Parallel Dependency Treebank Annotated with Interlinked Verbal Synonym Classes and Roles

Zdeňka Urešová **Eva Fučíková** **Eva Hajičová** **Jan Hajič**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranske nam. 25
11800 Prague, Czech Republic
{uresova,fucikova,hajicova,hajic}@ufal.mff.cuni.cz

Abstract

We present an ongoing project of enriching an annotation of a parallel dependency treebank, namely the Prague Czech-English Dependency Treebank, with verb-centered semantic annotation using a bilingual synonym verb class lexicon, CzEngClass. This lexicon, in turn, links the predicate occurrences in the corpus to various external lexicons, such as FrameNet, VerbNet, PropBank frame files, OntoNotes, and WordNet. We briefly describe the content of the CzEngClass synonym class lexicon and then we focus on its use for an enrichment of corpus annotation, which proceeds in two steps - automatic preprocessing and manual correction. This paper describes a first milestone of a long-term project; so far, approx. 100 CzEngClass classes, containing about 1800 different verbs each for both Czech and English, are available for such annotation. The corpus coverage at the moment is about 50%, allowing us to extract some basic statistics and discover a set of issues that appeared during the annotation process. The ultimate goal is to have a high-coverage, multilingual verbal synonym lexicon and corpora with all events annotated by such lexicon, to serve both theoretical studies in lexical semantic, translatology, corpus annotation studies etc. as well as a usable resource for training automatic semantic text processing systems for event/participant detection and linking and for general information extraction.

1 Introduction

While there are various richly annotated corpora linked to lexicons for several languages, such as OntoNotes (Pradhan et al., 2007) or the Prague Dependency Treebank projects (Hajič et al., 2006; Hajič et al., 2018), there are only a few that link substantial amount of annotated material to semantic lexicons, such as FrameNet, VerbNet, SemLink, PropBank or WordNet, and to our knowledge none that would link to all of them within a single corpus.

The project presented in this paper aims at filling this gap. The aim is to create a richly annotated corpus where each occurrence of a verb, for example *say*, is annotated by a (bi-lingual) synonym class *say, tell, disclose, report, ..., říci, sdělit, uvést, ...* and its dependents in the semantic representation (regardless of their syntactic realization) are labeled by semantic roles assigned to that class (Speaker, Addressee, Information).

Such a resource can be divided into two components:

- a semantically oriented bi- or multilingual verbal synonym lexicon, linked to all the other lexical resources, and
- the richly annotated corpus that contains references to entries in this semantic lexicon at every content verb (predicate) in the corpus.

The first component is covered by the existing CzEngClass lexicon[1] (Urešová et al., 2018c) which, while not complete and covering only Czech and English at this time, already provides enough synonym classes (and promises more coverage in the future) to approach the annotation task (the 2^{nd} point above).

In this paper, we start with a short description of the resources used directly or indirectly through the available lexicon and corpus (Sect. 2), then we show how we have proceeded with the annotation process

[1] http://hdl.handle.net/11234/1-2977.

(Sect. 3) and present the basic statistics of the automatic part of the annotation part of the process as applied to the whole corpus (Sect. 4). Manual corrections performed on a sample of 100 verb-pair occurrences are described in Sect. 5, giving a first glimpse of the effort needed to complete the manual part of the annotation for the whole corpus. Sect. 6 describes some lessons learned, summarizes the findings and provides some hints for future work.

2 Original Resources Used

The main resource is the CzEngClass lexicon. Its entries serve as the target of reference links attached to verb occurrences in the corpus (Urešová et al., 2018a).

2.1 The Lexicons

2.1.1 The CzEngClass Lexicon

The CzEngClass lexicon has the following structure (Fig. 1):

