The SynSemClass lexicon: A resource for multilingual synonymy

This paper presents ongoing work on the multilingual lexicon SynSemClass (henceforth, SSC). Other related projects have addressed the issue of synonymy in multilingual contexts from various perspectives, e.g., WordNet (Vossen 2004; Fellbaum and Vossen 2012) or Predicate Matrix (López de Lacalle et al. 2016). However, to the best of our knowledge, only SSC formalizes multilingual verbal synonymy in terms of syntactic and semantic properties.

For our purposes, *synonymy* is understood as *contextual synonymy* (Palmer 1981) and *context* is defined in terms of the set of semantic roles expressed by the arguments and adjuncts of a verb, either explicitly or implicitly and with possible restrictions. Based on these criteria, verbs are grouped into synonym classes, both monolingually and cross-lingually. Specifically, verbs are considered to belong to the same class if they convey the same meaning in a specific context, i.e., if the valency frame defined for each verb can be mapped to the set of roles (i.e., Roleset) defined for a particular class (Table 1). The same verb can be classified into two different classes depending on the meaning expressed by their arguments (Table 2).

SSC is built following a bottom-up approach and data are linked to a set of external resources available for each language (e.g., VerbNet for English or E-VALBU for German, among others). Translational equivalents are automatically extracted from parallel corpora (e.g., ParaCrawl for German-English) and annotated by human annotators. All annotators are (near-) native speakers of one of the languages included and proficient in English. For each language, the same set of classes is processed by two annotators and their annotations are monitored by a researcher. The task of the annotators consists in:

i) 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,
ii) when available, establishing links to external resources, and
iii) selecting relevant examples.

The latest release, SynSemClass 4.0 (June 2022) (reference excluded due to anonymity) covering Czech, English, and German, contains 1,200 classes with approx. 9,000 CMs. The Spanish-English part of the lexicon (planned to be included in the fifth version, July 2023) contains 26 classes enriched by 2,358 Spanish class members (as of January 2023). Ongoing work is being done to include Chinese and Korean verbal synonyms in SSC.

The resulting resource has a twofold value: it provides fine-grained syntactic-semantic information on multilingual verbal synonyms at the same time it links data to other existing monolingual and multilingual resources. Although the number of classes and languages available in SSC is still limited, we believe that the resource can provide relevant data for descriptive and computational purposes as it may be used for cross-linguistic research on verbal valency as well as curated data for NLP tasks, such as cross-lingual synonym discovery.

References

Fellbaum, C., Vossen, P. 2012. Challenges for a multilingual wordnet. *Lang. Resources & Eval*uation, 46(2): 313–326. https://doi.org/10.1007/s10579-012-9186-z

López de Lacalle, M., Laparra, E., Aldabe, I. *et al.* 2016. Predicate Matrix: Automatically extending the semantic interoperability between predicate resources. *Language Resources & Evaluation* 50, 263–289. https://doi.org/10.1007/s10579-016-9348-5

Palmer, F. R. 1981. *Semantics*. 2nd edn. Cambridge University Press.

Vossen, P. 2004. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *International Journal of Lexicography*, 17(2): 161–173. https://doi.org/10.1093/ijl/17.2.161

Appendix

Table 1. A simplified version of the role-argument mapping of class *allow* for English, Czech, German and Spanish

	Authority	Permitted	Affected
allow	ACT	EFF	PAT
dovolit	ACT	PAT	ADDR
erlauben	VA0	VA1	VA2
permitir	arg0	arg1	arg2

Table 2. An example of a verb (*meet*) included in two classes based on the different semantic roles expressed by the arguments it takes

‘A Participant_1 meets a Partipant_2’		
	Participant_1	Participant_2
encounter	ACT	PAT
meet	**ACT**	**PAT**
sejít se	ACT	PAT
setkat se	ACT	PAT

‘A Cognizer gets to know a Person’		
	Cognizer	Person
aquaint	ACT	PAT
meet	**ACT**	**PAT**
poznat	ACT	PAT
seznámit se	ACT	PAT