Proceedings of the Globalex Workshop on Linked Lexicography, pages 10–19
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

SynSemClass Linked Lexicon: Mapping Synonymy between Languages

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, Jan Hajič
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25
11800 Prague 1, Czech Republic
{uresova,fucikova,hajicova,hajic}@ufal.mff.cuni.cz

Abstract

This paper reports on an extended version of a synonym verb class lexicon, newly called SynSemClass (formerly CzEngClass). This lexicon stores cross-lingual semantically similar verb senses in synonym classes extracted from a richly annotated parallel corpus, the Prague Czech-English Dependency Treebank. When building the lexicon, we make use of predicate-argument relations (valency) and link them to semantic roles; in addition, each entry is linked to several external lexicons of more or less "semantic" nature, namely FrameNet, WordNet, VerbNet, OntoNotes and PropBank, and Czech VALLEX. The aim is to provide a linguistic resource that can be used to compare semantic roles and their syntactic properties and features across languages within and across synonym groups (classes, or 'synsets'), as well as gold standard data for automatic NLP experiments with such synonyms, such as synonym discovery, feature mapping, etc. However, perhaps the most important goal is to eventually build an event type ontology that can be referenced and used as a human-readable and human-understandable "database" for all types of events, processes and states. While the current paper describes primarily the content of the lexicon, we are also presenting a preliminary design of a format compatible with Linked Data, on which we are hoping to get feedback during discussions at the workshop. Once the resource (in whichever form) is applied to corpus annotation, deep analysis will be possible using such combined resources as training data.

Keywords: Linked Lexicon, Linked Data, Semantics, Syntax, Synonymy, Parallel Corpus

1. Introduction

The aim of the presented research is to create a linked lexicon of bilingual Czech-English synonyms, now openly available in version 1.0 (Urešová et al., 2019). Synonyms are extracted from translated texts of the Prague Czech-English Dependency Treebank corpus. A functionally adequate relationship in terms of translation must exist between the meaning of the English and the Czech verbs, i.e., the English and the Czech verb(s) are considered synonymous in the given context(s) if the translated verb adequately expresses the functional intent of the original. We aim for each synonym class to be characterized both meaning-wise (verb sense(s), semantic roles) and structurally (valency arguments) by linking (mapping) semantic roles and valency members (Role $\leftrightarrow$ Argument mapping).

This paper synthesizes previous work on the lexicon to comprehensively describe its version 1.0 published in connection with this paper (and under a new name that reflects future direction from bilingual to multilingual entries), but it also adds - on top of a comprehensive description of lexicon structure and the process of its creation - a number of interannotator agreement evaluation experiments (Sect. 4) and a first attempt at defining a Linked Data scheme for it (Sect. 5).

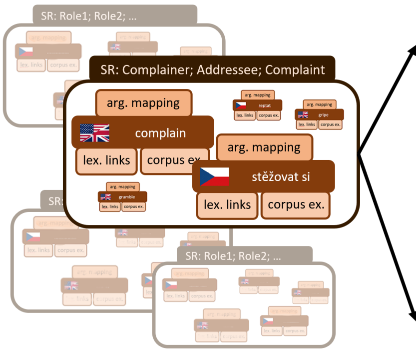
The paper is structured as follows. In Sect. 2, structure and content of the lexicon are described. The resources used and linked to are presented in Sect. 3. Sect. 4 contains a description of the process by which the lexicon has been created, i.e., the annotation process and interannotator agreement (IAA) analysis. The principles of (re)structuring and (re)formatting the links to Linked Data format are described in Sect. 5, and related work is described in Sect. 6. We summarize our work and outline future plans in Sect. 7.

2. Structure and Content of the Lexicon

The SynSemClass lexicon - formerly CzEngClass, whose previous preliminary versions as well as various aspects of its theoretical basis and the annotation process are described in (Urešová et al., 2019a; Urešová et al., 2018a; Urešová et al., 2018e; Urešová et al., 2018d; Urešová et al., 2018c; Urešová et al., 2018b) - builds upon the PCEDT parallel corpus (Sect. 3.1) and the existing internal resources, namely CzEngVallex, PDT-Vallex, and EngVallex lexicons (Sect. 3.2). On top of that, other lexical databases, namely FrameNet, VerbNet, PropBank, OntoNotes and WordNet (Sect. 3.2) are used as additional sources, and links are annotated and kept between their entries and the SynSemClass entries. The overall scheme of the lexicon with an example of one class entry is depicted in Fig. 1.