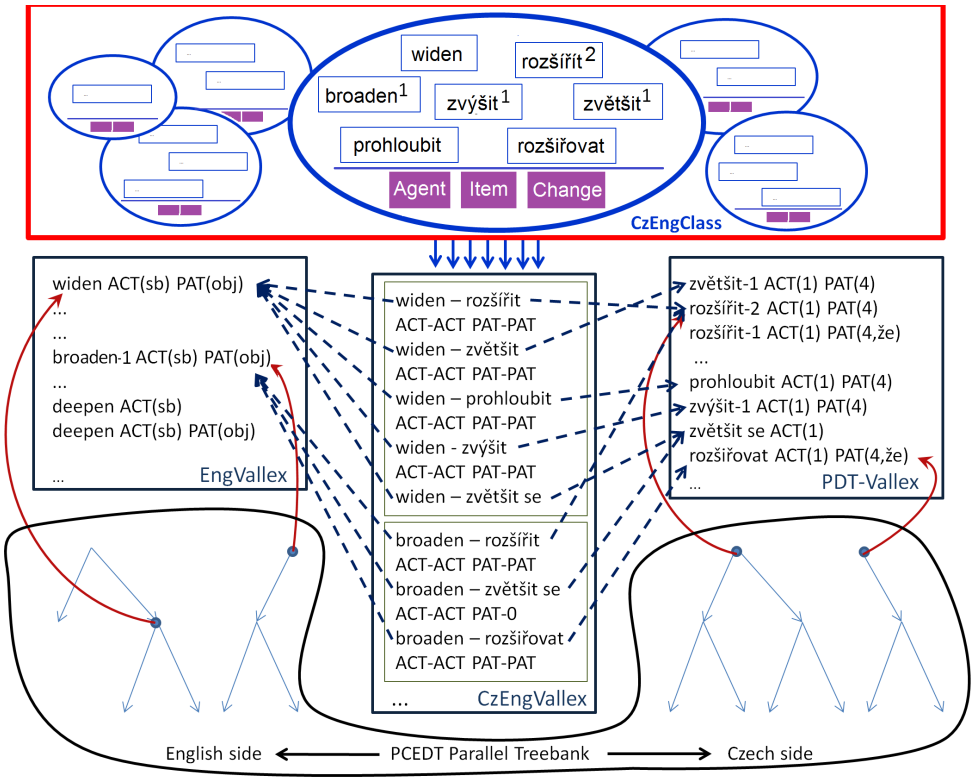

Figure 1: CzEngClass lexicon & other resources (from (Urešová et al., 2018a))

The lexicon consists of cross-lingual synonym classes which group verb senses (of different verbs) that have the same or similar meaning[2] *and* the (valency) arguments of which can be mapped to a common set of semantic roles, called a *Roleset*. The semantic roles (SRs) are assigned to all the members of one synonym class, and mapped individually to their valency arguments as captured for those verb senses in the EngVallex (Cinková et al., 2014) and PDT-Vallex (Hajič et al., 2003) lexicons.[3] In Fig. 1, the lexicons are depicted as the square boxes on the left and right, just below the CzEngClass core lexicon depiction on top in the big rectangular box, where each oval shows one synonym class; the purple rectangular boxes on the bottom of one class present the common set of semantic roles for that class: Agent, Item, Change).

[2]The notion of "same or similar meaning" is used here rather intuitively, as the understanding of "synonymy" itself varies quite substantially. However, adding the mapping condition between roles and arguments helps to use more substantiated and evidenced criterion for deciding which verbs belong to the particular class.

[3]The valency lexicons use labels like ACT for Actor – the first argument, and PAT for Patient – the second argument, (Sgall et al., 1986).

The SRs are simultaneously linked to the existing verbal pairings (translational equivalents) found in the CzEngVallex lexicon (Urešová et al., 2016). CzEngVallex (the large box in the middle of Fig. 1) is in turn linked to the PCEDT parallel corpus (see the bottom of Fig. 1). In addition, CzEngClass entries refer also to several existing semantic lexicons (Sect. 2.1.2). More details on the mapping of SRs to valency slots of the corresponding valency lexicon entries are presented later in this paper, with examples in Tables 1 and 2.

The examples show some of the basic properties of the entries in the CzEngClass lexicon, and illustrate some specific issues that had to be dealt with.

Class: *surprise – překvapit*

	Roleset (semantic roles)	
Class Member (sense ID)	Experiencer	Stimulus
surprise (EngVallex-ID-ev-w3269f1)	PAT	ACT and/or MEANS
překvapit (PDT-Vallex-ID-v-w4862f1)	PAT	ACT and/or MEANS
ohromit (PDT-Vallex-ID-v-w3015f1)	PAT	ACT and/or MEANS

Table 1: Example verbal synonym class for *surprise – překvapit* with role mappings (simplified)

In Table 1, a relatively "regular" (and small) synonym class is presented. This class (*surprise – překvapit*) bears two semantic roles: Experiencer and Stimulus. These roles are mapped, for each class member, to the valency slots associated with the individual verbs in the PDT-Vallex and EngVallex valency lexicons. In this class, the mapping is the same for all verbs in the class. We can also observe here a non-trivial (non-1:1) mapping, namely that the Stimulus is not always expressed by an ACTor alone. For example, in *Mr. X.* ACT *surprised Mr. Y.* PAT *by claiming.*MEANS *the prize for himself.*, the role Stimulus is formed by joining of both Mr. X and the "claiming event" (for which Mr. X is in fact the ACTor).[4]

In Table 2, the class *decline – odmítnout* exposes another frequent phenomenon, namely that the same role is mapped to different valency slots with different class members (ADDR vs. ORIG; in other classes, frequent pairs mapped to the same role are DIR1 and ORIG or ACT and LOC). This is mostly caused by the principles and conventions used in the underlying FGD valency theory (Sgall et al., 1986; Panevová, 1974), as reflected in the valency lexicons. At the same time, it exemplifies that *from the semantic perspective*, the valency slot labeling as determined by the rules and conventions of the FGD theory (or any other valency theory, for that matter) is not crucial, since the mapping provides the flexibility to relate them to the right semantic role(s). This example also shows that some verbs are included in that class, even if they do not express some semantic role from the Roleset by using a clearly assigned valency slot or any other modifier. Such a role (here, the Proposer) is deemed to be necessarily understood from a wider context.[5] For example, *deny* (more precisely, its sense identified by EngVallex-ID-ev-ev-w876f1) and *refuse* (EngVallex-ID-ev-w@2598f1) both being without an obligatory ADDR, display this behavior. The (semantic) complementation assigned to the role of Proposer is for such cases marked as #sb("somebody"), and it is assumed that in the process of corpus annotation, it will be inserted to the resulting representation and connected by a (semantic) co-reference link to the actual Proposer. The last column (Restrictions or requirements) of the Table 2 may contain additional requirements on the class member or its mapping to be valid, such as negation (*přijmout* in Czech means *accept*, cf. last line of Table 2).

