Syntactic-Semantic Classes of Context-Sensitive Synonyms Based on a Bilingual Corpus

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič(✉)

Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Malostranské nám. 25, 11800 Prague 1, Czech Republic
{uresova,fucikova,hajicova,hajic}@ufal.mff.cuni.cz

Abstract. This paper summarizes findings of a three-year study on verb synonymy in translation based on both syntactic and semantic criteria and reports on recent results extending this work. Primary language resources used are existing Czech and English lexical and corpus resources, namely the Prague Dependency Treebank-style valency lexicons, FrameNet, VerbNet, PropBank, WordNet and the parallel Prague Czech-English Dependency Treebank, which contains deep syntactic and partially semantic annotation of running texts. The resulting lexicon (called formerly CzEngClass, now SynSemClass) and all associated resources linked to the existing lexicons and corpora following from this project are publicly and freely available. While the project proper assumes manual annotation work, we expect to use the resulting resource (together with the existing ones) as a necessary resource for developing automatic methods for extending such a lexicon, or creating similar lexicons for other languages.

Keywords: Verbal synonymy · Multilingual synonym Lexicon · Semantic lexicon · Valency · Language resources · Czech · English

1 Introduction

The goal of the project is to group verbs used as synonyms in Czech and English into (cross-lingual) synonym classes representing a cross-lingual meaning of the state or event expressed by the set of verbs assigned to that class. Each class thus represents a "concept" of an event or state type, expressed by a set of synonyms in that particular class, their syntactic behavior and their links to existing lexical-semantic resources.

For the purpose of this work, we use the term "synonym" in the "loose" interpretation [15], i.e., the necessary semantic equivalence takes also wider context into account.

The novel feature is the use of a richly annotated bilingual corpus to get more insight into the usage of verbs (together with their arguments) in translation. In the present paper, we have extended the discussion as presented in an

Z. Vetulani et al. (Eds.): LTC 2017, LNAI 12598, pp. 242–255, 2020.
https://doi.org/10.1007/978-3-030-66527-2_18

earlier version as published at the LTC 2017 conference [27], where the initial results have been discussed based on a sample of 60 classes manually processed and linked to the existing resources. The current size covers about 200 classes processed in several steps. The relevant features of the classes and their verbal members are also described.

While not being the goal of this very project, the ultimate use of such resource is both for followup linguistic studies and for use in natural language applications. The resulting lexicon, together with the existing resources to which it will be linked, will be used as a "gold standard" for evaluating automatic methods that should mimic the laborious manual work performed in this project (and possibly also as training data for systems based on deep learning, depending on its final extent). That way, it will serve as a seed resource for future, automatically extracted, cross-lingual lexicons with the same properties.

Since the publication at LTC 2017, partial findings have been published at various workshops, some results of which are summarized here too [28–33].

2 Resources Used

While the corpora (Sect. 2.2) are the basis for providing the lexicon annotators with examples of real use of the verbs in question, as well as material for automatic pre-extraction of the classes, the existing lexicons (Sect. 2.1) are used to link the new resource to, in various ways described later in the article.

2.1 Lexical Resources

The lexical resources used are of two types: the valency lexicons associated with previously annotated corpora, the Prague Dependency Treebank [10] and the Prague Czech-English Dependency Treebank [8], which serve as the sense identification source, and the external lexicons (VALLEX, FrameNet, VerbNet, OntoNotes, PropBank and WordNet), which are being linked to from the individual SynSemClass Entries.

The actual versions of these lexical resources are as follows:

– PDT-Vallex (Czech) is a Czech valency lexicon used for the annotation of the Prague Dependency Treebank [11] family of treebanks [9,25][1] This lexicon is also published as part of the Prague Dependency Treebank 2.0 by the Linguistic Data Consortium.[2] PDT-Vallex is based on the Functional Generative Description theory [24]. PDT-Vallex contains 7,121 verbs structured into 11,933 valency frames (roughly corresponding to verb senses), and the latest version is available as part of the PDT 3.5 distribution [10]. For a detailed information about the actual structure of the PDT-Vallex lexicons and its entries, see [25].

[1] https://lindat.mff.cuni.cz/services/PDT-Vallex/.
[2] http://www.ldc.upenn.edu/LDC2006T01.

