifying external links, and selecting examples, we used the SynEd editor (Urešová et al. 2018a; Fučíková et al. 2023), a tool specifically designed for this purpose by the SynSemClass maintainers and refactored to adapt it to any number of languages.

4.2 Class membership

The first task of the annotators involves determining whether a given verb belongs to a specific class. A verb is considered a member of a class if i) the meaning conveyed by the Spanish verb in a given context is the same or similar to that

conveyed by its English equivalent, and ii) it is possible to map each role defined in the Roleset for that class to the arguments defined for the verb in the valency frame. In addition to verbal lemmas, SSC includes multi-word expressions, such as idioms and light verb constructions (LVCs). Tab. 1 provides a simplified example of the role-argument mapping of a set of multilingual verbal synonyms in class vec00012 ('An Authority allows an Affected entity to engage in a Permitted entity').

	Authority	Permitted	Affected
allow (EN)	ACT	EFF	PAT
dovolit (CS)	ACT	EFF	ADDR
erlauben (DE)	VA0	VA1	VA2
permitir (ES)	arg0	arg1	arg2

Tab. 1. Role-argument mapping for class members in English, Czech, German, and Spanish in class vec00012 (simplified)

An evaluation of the annotations was performed to assess the quality of the annotated data. The sample used for this aim contains the last set of 42 classes annotated by two pairs of annotators: A1 vs A2 (21 classes, 1,099 verbs) and A1 vs A3 (21 classes, 998 verbs).

In the step of defining class membership, annotators could choose from five labels: 'yes', 'rather_yes', 'no', 'rather_no', and 'deleted'. To facilitate the interpretation of the results obtained, the five labels have been reorganized into two categories: 'yes' (including 'yes' and 'rather_yes') and 'no' (including 'no', 'rather_no', and 'deleted'). Tab. 2 presents the results for the agreement rate and Cohen's κ value (Cohen 1960).

	A1 vs A2	A1 vs A3
Agreement	91%	94%
Cohen's κ	0.62	0.75

Tab. 2. IAA results for Spanish verbal synonyms class membership

The results indicate a high level of agreement among the annotators in both cases (91% and 94%), but these may indicate a biased representation since i) the percentage of agreement between two annotators is expected to be high, and ii) the data distribution is highly biased towards the label 'no'. On the other hand, Cohen's κ is a more informative measure for this purpose as it can correct such bias. For the first pair of annotators (A1 vs A2), κ= 0.62, and the second pair of annotators (A1 vs A3), κ=0.75, thus indicating substantial agreement in both cases.

When compared to the initial set of classes, the results show a notable improvement across subsequent batches as the annotators become more familiar

with the task at hand. Tab. 3 compares the results obtained with the values for the first set of classes annotated (14 classes, 939 verbs). Specifically, for the pair of annotators A1 and A2, the agreement rate remains the same (91%), and κ increases from 0.49 to 0.62. Similarly, for the second pair of annotators, A1 and A3, there is a slight increase in the agreement rate (from 92% to 94%), as well as for κ values (from 0.54 to 0.75).

	A1 vs A2		A1 vs A3	
	First set	Last set	First set	Last set
Agreement	91%	91%	92%	94%
Cohen's κ	0.49	0.62	0.54	0.75

Tab. 3. IAA results for Spanish verbal synonyms class membership for the first set (14 classes) and the last set (21 classes) of classes

One possible explanation for the observed discrepancies in the results obtained for the two pairs of annotators A1 vs. A3 and A1 vs. A2 is that the latter seems to follow a more inclusive approach when determining class membership. This is not unexpected given the semantic complexity of the task.

4.3 Additional information

In the ontology, class members are linked to related entries in other external resources available for Spanish. Specifically, the resources used are two monolingual lexicons (ADESSE and Spanish SenSem), the Spanish version of FrameNet, and the Spanish WordNet 3.0 integrated into the Multilingual Central Repository.

1. ADESSE (García-Miguel et al. 2005) contains 3,400 lemmas extracted from the corpus ARTHUS (1.5 million words). The lexicon provides information regarding argument structure and semantic roles.
2. Spanish SenSem (Alonso et al. 2007) contains the most frequent 250 verbs from the SenSem corpus (Fernández-Montraveta – Vázquez 2014) built on texts from newspapers and literary sources.
3. Spanish WordNet 3.0 is integrated within the Multilingual Central Repository (Gonzalez-Agirre et al. 2012) together with six languages, including English. The MCR is also enriched with semantically tagged glosses and contains ontology information from WordNet Domains, Top Ontology, and AdimenSUMO.
4. Spanish FrameNet (Subirats 2009) contains 1,000 lexical units based on frame semantics and supported by corpus data. It provides syntactic and semantic information for each sense automatically annotated and validated by human annotators.

The resources selected complement the information provided by AnCora in different ways. Specifically, both ADESSE and SenSem provide definitions and

valency frames for each sense that make easier the task of the annotators. Spanish FrameNet uses the same frames used in the original English version, thus facilitating the task of the annotators and giving consistency to the annotation. Similarly, the Spanish WordNet 3.0 also provides consistency and facilitates annotation by integrating Spanish synonyms in multilingual synsets where English equivalent verbs are included.

The last step of the annotation consists in selecting relevant examples to illustrate the meaning of the class. Whenever possible, the argument structure defined for that verb sense must be explicitly realized in the examples.

5 RESULTS

The Spanish part of the SynSemClass currently contains 257 classes enriched with 1,400 Spanish verbal synonyms. It will be available in the next release, SynSemClass 5.0 (planned for 2023). Although classes with Spanish members still represent a small part of the total number of classes (29%), the results obtained so far are relevant for the development of the ontology in several aspects.