2.1.2 External Lexicons

The CzEngClass lexicon, as mentioned above, indirectly refers each verb at each entry to the following external lexicons (Urešová et al., 2018a):

- The Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2006; Fillmore, 1976; Fillmore, 1977), a lexical database of English,[6]

[4]This example does not claim anything special about (non-)agentive subjects - it rather shows that mapping between SRs and valency slots does not have to be necessarily 1:1.

[5]For such cases, CzEngClass uses specific pseudo-functors, such as #any,#sb, and #sth.

[6]`https://framenet.icsi.berkeley.edu`

Class: *decline – odmítnout*

Class Member (sense ID)	Roleset (semantic roles)			Restrictions
	Authority	Proposer	Proposal	or requirements
decline (EngVallex-ID-ev-w829f1)	ACT	ADDR	PAT	
deny (EngVallex-ID-ev-ev-w876f1)	ACT	#sb	PAT	
deny (EngVallex-ID-ev-ev-w876f2)	ACT	ADDR	PAT	
odmítnout (PDT-Vallex-ID-v-w2785f1)	ACT	#sb	PAT	
refuse (EngVallex-ID-ev-w@2598f1)	ACT	#sb	PAT	
přijmout (PDT-Vallex-ID-v-w5161f3)	ACT	ORIG	PAT	negation

Table 2: Example class for *decline – odmítnout* with role mappings (simplified, from 24 verbs)

- VerbNet (Schuler, 2006; Duffield et al., 2007; Kipper et al., 2006), a class-based verb lexicon[7] with syntactic and semantic information of English verbs,
- PropBank (Palmer et al., 2005), linked to the OntoNotes corpus,[8]
- SemLink (Palmer, 2009)[9] which connects the above lexicons, and
- WordNet (Miller, 1995; Fellbaum, 1998).[10]

The external links allow to compare and use these lexical resources with the annotated corpus, but for the proper semantic annotation are not used except as a source of secondary knowledge for the annotator when making annotation decisions.[11]

2.2 The PCEDT Corpus

For this project, the Prague Czech-English Dependency Treebank (PCEDT), as available from LDC[12], described in (Hajič et al., 2012), is used. It contains about 55,000 sentences on each language side. The English side is the Wall Street Journal part of the Penn Treebank and the Czech side is its translation. Both sides are annotated for the so-called *Tectogrammatical Representation* (TR), used for the Prague Dependency Treebank family of projects (Hajič et al., 2006). Most importantly, content verbs are annotated by their corresponding valency lexicon entries as captured in the PDT-Vallex lexicon (Urešová et al., 2014) on the Czech side and in the EngVallex (Cinková et al., 2014; Cinková, 2006) on the English side.

As described in (Urešová et al., 2018a) (and in Sect. 2.1 above), the PCEDT corpus has been used as a source information for building the CzEngClass lexicon, raising the question of why the annotation cannot be fully deterministic. However, the classes are in principle independent of the PCEDT data and they underwent manual pruning; it is thus likely we get ambiguous (or even no) annotation by simply following the links from the CzEngClass entries directly to the corpus. In any case, the coverage of the CzEngClass entries will be relatively high, since they have been extracted from the same corpus in the first place.

3 The Annotation Process

3.1 Data Structure for Added Node Attributes (Technical Description, for Reference)

For the annotation process, we have used the valency reference IDs of individual verbs captured in both appropriate valency lexicons and also in the CzEngClass classes. In the PCEDT corpus, both on the Czech as well as English side, each occurrence of a content verb is annotated with such a valency frame ID. It is thus straightforward to "inverse" the mapping automatically, and include a reference to the class ID with

[7] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[8] http://propbank.github.io/
[9] https://verbs.colorado.edu/semlink/
[10] https://wordnet.princeton.edu/
[11] It should be noted that the coverage of these lexicons, at least based on the links found in CzEngClass, is not sufficient to allow for a systematic use in the annotation process, both automatic and manual.
[12] https://catalog.ldc.upenn.edu/LDC2012T08

each content verb (or at least to those contained already in CzEngClass). The "inversion" is meant with reference to the Fig. 1, where the arrows are driven top down. In the present project, the goal is to have direct reference links *from* the corpus at the bottom of Fig. 1 to the individual entries in the red box on the top. The CzEngClass resource in its current version 0.3[13] assigns IDs to each verb in every class; these IDs are being used as the final reference of the verbs in the PCEDT. The annotation schema (at the highest, TR level) has been extended by the items (attributes) listed in Table 3.