- EngVallex (English)[3] is an English valency lexicon with 7,148 valency frames for 4,337 verbs, using the same valency framework as PDT-Vallex. It was built by an (largely manual) adaptation of the PropBank Lexicon [19] to the PDT labeling standards and principles [2].
- CzEngVallex (Czech-English) [6,26,34] is a Czech-English bilingual valency lexicon. It contains 20,835 explicitly linked verb senses (frame-to-frame pairs) and their aligned arguments (argument-to-argument pairs). It is linked, entry by entry and frame by frame, to the Prague Czech-English Dependency Treebank [7][4] and to the two monolingual valency lexicons mentioned above: PDT-Vallex and EngVallex.
- Berkeley FrameNet (English) [1,22] is a lexical database of English[5], containing about 13,000 word senses from more than 200,000 manually annotated sentences linked to more than 1,200 Semantic Frames. FrameNet is based on the Frame Semantics theory [5]; each lexical unit evokes a Semantic Frame (SF) which lists relevant Frame Elements (FEs), or Semantic Roles (SRs).
- OntoNotes [20,21,23] is a large-scale, multi-genre, multilingual corpus manually annotated with syntactic, semantic and discourse information. In the SynSemClass lexicon, we have used the OntoNotes Sense Groupings as published e.g., in the SemLink database (see below). They provide verb sense resource at a middle level of granularity. In these Groupings, each verb sense is identified and marked with a numeric index (1, 2, etc.). These indexes then serve as the distinguishing factor in the links from the individual class members in SynSemClass to the OntoNotes Grouping sense.[6]
- VerbNet (English) [3,13,23] is a class-based verb lexicon[7] with mappings to other lexical resources such as WordNet or FrameNet. VerbNet contains syntactic and semantic information on English verbs. It extended Levin [14] verb classes by refinement and addition of subclasses [13]. Each verb class is described by thematic roles, selectional restrictions on the arguments, and frames. Currently, VerbNet contains about 5,257 verb senses structured in 274 classes.
- PropBank (English) [19] is not only a lexicon but also a corpus[8] of one million words of English text, annotated with argument role labels for verbs (113,000 tokens, 3,324 frames files/types). Arguments are linked to their semantic roles [19].[9]

[3] http://hdl.handle.net/11858/00-097C-0000-0023-4337-2.

[4] PCEDT 2.0 is available from the LINDAT/CLARIAH-CZ repository at http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4).

[5] https://framenet.icsi.berkeley.edu.

[6] OntoNotes Sense Groupings are also viewable at the Unified Verb Index Reference Page at http://clear.colorado.edu/compsem/index.php?page=lexicalresources&sub=ontonotes.

[7] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

[8] http://propbank.github.io/.

[9] The PBs semantic roles are not the same as SynSemClass lexicon roles.

– SemLink (English) [18][10] links together different lexical resources (PropBank, VerbNet, FrameNet) through sets of mappings. The Semlink lexicon can be browsed online using the Unified Verb Index.[11]
– WordNet(s) [4,16] is a semantic network[12] of English. Words (nouns, verbs, adjectives, adverbs) are hierarchically grouped into sets of synonyms (117,000 "synsets"). Each synset contains word forms (referring to a given concept), a definition gloss and an example sentence. Czech WordNet 1.9[13] [17] will be used in future work when extending the classes on the Czech side.

2.2 Corpus Resources

The Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) [8] is a parallel treebank with over 1.2 million tokens in almost 50,000 sentences for each side. The PCEDT is based on the texts of the Wall Street Journal part of the Penn Treebank and their manual translations. Each language part is annotated in the Prague Dependency Treebank style, i.e., the annotation is dependency-style with argument structure of verbs (syntactic and semantic labeling), which corresponds to the associated valency lexicons for both languages: the PDT-Vallex (for Czech) and the EngVallex (for English); see Sect. 2.1.

In addition, we also use various monolingual corpora, such as the COCA corpus[14], corpora available in the SketchEngine[15] and corpora accessible and searchable through the KonText tool in the LINDAT/CLARIAH-CZ repository.[16]

3 Structure of the SynSemClass Lexicon

As the first thing in building the SynSemClass lexicon (called CzEngClass at that time), a structure of the lexicon (Fig. 1, from [27]) has been designed.