Each synonym class contains Czech and English verbs (verb senses) that have similar meaning. The latest version of SynSemClass captures 3515 verb senses with 2027 on the English and 1488 on the Czech side. The synonymous senses are represented as valency frames (of generally different verbs) and they are called Class Members.

Each class is assigned a common set of semantic roles, called a Roleset. A Roleset contains the core "situational participants" common for all the Class Members in one class. When determining Class Membership for a potential candidate verb (sense), the Roleset also serves as a source of context information: if all the semantic roles from the Roleset can be mapped to valency slots (labeled by a "functor" in the valency theory (Panevová, 1974) within the Functional Generative Description, or FGD (Sgall et al., 1986), framework) for the given verb sense as recorded in its valency frame in the appropriate valency lexicon (and vice versa), it is deemed–together with the approximate sense match to the other Class Members–as belonging to



Class ID: vec00132

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

Figure 1: The overall scheme of the SynSemClass lexicon and an example of a class (“complain-stěžovat si”)

that synonym class. As a result, each Class Member has its valency frame slots mapped to the semantic roles kept in that class’ Roleset.[1] The valency frames of different Class Members of one SynSemClass class are thus “compatible” through the mapping to the Roleset, even if they possibly differ in their argument labels (valency slot functors).

Some typical Rolesets:

- Class “klesnout-decline”: Item, Initial_value, Final_value, Difference
- Class “nabídnout-offer”: Offerer, Recipient, Entity_offered, Entity_received
- Class “chránit-defend”: Defender, Asset_Patient, Harmful_situation

While the role labeling system is still preliminary, we strive to have the names semantically descriptive in nature (i.e., Offerer instead of Agent), and we keep the convention that for better human understanding, we use an underscore for signalling an “or” when the following word is capitalized (as in Asset_Patient), as opposed to a mere use of multiple words to describe the role, in which case the next word is not capitalized (Final_value).

When mapping the roles from a given Roleset, each of the roles must be mapped to “something” from the valency frame of a verb in that class; that “something” may be

- either a member of the valency frame,
- or any other free modification to which the given semantic role might be mapped,
- or a proxy semantic participant (#any, #sb, #sth).

Conversely, each member of the valency frame of a verb listed in the given synonymous class must be mapped to a semantic role from the assigned Roleset.[2] If any member of the valency frame of a potential Class Member of the given synonymous class really cannot be mapped to the chosen Roleset of that class, then that candidate Class Member cannot be included in the class.

Class Members are further linked to the original resources used (the parallel Czech-English treebank and the Czech and English valency lexicons) and also to other external resources (see Sect. 3).

A simplified example of the synonym class “complain–stěžovat si” is shown in Fig. 1. It schematically shows the SynSemClass lexicon on the left with its entries (= classes), and an example synonym class in the form of a table with the additional annotation available for this entry. Most importantly, the table shows the Argument mapping between the roles in the common Roleset for this class (which in this case contains three semantic roles: Complainer, Addressee (of the complaint) and Complaint) and the individual Class Members’ arguments from the PDT-Vallex and EngVallex lexicons. While in most cases the mapping is straightforward (as is the case of the valency slot ACT, which is mapped to Complainer for all the Class Members shown in Fig. 1), in some cases there is a need to specify certain restrictions (e.g., restrict the mapping between ADDR and Addressee to the use of the preposition “to”) or a combination of arguments (slot names) mapped to a single semantic role (PAT|EFF is mapped to Complaint for the Czech Class Member “stěžovat si”). For more details about this mapping and its annotation, see Sect. 4.3.

The links to external resources follow - for English, they contain the OntoNotes sense number (e.g., 1, or a hyphen if no OntoNotes sense to map to is available for the given verb sense), FrameNet frame name (Complaining), PropBank roleset number (e.g., gripe.01) and WordNet sense number (e.g., #1). Alternatives may exist (e.g., for WordNet senses - see grumble, which maps to both grumble#1 as well as grumble#2 in WordNet). Czech verbs are linked only to the VALLEX lexicon (Lopatková et al., 2016), and in the future, they will also map to Czech WordNet.