Attribute	Description
`syn_class`	root attribute container for the semantic reference
`syn_class/class`	class container (of current node)
`…/class/class.rf`	ID of the class
`…/class/rep`	human-readable class name(s)
`syn_class/semrel`	sem. role container (rel. to parent node)
`…/semrel/semrole`	semantic role
`…/semrel/form`	required form
`…/semrel/spec`	additional information
`…/semrel/fromclass.rf`	class to which the role belongs

Table 3: The attributes for semantic (synonym classes) extension of Tectogrammatical Representation

The `syn_class` structured attribute contains either the class reference (`class.rf`), or the appropriate semantic role (`semrole`), or both (for predicates that are at the same time arguments to other predicates, typically roots of embedded clauses). The `fromclass.rf` attribute of each semantic role is necessary for temporarily ambiguous assignment of classes to its effective parent predicate, or in case of multiple effective parents (in coordination structures etc.), for keeping the same distinction permanent in the annotation.

3.2 Assignment of Classes

The initial assignment of CzEngClass classes to the verb occurrences in the corpus has been done automatically. First, for each occurrence of a content verb in the corpus, its valency reference is retrieved and searched for in the CzEngClass lexicon (checking each member in each class). If found, the class to which this class member belongs is recorded with the content verb node in the corpus (the `syn_class/class/class.rf` attribute) and the other attribute (`rep`) is filled appropriately.

As already explained in part in the last paragraph of Sect. 2.2, it is possible that a verb (as identified by its valency frame ID) is found in more than one synonym class, and therefore that each occurrence of that verb is annotated by several classes. In theory, this should not happen if the valency lexicon entries distinguish all possible verb senses, as they in principle should (Hajič et al., 2003). However, as the "senses" are defined, within the valency theory used, at the linguistic meaning level (and not fully semantically), there is no contradiction in the fact that some valency entries appear in several (semantically defined) synonym classes. Thus, there is no reason to "blame" the valency lexicons in such a case, and semantic differences will have to - naturally - be resolved during (semantic) corpus annotation.

3.3 Assignment of Semantic Roles

After an appropriate CzEngClass class is assigned to a verb node in the semantic representation of the corpus, the arguments are mapped to the semantic roles associated with that class and they are also filled into the `syn_class/semrel` attributes of the `syn_class` structure of the argument's nodes. The automatic part of the assignment proceeds as follows:

- all the argument nodes of the given predicate are identified, based on the valency frame of the predicate verb as originally recorded in the corpus;

[13] `http://hdl.handle.net/11234/1-2977`

- the functor of every identified argument node is assigned the appropriate semantic role based on the CzEngClass lexicon mapping of semantic roles to arguments for the appropriate verb (class member) entry;
- the role, the form, additional information and ID of the class to which the semantic role belongs are stored in the `syn_class/semrel` attributes of the appropriate argument node.

This process might not, however, lead to all roles being represented in the corpus. For roles that are - for the given predicate (class member) - not mapped to any argument, it is necessary to introduce a new node in the semantic representation. This concerns roles mapped to pseudo-functors `#sb`, `#sth`, and `#any`. In those cases, this new node gets a special "lemma" #SitRef ("situational reference"), its (pseudo-)functor is copied from the CzEngClass mapping and its semantic role is filled into this node's `syn_class` attributes. Similarly, for optional arguments or free modifications (adjuncts) listed in the CzEngClass entry mappings (for the given verb) which have not been found in the TR of the sentence in the corpus, a new artificial node is inserted. This node gets also the #SitRef "lemma", the appropriate functor based on the CzEngClass mapping, and the corresponding semantic role; all filled into this new node's `syn_class` attributes. This is similar to the approach to implicit semantic roles ("Null Instantiations") described in (Ruppenhofer et al., 2009); the difference is that in our approach, licensing of such elements is based strictly on the set of roles assigned to the class (not on English - or any other language's - grammar, given that we aim at multilingual classes). Also, we do not distinguish types of such situational references (indefinite, definite, ...) and defer this to the future process of discourse-based linking of the #SitRef to their referents, again without taking (grammatical) licensing into account; existence of a link will then correspond to definite null instantiations.

The #SitRef nodes are not created when two (or more) mappings exist for a given semantic role, and not every functor from these mappings is found in the corpus sentence representation. In such a case, only those present in the corpus are assigned a semantic role from the CzEngClass entry, and no new nodes are created.

However, if, for a given semantic role, no node in the corpus exists to which it is mapped in the CzEngClass lexicon, only one #SitRef node is created, namely the one corresponding to the functor with the highest precedence, where precedence is heuristically defined in the following way:

- core arguments in the order ACT, PAT, ADDR, EFF and ORIG;
- free modifications in the order BEN, SUBS, RCMP, MANN, LOC, TWHEN, DIR3, CAUS, AIM;
- all other free modifications in alphabetical order.

In the case of repeated free modifications in the corpus, only the first one (leftmost) is assigned the appropriate mapped semantic role.

4 Properties of the Enriched Corpus

After the automatic assignment of the "inverted" references, statistics have been collected. For certain configurations, examples have been extracted and an initial manual inspection performed.

4.1 Basic Statistics

About 50% of verbs annotated with the original valency lexicon entry in the corpus have received a CzEngClass ID (67,733 out of 130,079 on the English side and 48,445 out of 118,029 on the Czech side).

However, only 33,005 English (32,560 Czech) verbs are aligned with a Czech (English) verb found in CzEngClass.[14] In conclusion, the coverage of the corpus by the current version of CzEngClass is about half of the corpus in terms of independent coverage of its Czech and English side, but only slightly above 25% when also the bilingual alignment is taken into account.