The SynSemClass lexicon builds upon the existing resources, as described above: CzEngVallex, PDT-Vallex and EngVallex lexicons and the PCEDT parallel corpus. In addition, the other lexicons listed (FrameNet, VerbNet, PropBank, OntoNotes and WordNet) are used as additional sources and links will be kept between their entries and the SynSemClass entries.

At the core of the SynSemClass lexicon, there are Synonym Classes, which are, for the purpose of this project, defined as (multilingual, or rather cross-lingual)[17] groups of verb senses (of different lexemes/words) that have the same meaning *and* the arguments of which can be mapped to a common set of semantic

[10] https://verbs.colorado.edu/semlink/.
[11] http://verbs.colorado.edu/verb-index/.
[12] https://wordnet.princeton.edu/.
[13] http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.
[14] http://corpus.byu.edu/coca/.
[15] https://www.sketchengine.co.uk/.
[16] http://lindat.mff.cuni.cz/services/kontext/.
[17] For the time being, bilingual: in Czech and English.

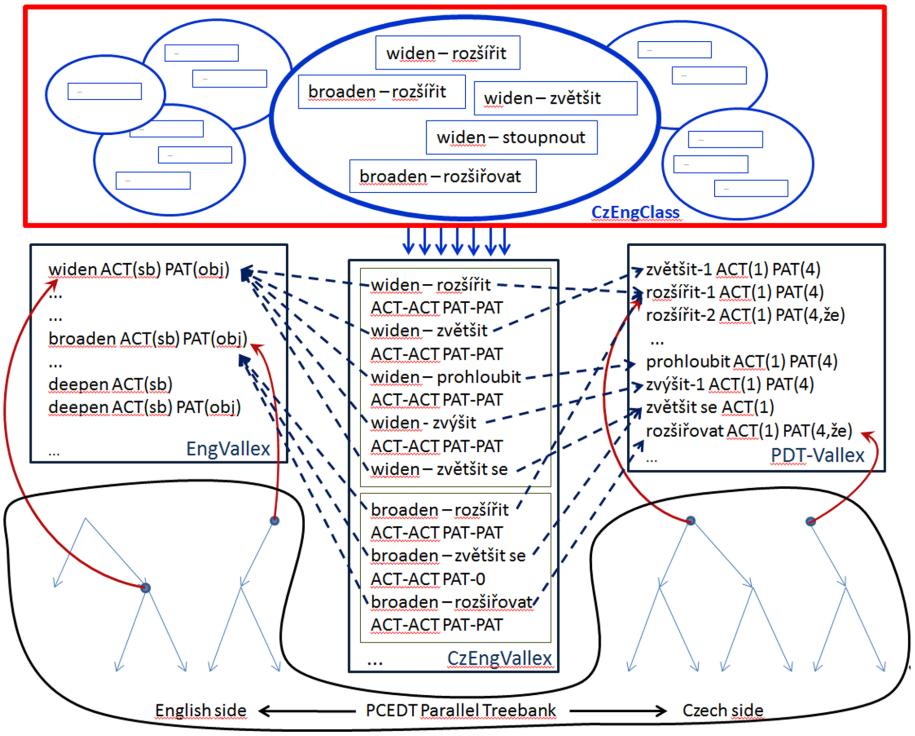


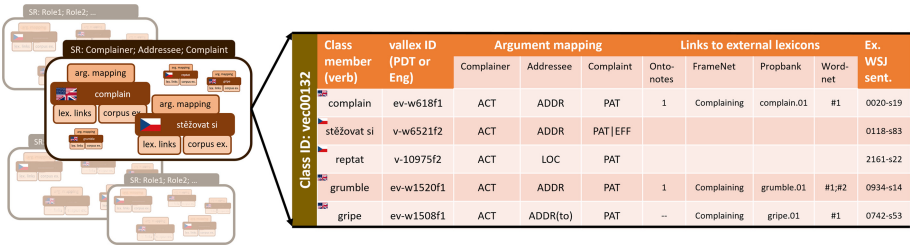
Fig. 1. SynSemClass lexicon & relation to core resources (Color figure online)

roles.[18] The term "same meaning" is understood with regard to a context, the relevant information about which is expected (at least in some cases) to be part of the argument mapping in the form of certain restrictions (lexical, syntactic, semantic) put on the arguments or even on a wider context. This is the case of most light verb constructions (*hold a meeting - meet*), idiomatic verbal MWEs (*cut loose - sever*), in some cases clear cases of hyperonymy (*pay [back] - repay*), and more (e.g., *return [call] - call back*), with clear patterns emerging.