Examples are selected from the available corpora, in this case from the Prague Czech English Dependency Treebank (PCEDT), which is a parallel version of the WSJ part of the Penn Treebank (WSJ section and sentence ID number is used in Fig. 1, see also Sect. 3.1). These examples are selected so that they best characterize the corresponding verb sense as included in the particular class.

The extended version of SynSemClass (Urešová et al., 2019) is openly available in the LINDAT/CLARIN repository[3] contains 145 synonym classes with 3515 verbs fully

[1] This ‘perfect’ 1:1 mapping has to be relaxed in specific cases, see e.g., (Urešová et al., 2018a).

[2] There is only one exception to this rule: If the valency frame of an English Class Member includes a non-obligatory free modification (which is not in line with the FGD rules), it may not (but can) be taken into account in the mapping and when the Roleset is created.

[3] `http://hdl.handle.net/11234/1-3125`

annotated in Step 2, out of which 57 classes are also annotated in Step 3. For more details on the annotation process and its Steps see Sect. 4.

3. Resources Used and Linked to

In this section, we describe the main corpus used as the source of evidence for creating the SynSemClass lexicon entries (= the synonym classes), and the lexical resources used for both identifying the Class Members as well as linking them to the external lexicons.

3.1. The Corpus

As described in previous papers on this resource (Urešová et al., 2019a; Urešová et al., 2018a; Urešová et al., 2018e; Urešová et al., 2018c; Urešová et al., 2018b), for evidence examples, we use the parallel Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0) (Hajič et al., 2012). This corpus contains approx. 50 thousand aligned sentence pairs. The English side is the WSJ part of the Penn treebank (Marcus et al., 1993); it has been translated to Czech by professional translators. Each language part is enhanced with a rich manual linguistic annotation in the Prague Dependency Treebank (PDT 2.0) style (Hajič et al., 2006; Hajič et al., 2018) which is based on the Functional Generative Dependency (FGD) framework (Sgall et al., 1986). For the purpose of our work, it is important that the annotation captures aligned surface dependency trees and deep syntactico-semantic (tectogrammatical) trees across the two languages on sentence and node levels. Moreover, at the deep (tectogrammatical) layer, each verb node (occurrence) is assigned a valency frame, also representing a verb sense, by way of using its ID which identifies it in the associated valency lexicons, PDT-Vallex and EngVallex (Sect. 3.2).

3.2. Linked Lexical Resources

When building the synonym classes, we proceed from the PDT-style valency lexicons which are an integral part of the PCEDT. The existing annotation of PCEDT by the valency lexicon entries has helped to seed the SynSemClass lexicon and also to get real-world examples. For Czech verbs, PCEDT uses the Czech valency lexicon called PDT-Vallex (Urešová et al., 2014), (Urešová, 2011), while for English verbs, the English valency lexicon EngVallex (Cinková et al., 2014) is used. The most important links come from the CzEngVallex lexicon (Urešová et al., 2015), (Urešová et al., 2016), a bilingual valency lexicon which combines PDT-Vallex and EngVallex entries and contains not only Czech and English verbs which are translation equivalents to each other but it also captures mapping among their valency arguments.
The individual Class Members in SynSemClass are further mapped to the following external lexical resources: FrameNet (Baker et al., 1998; Fillmore et al., 2003), VerbNet (Schuler, 2006), PropBank (Palmer et al., 2005), senses from OntoNotes Groups (Pradhan and Xue, 2009), English WordNet[4] and Czech Vallex (Lopatková et al., 2016).[5]

[4] https://wordnet.princeton.edu
[5] https://ufal.mff.cuni.cz/vallex/3.5

4. Creating the Lexicon

SynSemClass is being built strictly "bottom-up", i.e., from the corpus and existing lexical resources towards the new synonym lexicon. Since the lexicon is a complex resource, we divide its creation and annotation of its entries into three "areas": (1) determining which verbs should go into one class (Class Members), (2) determining the common set of semantic roles for each class and mapping it to valency for each Class Member, and (3) adding links to other existing lexical resources.
These three areas are intertwined and influence each other - for example, while linking a Class Member to the other lexical resources the annotator might realize that the Class Member should go to a different class, or that the class should be split into two, or merged etc., but overall, this "division of work" allows us to describe the structure of the lexicon and the annotation process more clearly.
The tasks to be performed to get full annotation and meet all the objectives in all of the three areas are even more complex. Going "bottom-up", i.e., starting from the PCEDT corpus, we proceed in four steps, interspersing automatic and manual phases.
In the automatic phases, the PCEDT corpus is used to get preliminary Class Membership and valency information for both Czech and English verbs.
In the manual phases, many (sub)tasks are performed for each class, all of them for verbs in both languages (Czech and English):