Up to five classes have been assigned to a single verb node in the corpus; i.e., there are verbs (verb senses) in both Czech and English that appear in (up to) five CzEngClass synonym classes. While the 1:1 alignment

[14] The asymmetry between the two last numbers is due to non-1:1 verb alignments.

prevails (in about one third of the cases where the aligned verbs are both found in CzEngClass), there are nontrivial numbers of occurrences of a 2:2, 1:2, 2:1, 3:2 etc. alignments.

Finally, of those 27,242 pairs aligned n:n (only 1:1, 2:2 and 3:3 alignments found), 21,050 CzEngClass class pairs fully matched between the two languages.

4.2 Manual Inspection and Examples of Mismatch

The fully matching pairs (i.e., those 21,050 matching pairs, or more precisely the 16,825 1:1-aligned full matches) are in fact what is to be expected, should the bilingual semantic synonym lexicon be "nice and clean." However, not unexpectedly, language(s) do(es) not behave that nicely. It is therefore interesting to investigate the other cases.

Manual inspection of the non-1:1, non-matching cases revealed the following:

- for any non-1:1 alignment, i.e., for cases when the Czech or English verb or both are in more than one class: either the classes should be merged (as is often the case, e.g., for the verbs of communication, as we have acknowledged in (Urešová et al., 2018b) while describing an independent manual annotation experiment), or the verb sense distinctions as represented by the valency frames in the PDT-Vallex and EngVallex lexicons, as used for the PCEDT corpus annotation, are too coarse-grained and should be split into more verb senses;

- for an 1:1 alignment where the classes do not match, there are two possible causes:

  – the classes/alignments are plain wrong,
  – or the alignment is (sort of) OK, but the original sentence has been translated too freely, reformulating the source text to the extent that synonymy between the two "corresponding" verbs does not hold as defined in the CzEngClass specifications (Urešová et al., 2018a).[15]

For example, the aligned verb pair *say* - *uvést* has been (in many sentences) assigned three classes on the English side and two of them at the Czech side; closer inspection shows that these are to be merged.[16]

Example of a non-matching 1:1 class alignment is the pair *suggest* - *ukázat*: each side has been assigned a different class. Closer inspection shows that the Czech verb has semantically two different meanings - one is close to *suggest*, and the other one corresponds to *prove, implicate, establish, demonstrate, ...* which suggests that two senses of the Czech verb *ukázat* should be established (pun intended).

We are leaving out the cases where the original alignment does not strictly pair verbs (e.g., the translation is not literal, nominalization has been used, alignment error).

5 Manual Corrections

After the automatic part of the semantic annotation process has been completed, manual effort is needed to disambiguate and correct it.

One hundred paired verb occurrences in the parallel corpus have been selected from section 00 of the PCEDT (continuously to have wider context available).[17] Only those aligned pairs that have been both automatically assigned at least one class have been considered. Out of these 100 pairs, i.e., 200 verb tokens (100 on the Czech side, 100 on the English side), 117 verb occurrences have been assigned multiple classes (up to 4 different ones), 48 in Czech and 69 in English. As a first step, these had to be manually disambiguated.

5.1 Removing Duplicate Classes

As it appears, many of the multiply assigned classes have in fact been the artifact of the CzEngClass lexicon construction - some classes are clearly duplicates and should have been merged. This concerns mostly the class *say* - *říci*, which is in fact assigned (to various verbs occurring in the corpus) in almost half the

[15] In fact, the translation could also be plain wrong, but we have not found such a case.

[16] It should be noted that the reason for having very similar classes that in fact should be just one class (i.e. to be merged in the process) is merely the way the classes have been created: each has been seeded by a randomly chosen verb, but some could have been synonyms, which could only be revealed later by the annotation process as described in (Urešová et al., 2018b) and here.

[17] Files `wsj0006` to `wsj0020`.

sentence pairs (46 out of the 100 pairs). The other class identified for such a merge in the lexicon is *require - vyžadovat*. After removing such duplicates, there remained 27 Czech and 60 English verb occurrences to manually disambiguate.

5.2 Manual Class Disambiguation

After the remaining 27 Czech and 60 English verbs have been disambiguated, statistics on the type of the disambiguation have been collected (Table 4).[18]

Disambiguation type:	More specific class selected	More general class selected	Competing selected
Czech	13	7	7
English	51	2	7
Total	64	9	14

Table 4: Statistics on manual class disambiguation results, by type

Closer inspection shows the following:

- most of the English cases where more specific meaning has been selected applies to the verb *say*, which appears both in the class *say - říci* as well as in the more general class *talk - mluvit*.[19]

- other examples where the more specific class has been selected are the classes *offer - nabídnout* and *provide - poskytnout*, which share, e.g., the verb *offer* itself, used in two different contexts: if the entity offered is a true offer which can be refused, it belongs to the more general class *offer - nabídnout*, whereas if the offer also means that it has to be 'accepted' unconditionally, then it has to be annotated by the class *provide - poskytnout*, as in: *"[it] is the second incentive plan the magazine has offered advertisers in three years"*[20], where the advertisers have no choice but to use this new plan (if they want to advertise in this particular magazine).