In Fig. 1, the red box shows the SynSemClass proper, with its links to external lexicons. Below the red box, the resources directly used for its creation, extraction of examples, etc. are shown: in the middle, the CzEngClass lexicons links the original valency frame pairs, taken from the two monolingual valency lexicons, PDT-Vallex and EngVallex. These two lexicons are in turn linked to the parallel Czech-English corpus (see the bottom of the figure), the PCEDT 2.0.

Another view of the resulting lexicon in tabular form, depicting (a small sample of) one synonym class ("complain") is shown in Fig. 2 (taken from [33]).

[18] The term "sense" is used here for the differentiation of a single verb lexeme ("word") into one or more senses, represented technically by its valency frame ID, as it is done in the original valency lexicons (PDT-Vallex and EngVallex).



Class member (verb)	vallex ID (PDT or Eng)	Argument mapping			Links to external lexicons				Ex. WSJ sent.
		Complainer	Addressee	Complaint	Onto-notes	FrameNet	Propbank	Word-net	
complain	ev-w618f1	ACT	ADDR	PAT	1	Complaining	complain.01	#1	0020-s19
stěžovat si	v-w6521f2	ACT	ADDR	PAT|EFF					0118-s83
reptat	v-10975f2	ACT	LOC	PAT					2161-s22
grumble	ev-w1520f1	ACT	ADDR	PAT	1	Complaining	grumble.01	#1;#2	0934-s14
gripe	ev-w1508f1	ACT	ADDR(to)	PAT	--	Complaining	gripe.01	#1	0742-s53

Fig. 2. SynSemClass class example ("complain"), taken from [33]

In this example, the class "complain" (ID `vec00132`) contains five verbs, identified in turn by their PDT-Vallex or EngVallex frame IDs. For each of these verbs (verb senses), the argument mapping between their syntactic properties (the valency frame slots, labeled by the PDT/FGD-style arguments such as `ACT, PAT, EFF` and others) and the unique set of Semantic Roles defined for the whole class (here: `Complainer, Addressee, Complaint`). The argument mapping part is followed by links to external lexicons, for each class member separately, since the nuances in the meaning of the particular senses can lead to quite different targets in the external lexicons. While in this case, the FrameNet frame is the same for the three English members of the class, it often happens that several FrameNet classes are linked to from a single SynSemClass class. Finally, several example sentences from the PCEDT parallel corpus are attached too (here identified by their WSJ sentence IDs).

4 Data Preparation

The work so far has been done in several steps. First, we have randomly selected a set of 200 Czech verbs (verb senses) from three categories based on their frequency in the PCEDT corpus (high, medium, low). We have used the bilingual valency lexicon CzEngVallex [34] to determine a set of candidate English verbs for one synonym class, based simply on their pairings with the original Czech verb. Since CzEngVallex is linked to the PCEDT corpus, this gave us also a set of usage examples of these verb pairs, i.e., the context in which they have been used in the original English sentence and in its Czech translation.

For each of the English verbs in the candidate synonym group we have extracted links from the Unified Verb Index.[19] leading to PropBank, VerbNet, FrameNet and WordNet. These links (readily available to the annotators in the annotation software - cf. e.g., [31]) have been used for guiding the subsequent manual annotation pass.

There have been three manual annotation passes, gradually expanding the classes by new verbs, again taken from the PCEDT corpus using the parallel links in the CzEngVallex lexicon and the PCEDT corpus. After each expansion, a manual pass followed (Fig. 3), as described below.

[19] https://verbs.colorado.edu/verb-index.

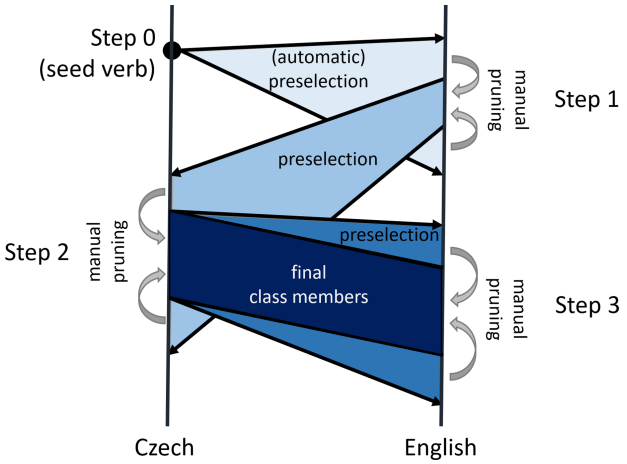


Fig. 3. Scheme of the data preprocessing and the manual annotation passes, taken from [33]