In terms of organization, Spanish has offered the opportunity to 'simulate' a scenario where an 'external' team works on the addition of a new language only with central support from the original maintainers.

From a methodological perspective, while the procedure followed for Spanish partly relies on previous work in German, it has been necessary to make some changes in the annotation tool and annotation process in other to adapt to some specifications of Spanish language. The results obtained so far may serve as the basis for future work in Spanish but also for other languages as the tool is expected to continue evolving to adapt to new languages.

In terms of multilinguality, adding a new language (and the first from a different family) contributes to enriching and refining synonymy classes in the ontology by providing linguistic evidence (including special cases, such as LVCs). With the addition of Spanish, it is already clear that as more data from more languages are added, classes will need to be hierarchized in the future (modified by splitting, merging, etc.).

Regarding Spanish resources, to the best of our knowledge, SSC has become the first multilingual richly annotated resource of a general ontology type that includes Spanish. It is also the first to link various existing Spanish lexical resources, in line with other initiatives such as the Unified Verb Index (UVI)[2] for English.

As for the limitations, the development of the ontology still heavily relies on manually processed data with respect to both data filtering and annotation.

[2] https://uvi.colorado.edu/

6 CONCLUSIONS AND FUTURE WORK

This paper has described the progress made so far in the addition of a new language to the multilingual event-type ontology SynSemClass. Although part of the method employed for the inclusion of Spanish is based on previous work in German, some steps needed to be added or modified in the data extraction and preparation phase to accommodate the specific features of the resources used. Similarly, the tool employed for annotation has also undergone some refactorization that allows to include a new language. In terms of the annotation process, some aspects of Spanish required a specific treatment, such as pronominal verbs or differences across geographical varieties. Furthermore, and specifically for Spanish, the result of adding Spanish is a twofold contribution in that it does not only enrich the ontology but also provides a resource that links data from several resources developed for Spanish that were independent up to now and complements them by including senses that are not captured in the resources available. The resulting resource has significant implications for both multilinguality and contrastive language research.

The addition of Spanish to the ontology is one step more towards the creation of a collaborative multilingual event-type ontology. From a more global perspective, plans in the near future include continuing work on the multilingual character of the ontology by adding more languages, which would necessarily imply that the lexicon and the tools continue evolving and adapting to the specification of new languages.

In the long term, the project is a part of a larger project for multilingual knowledge representation, where the SynSemClass classes will serve as a grounding for all events and states by relating all other entities in the resulting representation, which will also be grounded using other means. Although some verb annotation experiments have been done for the previous versions of the ontology, the full specification 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).

References

Alonso, L., Capilla, J. A., Castellón, I., Fernández-Montraveta, A., and Vázquez, G. (2007). The SenSem project: Syntactico-semantic annotation of sentences in Spanish. Recent Advances in Natural Language Processing IV, pages 89–98.

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

Bonn, J., Myers, S., Gysel van, J. E. L., Denk, L., Vigus, M., Zhao, J., Cowell, A., Croft, W., Hajič, J., Martin, J. H., Palmer, A., Palmer, M., Pustejovsky, J., Urešová, Z., Vallejos, R., and Xue, N. (2023). Mapping AMR to UMR: Resources for Adapting Existing Corpora for Cross-Lingual Compatibility. In Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 74–95, Washington, D.C. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), pages 37–46.

Daza, A., and Frank, A. (2020). X-SRL: A parallel cross-lingual semantic role labeling dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3904–3914. Online. Association for Computational Linguistics. Accesible at: https://aclanthology.org/2020.emnlp-main.321.

Fernández-Alcaina, C., Fučíková, E., and Urešová, Z. (2022). Annotation guidelines for Spanish verbal synonyms in the SynSemClass lexicon. Technical Report 72, ÚFAL MFF UK, 52 p.

Fučíková, E., Hajič, J., and Urešová, Z. (2023). Corpus-Based Multilingual Event-type Ontology: Annotation Tools and Principles. In Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023), pages 1–10, Washington, D.C. Association for Computational Linguistics.

García-Miguel, J. M., Costas, L., and Martínez, S. (2005). Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintáctico-semánticos en el proyecto ADESSE. In G. Wotjak – J. Cuartero Otal (eds.): Entre semántica léxica, teoría del léxico y sintaxis. Berlin: Peter Lang Verlag, pages 373–384.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (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).

Lopatková, M., Kettnerová, V., Bejcek, E., Vernerová, A., and Žabokrtský, Z. (2017). Valenční slovník českých sloves VALLEX. Nakladatelství Karolinum, Praha.

Lopatková, M., Kettnerová, V., Vernerová, A., Bejcek, E., and Žabokrtský, Z. (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. Accessible at: http://hdl.handle.net/11234/1-3524.

Palmer, F. R. (1981). Semantics. 2nd ed. Cambridge: Cambridge University Press.

Taulé, M., Martí, A., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), pages 96–101, Marrakech, Morocco. European Language Resources Association (ELRA).

Schuler, K. K., and Palmer, M. S. (2005). Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis. University of Pennsylvania, USA.

Subirats, C. (2009). Spanish FrameNet: A frame semantic analysis of the Spanish lexicon. In H. C Boas (ed.): Multilingual FrameNets in Computational Lexicography. Berlin/New York: De Gruyter Mouton, pages 135–162.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018a). 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).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018b). 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.

Urešová, Z., Zaczynska, K., Bourgonje, P., Fučíková, E., Rehm, G., and Hajič, J. (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.

Xue, N., Bonn, J., Cowell, A., Hajič, J., Palmer, A., Palmer, M., Pustejovsky, J., Sun, H., Urešová, Z., Wein, S., and Zhao, J. (2023). UMR Annotation of Multiword Expressions. In The 4th International Workshop on Designing Meaning Representation (DMR 2023), June 20, 2023, Nancy, France.