- examples of where the more general class of competing classes in a hierarchical relation is *pay - platit* vs. *repay - splatit*; these classes share a number of verbs, but depending on context, the more specific or the more general class must be selected. In the sentence *"... to pay for the plant"*[21] the more general interpretation has been selected, since not even the textual context (discourse) makes it clear whether this payment is a "simple" payment, or a repayment of a loan or similar debt.[22]

- an example of a selection among competing classes, where no hierarchy among them could be identified, are the following three classes, found three times in the annotated sample: *expect - čekat* (something from somebody), *anticipate - předpokládat* and *predict - předpovídat*, which share the verbs *expect, suppose, believe, očekávat, čekat* and others. While these verbs share valency frames across usages and interpretations, they are used in non-synonymous contexts - expecting (that someone does something which he/she should do or is planned to do) is not the same as predicting/forecasting (that something happens) and that is in turn not the same as anticipating (that something I believe happens actually happens). However, in this case no hierarchy could be determined among these classes.

[18] The purpose of this experiment was to find interesting cases, as opposed to measuring statistical variables such as inter-annotator agreement, for which there was not enough data. For larger but simpler annotation sample and its evaluation, see (Urešová et al., 2018b).

[19] There is no explicit hierarchy recorded in the CzEngClass lexicon yet, but the analysis of (not only) the ambiguous cases clearly shows that there must be one to be taken into account in the future.

[20] File wsj0012, sentence 2

[21] File wsj0015, sentence 3

[22] The actual full sentence reads: *"In a disputed 1985 ruling, The Commerce Commission said Commonwealth Edison could raise its electricity rates by $49 million to pay for the plant."*; we might speculate that from the fact that the company is acquiring the plant *now* while rates can only be raised *in the future*, it follows that it has to borrow money now and *repay* those $49 million later, but that would be reaching too much into the world knowledge and even then, one cannot exclude that the conditions of the contract will be different and no borrowing in fact occurs. In such cases, the rule that we have adopted reads "use the more general class if no clear evidence is found to opt for the more specific one".

The above points confirm, as already noted, that it will be practical to introduce a hierarchy into the (so far) flat list of classes in the CzEngClass lexicon. However, it seems that such a hierarchy will be slightly different than the one found e.g., in FrameNet, mainly because it will have no "container-only" classes (such as those found in FrameNet). The nodes in the "hierarchical tree" will be the classes themselves, and the tree will be used for guiding the annotation. We are leaving for future work to see if the evidence from the corpus annotation, as exemplified above, has any theoretical or methodological implications in the area of synonymy or lexical semantics in general.

5.3 Reassigning Semantic Roles

In some cases, the automatic assignment of semantic roles based on the CzEngClass lexicon failed; out of the 100 predicated verb occurrences examined (corresponding to 290 pairs of dependent aligned nodes holding semantic roles), 21 displayed a problem (7.2%). After an inspection, we determined a few types of failures:

- structural splitting of a semantic role:

  This is a typical case for verbs of communication, where one semantic role can be expressed either as one valency argument (typically as PATient) or as two valency arguments (typically split into PATient and EFFect): *Paul said that he is*.PAT-Information *a liar.* vs. *Paul said about him*.PAT-Information *that he is*.EFF-Information *a liar*.

- multiple structural expression of a semantic role:

  Some semantic roles can be syntactically expressed in multiple ways that are not mirrored in the valency frame. This is partially due to the valency theory used for the Tectogrammatical Representation. For example, the semantic role "Speaker" can be structurally expressed in multiple ways, such as ACTor or LOCcation: *He*.ACT-Speaker *called him a liar.* vs. *In The New York Times*.LOC-Speaker, *he was called a liar*.[23]

- semantic roles reassigned to other nodes (not directly dependent on verb); for example, in a sentence containing *"... expect regulatory approval"*,[24] the semantic role Source for the verb *expect* is the *regulatory body*, but since it syntactically depends on *approval*, and not on the verb *expect* itself, the automatic assignment of roles based on the valency to role mappings could not identify it correctly.

5.4 Situational Reference

The newly introduced nodes identified by the "lemma" #SitRef are meant to be linked to the actual situational participant in the current sentence, or more likely somewhere else in the annotated document, much like textual co-reference is being marked in the tectogrammatical annotation of the corpus.[25] This situation appears not to be frequent, except for errors in the original annotation or in the underlying valency lexicons and for certain frequent verbs in classes with the Benefactive and Addressee roles. We are leaving this to the future work.[26]

5.5 Example Annotation

Figs. 2 to 5 show two (occurrences of) English verbs in the corpus, *say* and *expect*, both before and after the manual annotation as described in Sect. 3 and 5, for the sentence *"The thrift holding company* **said** *it* **expects** *to obtain regulatory approval and complete the transaction by year-end."* (File wsj0006, sentence 2).

In these Figures, the underlying section of the deep dependency tree as originally annotated using the tectogrammatical representation specification is shown in gray and node labels (lemma, functor) in black.

[23] Although one can consider to use two different labels for one semantic role distinguishing the animacy here (Speaker vs. Medium), due to a similar phenomenon occurring at other verbs and synonymous classes the authors decided to keep only one label and in sentences like *NYT* .ACT *called him a liar*, the ACTor is still labeled as Speaker and not Medium.