5 Annotation

5.1 Class Membership

At the beginning of each pass, including the initial one, the automatically preselected verbs (class members) had to be examined in order to determine if they actually are synonymous with each other. The reasons for exclusions have been numerous: bad automatic alignment in the corpus, very free or liberal translation leading to good sentence-level translation, but not word- (verb-) level translation, translation using hyperonyms or hyponyms, translation using negation of antonyms, etc.

Of course, the inclusion of any verb has been subject to further annotation, as described below; even if a verb has been retained after the first pruning based on intuitions about synonymy and verb senses, it could be excluded later, e.g., because argument mapping between valency slots and the Semantic Roles was deemed impossible.

5.2 Sense Determination

We have linked each English verb in the initial sample (in all three passes) to the OntoNotes sense as available from OntoNotes Groupings, in order to get more precise (even if not unique) links to the corresponding entry(ies), PropBank ID(s), VerbNet ID(s), FrameNet frame(s) and WordNet synset(s).[20] For example, for the English verb *set up* in a group extracted from the translation pairs linked to the Czech verb *budovat* (lit. *build*), OntoNotes sense No. 4 of `set-v` has

[20] Using the Unified Verb Index, http://verbs.colorado.edu/verb-index.

been assigned ("prepare (something) for a particular purpose"), attaching it to FrameNet frames ARRANGING and INTENTIONALLY_CREATE, VerbNet classes `braid-41.1.2` and `preparing-26.3-2`, PropBank IDs `set.03` and `set.08` and WordNet senses `6, 7, 21, 22, 25` (of *set*).

5.3 Common Semantic Roles (SRs)

In the next phase of the initial step,[21] we have devised a common set of SRs for each group (candidate synonym class) and mapped them to the valency frame slots from the PDT-Vallex and EngVallex lexicons for the Czech source verb and all the English verbs. In devising these roles, we have used FrameNet's Frame Element (FE) labels and descriptions as the initial pool of roles.[22] Many of the verbs (verb senses) in our candidate synonym sets have been found in FrameNet, so that we have started with the core FEs in the frame(s) associated with these verb senses (to be pruned later during a reconciliation phase also with VerbNet thematic roles). For example, for the *set up* example, we have first listed `Agent`, `Configuration` and `Theme`, `Creator` and `Created_Entity` from the ARRANGING and INTENTIONALLY_CREATE frames, finaly the SRs `Agent`, `Components` and `Created_Entity` are used; see Sect. 5.4. In Fig. 2, class "complain", the SRs are, as has already been mentioned, `Complainer`, `Addressee` and `Complaint`.

5.4 Argument - Role Mapping

For each verb in the candidate class, we have then paired each of the SRs to a valency slot as found in PDT-Vallex (for the Czech verb) and to EngVallex (for the English verbs). This has not, as expected, been straightforward, as will be described in more details in the next section. We have also used the other English resources to help clarify the relations if necessary. For example, some of the SRs initially listed have to be merged or deleted; in our "set up" example, it is clear that `Agent` and `Creator` are to be merged to one role.

6 Analysis of the Current Version

This section is based on a (sub)sample of 60 candidate synonym groups that have been created and mapped initially so far during the annotation process. More thorough examination will follow when all the 200 classes are fully checked.

[21] With possible modifications in the subsequent two steps.

[22] Using FrameNet v1.7, there are 1,168 different FE labels available across all frames. Later, VerbNet's thematic roles will be compared with the selected FEs and a common set used, provided a suitable common theoretical framework can be found.

6.1 Synonym Classes Composition

While the translation pairs extracted from the parallel corpus should have been clear synonyms, in some cases, even if the particular context has been taken into account, verbs had to be deleted from the group. For example, sometimes the parallel corpus correctly identified hyperonyms as translation equivalents, but there was no specific context that would restrict the hyperonym to the particular sense on the other side of the translation (this has happened in both directions):

- *That would hold.*PRED *spending.*PAT *on the program at about the previous year's level.*
- *To by znamenalo.*PRED *investice.*PAT *do programu přibližné ve stejné výši jako loni.*

In the above example, the Czech verb *znamenat* (form *znamenalo*) (in the sense of lit. *mean*, *imply*, *indicate*) is aligned with English *hold* (*... hold spending ... at about the same level*, Czech lit. *... mean spending ... etc.*). This is considered a functional equivalent translation in this context,[23] but since the context cannot be described just in terms of verb arguments, it has been decided to delete *hold* from this synonym class.

6.2 Roles and Argument Mapping

The initial set of roles for each class has been a union of FrameNet's Frame Elements (FEs) of all frames in which the appropriate English verbs have been found (see also Sect. 5.4). The goal was to establish a common set of roles for a given class, carefully considering both the FrameNet's FEs and the corresponding valency slots in the valency lexicons associated with the parallel corpus, including their use in the bilingual texts themselves.

Merging FrameNet-provided (or -inspired) SRs has been the most frequent operation, even within the same frame. For example, verbs inheriting from the STATEMENT frame might in some cases have merged the `Topic` SR (the subject matter to which the `Message` pertains) with the `Message` FE (what the `Speaker` is communicating to the `Addressee`)[24], since in our view, these typically occupy just one "slot" [12], with `Topic` being often part of the `Message`.[25] For example: *She said about her past* (`Topic`) *that it was wild* (`Message`) - *She said that her past was wild* (`Topic`+`Message`). Similarly, SRs differing only in animateness

[23] We can only speculate why the translator has used *znamenat* here; possibly because literal translations of *hold* are awkward in Czech (in this context), and the translator also determined that in fact the semantics of *hold* is already contained in the phrase *previous year's level*, and thus a translation of a hyperonym of *hold* can be used instead.

[24] https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Statement.

[25] In fact, the "Topic" part should be annotated within the information structure "layer" (topic/focus), not using semantic roles.

have often been merged, such as in the `Agent/Cause` case in the class represented by "widen" in Fig. 1.

The mapping of SRs to valency slots is mostly 1:1, as in the example of the synonym class corresponding most closely to the FrameNet's COMMERCE_PAY frame (Table 1): the arguments of *cover, pay, reimburse* are mapped 1:1 (if we tentatively add some of the missing ones into EngVallex, e.g., EFF/Effect to *reimburse* and EFF/Effect and BEN/Benefactor to *settle*). `Buyer` typically maps to the valency frame argument Actor, `Goods` to Effect, `Seller` to Addressee and `Money` to Patient.[26]

However, the correspondence of SRs and valency arguments is not necessarily always 1:1 - SRs have been occasionally merged (cf. `Topic` and `Message`) or split. For example, in BECOMING_AWARE, `Phenomenon` is mapped either to valency frame argument Effect or to valency frame argument Patient as shown in the following example *...to know details.*Effect *of one side only*, where `Phenomenon` is mapped to Effect and for another class member of the same class, the verb *hear*, `Phenomenon` is mapped to Patient: *she heard about the artery-clogging hazards.*Patient. Similarly, the mapping of `Goods` in the *pay* class is either to Patient (for *cover*) or Effect for PAY and other verbs; see also Table 1).

There are also examples with some specific context restrictions when a mapping can be applied. E.g., for the idiomatic verb *foot [the bill]* of the PAY synonym class (Table 1), the restriction to this idiomatic meaning (using *bill*) must be recorded. As another example, the Patient mapping of the verb *drill* in the BUILDING-related synonym class must be restricted to *drill a well (or other [large] hole-like thing)*. For light verb constructions, the nominal argument (labeled CPHR in the valency lexicons) often maps to the same role for the light verb argument as the Patient argument does for "non-light" verbs.

7 SynSemClass Size and Other Statistics

The current version of SynSemClass (version 2.0) contains 200 classes.[27] As of end of April, 2020, it contains 200 classes, which has been processed at least by 2 steps of the manual annotation (or better to say, they also contain additional annotation so that they are close to be through all three steps). Finished are 157 classes with 1567 Czech and 2836 English verbs (both relatively small as well as quite rich classes, e.g., "build", "wait", "allow", "learn", "invest", "think",

[26] Later investigation, as well as testing the addition of new verbs into this class, however revealed that most of the verbs should have had only three valency slots in EngVallex: ACT, PAT and ADDR, where PAT corresponds to either "payment" or "obligation", but not both. In addition, the EFF does not seem to be core for the "reimbursement" concept. Therefore the Roleset has been reduced (or, generalized) to three SRs only, namely `Payee` (mapped to ADDR), `Obligation_Payment` (mapped to PAT or EXT) and `Payer` (mapped to ACT). In any case, this is still an example of a prevailing 1:1 valency slot:SR mapping.