[24] File wsj0015, sentence 23

[25] Unless they were reassigned in one of the previous steps of the manual revisions of the corpus.

[26] For the purpose of the experiment described here, the #SitRef nodes have not been part of the evaluation.

Blue and brown arrows show textual and grammatical coreference links. The CO label suffix denotes parts of coordinated structure. The most relevant for the discussion here are the red and green node attributes: the classes assigned by the CzEngClass lexicon are in red, and the semantic roles are in green (please note that the semantic roles are complemented by the class identifier, also in green, to determine to which class this role belongs in case of multiple classes assigned to their parent verb node).

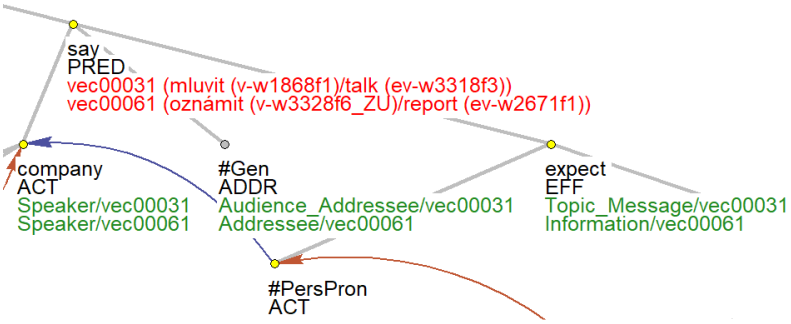

Figure 2: Automatic assignment of classes to *say* - ambiguity between classes 31 and 61

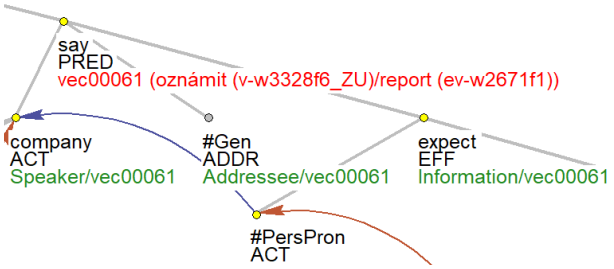

Figure 3: The verb *say* after disambiguation - class 61 and its roles selected

Figures 2 and 3 show a simple case where it was sufficient to disambiguate the appropriate class assigned automatically to the verb *say* in the example sentence, based on CzEngClass lexicon. Class 31 (*talk*) does not fit since the embedded clause is not just a topic, but a full information conveyed. Once disambiguated, the roles fit correctly.

Figures 4 and 5 show the more complex case of *expect*. First, there are three synonym classes to disambiguate: 2 (*expect*), 92 (*await*) and 93 (*assume*); the correct one in this case is class 2 (*expect*). For roles, it is necessary to link the newly generated #SitRef node to its (cognitive) antecedent, which is the *regulator*.

6 Conclusions and Future Work

We have described an experiment that enriched an existing annotated corpus by verb synonym classes, using a preliminary version of the CzEngClass lexicon. Even after the full lexicon is available, it is however expected that even the approx. half of the corpus (if the percentages from Sect. 4.1 can be extrapolated) which could have been automatically pre-annotated with the CzEngClass entries will need some manual inspection.

The verbs and their translations will, naturally, be never perfectly 1:1 aligned; we have shown some of the reasons and examples in Sect. 4.2. Assuming the identified duplicate classes are removed from the lexicon, there is about one-quarter of verb occurrences on the Czech side that are ambiguously annotated and need manual disambiguation; on the English side, this number was higher (60 out of the 100 sample occurrences), but this was due to a single verb - *say* - which might be perhaps tackled by introducing heuristics into the automatic pre-annotation procedure (e.g., select the class assigned on the Czech side if the aligned Czech verb is unambiguous and its class matches one of the English classes, possibly with additional restrictions).