[27] http://hdl.handle.net/11234/1-3215; previous version was available as SynSemClass 1.0, http://hdl.handle.net/11234/1-3125.

Table 1. Argument mapping for PAY class

Roles				
	`Buyer`	`Goods`	`Seller`	`Money`
Hradit	ACT	EFF	ADDR	PAT
Cover	ACT	PAT	BEN?	MANN
Foot	ACT	DPHR(*bill*)	BEN?	MANN
Pay	ACT	EFF	ADDR	PAT
Reimburse	ACT	EFF?	ADDR	PAT
Settle	ACT	EFF?	BEN?	PAT

etc.). In the last phase of step 3 annotation (i.e., almost finished) are another 37 classes, with 404 Czech and 208 English verbs, such as "hold", "speak", "watch", "announce", etc.). The remaining six classes from the 200 have been in fact merged with others, since after the expansion and annotation, they grew to be in fact identical with another class (created initially from a different seed verb).

These numbers are summarized in Table 2.

Table 2. SynSemClass 2.0 size and statistics

	In Step 3	Finished	Merged	Total
Classes	37	157	(6)	194
Czech verbs	404	1567	(110)	1971
English verbs	208	2836	(92)	3044

8 Conclusions and Next Steps

We have described some findings on synonymy of verb senses of generally different verbal lexemes in a bilingual setting, and specifically, we focused on their valency behavior and common Semantic Roles. Our future research will be aimed at extending it to more verbs, at further refinement of our semantic roles and their explicit mappings from valency arguments to the semantic roles, and at formalizing the additional restrictions. We will analyze in more detail the relation of valency and semantic roles also from their morphosyntactic realization point of view. We will also confront the findings as supported by the corpus material to the underlying theoretical framework(s), in order to possibly refine them in their approach to verb sense distinctions, valency and argument description. We will also compare our results with automatic approaches to cross-lingual semantic similarity detection, such as in [35], which is very much related to our work.

Finally, the resulting lexicon is now openly available in the LINDAT/CLARIAH-CZ repository at http://lindat.cz[28].

Acknowledgments. This work has been supported by the grants No. GA17-07313S and GX20-16819X of the Grant Agency of the Czech Republic, and it uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure, project No. LM2018101, supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL 1998, vol. 1, pp. 86–90. Association for Computational Linguistics, Stroudsburg (1998). https://doi.org/10.3115/980845.980860. http://dx.doi.org/10.3115/980845.980860
2. Cinková, S.: From propbank to engvallex: adapting the propbank-lexicon to the valency theory of the functional generative description. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2170–2175. ELRA, Genova (2006)
3. Duffield, C.J., et al.: Criteria for the manual grouping of verb senses. In: Proceedings of the Linguistic Annotation Workshop, LAW 2007, pp. 49–52. Association for Computational Linguistics, Stroudsburg (2007). http://dl.acm.org/citation.cfm?id=1642059.1642067
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication. MIT Press, Cambridge (1998)
5. Fillmore, C.J.: Frame semantics and the nature of language. Ann. New York Acad. Sci.: Conf. Origin Dev. Lang. Speech **280**(1), 20–32 (1976)
6. Fučíková, E., Hajič, J., Šindlerová, J., Urešová, Z.: Czech-English bilingual valency lexicon online. In: 14th International Workshop on Treebanks and Linguistic Theories (TLT 2015), pp. 61–71. IPIPAN, Warszawa (2015)
7. Hajič, J., et al.: Announcing Prague Czech-English dependency treebank 2.0. In: Proceedings of the 8th LREC 2012), pp. 3153–3160. ELRA, Istanbul (2012)
8. Hajič, J., et al.: Prague Czech-English dependency treebank 2.0 (2012). https://catalog.ldc.upenn.edu/LDC2004T25. http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4
9. Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P.: PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In: Nivre, J., Hinrichs, E. (eds.) Proceedings of The Second Workshop on Treebanks and Linguistic Theories. Mathematical Modeling in Physics, Engineering and Cognitive Sciences, vol. 9, pp. 57–68. Vaxjo University Press, Vaxjo (2003)
10. Hajič, J., et al.: Prague dependency treebank 3.5 (2018). http://hdl.handle.net/11234/1-2621
11. Hajič, J., et al.: Prague dependency treebank 2.0 (2006)
12. Kettnerová, V.: Konstrukce s rozpadem tématu a dikta v češtině (constructions with topic and message separation in Czech). Slovo Slovesnost **70**(3), 163–174 (2009)

[28] http://hdl.handle.net/11234/1-3215.

13. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending VerbNet with novel verb classes. In: Proceedings of LREC, p. 1 (2006)
14. Levin, B.: English verb classes and alternations. The University of Chicago Press, Chicago and London (1993)
15. Lyons, J.: Introduction to Theoretical Linguistics. Cambridge University Press, Cambridge (1968)
16. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748. http://doi.acm.org/10.1145/219717.219748
17. Pala, K., Smrz, P.: Building Czech Wordnet. Roman. J. Inf. Sci. Technol. **7**(1–2), 79–88 (2004). http://nlp.fi.muni.cz/publications/romjist2004_pala_smrz/
18. Palmer, M.: Semlink: linking PropBank, VerbNet and FrameNet. In: Proceedings of the Generative Lexicon Conference, pp. 9–15 (2009)
19. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (2005). https://doi.org/10.1162/0891201053630264. http://dx.doi.org/10.1162/0891201053630264
20. Pradhan, S.S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: OntoNotes: a unified relational semantic representation. Int. J. Semant. Comput. **01**(04), 405–419 (2007). https://doi.org/10.1142/S1793351X07000251. https://www.worldscientific.com/doi/abs/10.1142/S1793351X07000251
21. Pradhan, S.S., Xue, N.: 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, pp. 11–12. Association for Computational Linguistics, Boulder, May 2009. https://www.aclweb.org/anthology/N09-4006
22. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: extended theory and practice. Unpublished Manuscript (2006). http://framenet.icsi.berkeley.edu/
23. Schuler, K.K.: VerbNet: a broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania (2006). http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf
24. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. D. Reidel, Dordrecht (1986)
25. Urešová, Z.: Valence sloves v Pražském závislostním korpusu. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia (2011)
26. Urešová, Z., Fučíková, E., Hajič, J., Šindlerová, J.: Czengvallex - Czech English valency lexicon (2015)
27. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: Syntactic-semantic classes of context-sensitive synonyms based on a bilingual corpus. In: Vetulani, Z., Mariani, J. (eds.) Proceedings of 8th Language and Technology Conference, pp. 201–205. Fundacja Uniwersytetu im. Adama Mickiewicza, Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań (2017)
28. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: Creating a verb synonym lexicon based on a parallel corpus. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, May 2018
29. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: A Cross-lingual synonym classes lexicon. Prace Filologiczne **LXXII**, 405–418 (2018)

30. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: Defining verbal synonyms: between syntax and semantics. In: Haug, D., Oepen, S., Øvrelid, L., Candito, M., Hajič, J. (eds.) Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), pp. 75–90. Universitetet i Oslo, Linköping University Electronic Press Pub No. 155, Linköping (2018)
31. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: Tools for Building an Interlinked Synonym Lexicon Network. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, May 2018
32. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: Meaning and semantic roles in CzEngClass lexicon. Jazykovedný časopis/J. Linguist. **70**(2), 403–411 (2019)
33. Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J.: SynSemClass linked lexicon: mapping synonymy between languages. In: Proceedings of the Globalex 2020 Workshop at the 12th International Conference on Language Resources and Evaluation (LREC 2020). European Language Resources Association (ELRA), Marseille, May 2020
34. Urešová, Z., Fučíková, E., Šindlerová, J.: CzEngVallex: a bilingual Czech-English valency lexicon. Prague Bull. Math. Linguist. **105**, 17–50 (2016)
35. Wu, S., Choi, J.D., Palmer, M.: Detecting cross-lingual semantic similarity using parallel PropBanks. In: Proceedings of the 9th Conference of the Association for Machine Translation in the Americas, AMTA 2010, Denver, CO (2010). https://amta2010.amtaweb.org