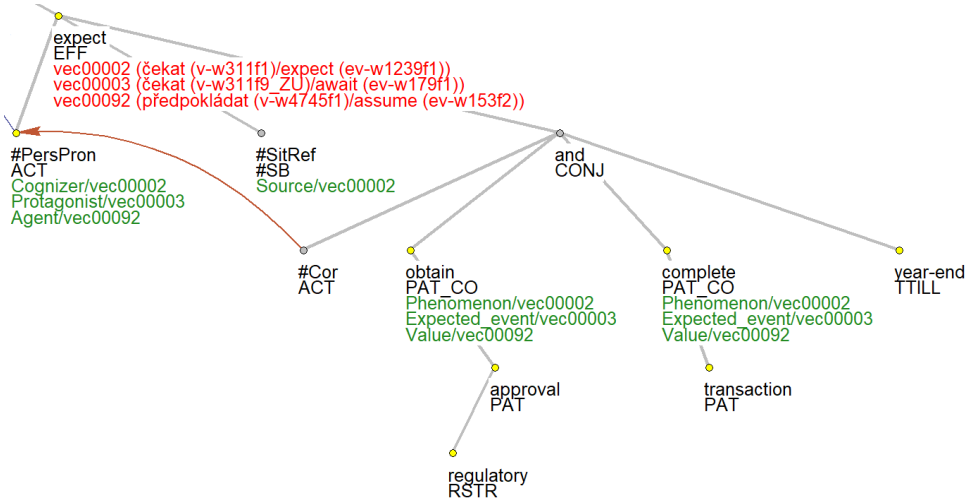

Figure 4: Automatic assignment of classes to *expect* - classes 2, 92 and 93

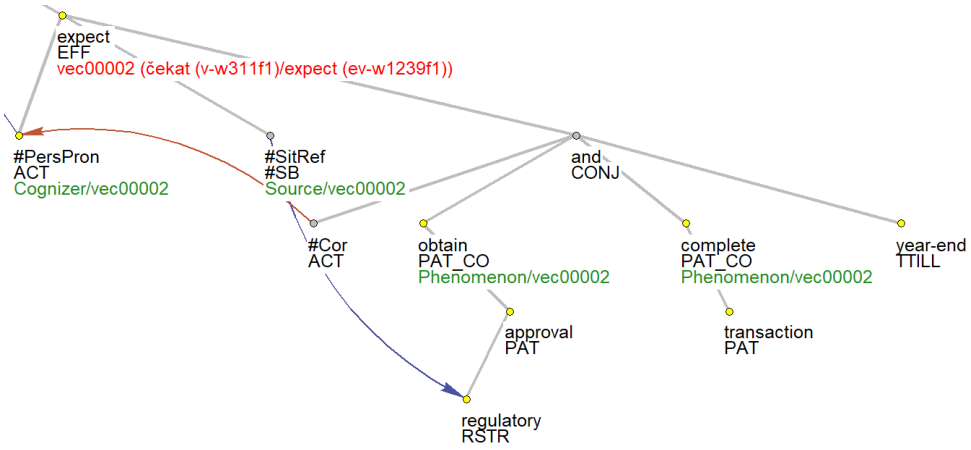

Figure 5: The verb *expect* after disambiguation - class 2 and its roles selected, #SitRef linked to *regulatory*

Similarly, the assignment of semantic roles to verb arguments and adjuncts as annotated in the corpus will need a manual pass. However, as the sample annotation has shown, the expected effort to correct errors in the semantic roles labeling part is relatively small - in the sample's 290 pre-assigned semantic roles, only 7.2% had to be corrected.

In the future, once the CzEngClass lexicon is published in full covering most, if not all, of the PCEDT corpus, the process described in Sect. 3 will be rerun, all the class alignments checked, and the resulting corpus will be published. This interlinked pair of resources will then be used for comparative lexical-semantic studies (also thanks to the links to the external lexicons, such as FrameNet, VerbNet, PropBank and WordNet), for study of translation from the lexical equivalence and synonymy perspective, and for machine learning experiments, e.g., for automatically extending the verb class synonym lexicon, and eventually for fully automatic annotation of (mono-, bi- and multilingual) corpora.

Acknowledgements

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic. It uses resources hosted by the LINDAT/CLARIN (LINDAT/CLARIAH-CZ) Research Infrastructure, projects No. LM2015071 and LM2018101, supported by the Ministry of Education of the Czech Republic. Some of the resources have been also funded or co-funded by the European Commission through several projects of the 6^{th} and the 7^{th} Framework Programmes.

References

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.

Silvie Cinková, Eva Fučíková, Jana Šindlerová, and Jan Hajič. 2014. *EngVallex - English Valency Lexicon*. LINDAT/CLARIN digital library. `http://hdl.handle.net/11858/00-097C-0000-0023-4337-2`.

Silvie Cinková. 2006. From PropBank to EngValLex: adapting the PropBank-Lexicon to the valency theory of the functional generative description. In *Proceedings of LREC 2006, Genova, Italy*.

Cecily Jill Duffield, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E. Vieweg, Jenny Davis, and Martha Palmer. 2007. Criteria for the manual grouping of verb senses. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 49–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA. 423 pp.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Charles J. Fillmore, 1977. *Scenes-and-frames semantics*, chapter 3, page 55 – 79. Number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

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 Erhard Nivre, Joakim//Hinrichs, editor, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57—68, Vaxjo, Sweden. Vaxjo University Press.

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, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2018. *Prague Dependency Treebank 3.5*. Charles University, Prague, Czech Republic. LINDAT/CLARIN digital library. `http://hdl.handle.net/11234/1-2621`.

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*. LDC, Philadelphia, PA, USA. `https://catalog.ldc.upenn.edu/LDC2006T01`, Catalog No. LDC2006T01.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending verbnet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1027–1032, Genoa, Italy, May. European Language Resources Association (ELRA).

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

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

Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9 – 15.

Jarmila Panevová. 1974. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. *International Journal of Semantic Computing*, 01(04):405–419.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended theory and practice. *Unpublished Manuscript*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW '09, pages 106–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.

Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevová, and Marie Mikulová. 2014. *PDT-Vallex*. LINDAT/CLARIN digital library. `http://hdl.handle.net/11858/00-097C-0000-0023-4338-F`.

Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

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'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. Defining verbal synonyms: between syntax and semantics. In Dag Haug, Stephan Oepen, Lilja Ovrelid, Marie Candito, and Jan Hajič, editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018) (Pub. No. 155)*, pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018c. Synonymy in Bilingual Context: The CzEngClass Lexicon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2456–2469.