- pruning the preliminary Class Members in each class, eliminating clear misalignments and/or sense mismatches,
- creation (Step 1) and possible amendment (Steps 2 and 3, see below) of the set of semantic roles for each class (the Roleset),
- linking (mapping) semantic roles to valency members for each verb in the class, with possible restrictions on the semantics of the arguments,
- selecting the most appropriate examples from the corpus to accompany each Class Member,
- adding links to the external lexical resources.

4.1. The Annotation Process

The annotation process has been sequenced into an initial automatic seed selection step (Step 0) and three followup steps (Steps 1-3), each consisting of an automatic phase (pre-assignment of verbs from the aligned parallel corpus to the classes, as populated in the previous step), and a manual pruning and annotation phase (Fig. 2).
We will refer to these Steps later when describing the results, including inter-annotator agreement in the three annotation "areas".
These steps can be briefly described as follows (Fig. 2):

- Step 0: An automatic semi-random selection of 200 Czech verbs (frames or verb senses from the Czech valency lexicon) which provisionally denote class names and form the initial set of classes, and which represent verbs (valency frames) of various frequencies in the parallel PCEDT corpus.

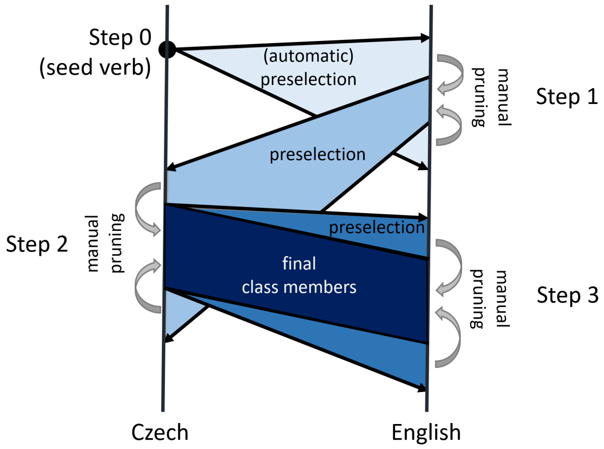


Figure 2: SynSemClass annotation process - major steps

- Step 1: For each of the provisional classes from Step 0 (containing just one Czech verb at this point), English verb translation counterparts have been automatically added based on the PCEDT corpus' (automatic) word alignments. This pre-selection has been followed by the manual phase, when Class Members have been pruned (using a five-point Likert scale, transformed then to binary membership decisions (Sect. 4.2.1)), a common Roleset has been created for the class, argument mapping and external links to English resources have been built as well as mapping restrictions and notes, and the appropriate English examples from the PCEDT have been chosen.
- Step 2: Czech translation counterparts of English verbs retained and annotated in Step 1 have been added in the automatic phase of this Step. Similarly to Step 1, these verbs have then been manually pruned, mappings to their arguments from the Roleset has been established (possibly amending the Roleset when necessary), and this argument mapping, plus restrictions, notes, external links and examples have been added for the newly selected Czech verbs.
- Step 3: English translation counterparts of the Czech verbs retained as Class Members for all the classes remaining after Steps 1 and 2 (after merging certain duplicate classes in the process, as described in (Urešová et al., 2019b)), except for the initial Czech verb from Step 0, have been added. Again, these preselected Class Members have been manually pruned, mappings to their arguments from the Roleset have been established (possibly amending the Roleset again when necessary), and argument mapping restrictions, notes, external links and examples have been added for the newly selected English verbs. This third step has been added after inspecting the results of Step 2 - it was apparent that the composition of each class was skewed towards containing more Czech than English verbs, which has been corrected by adding the additional English verbs in this Step. While it might seem that there is now again more English verbs than the Czech ones, manual inspection shows that this is due to the richer verb vocabulary used in the original text, while the Czech translation has been more uniform. No "Step 4" is thus planned.

In each Step, adjustments in the results of the previous Step(s) have been allowed. Specifically, after Step 2, it was clear that some of the original seeds (Czech verbs from Step 0) have been expanded to very similar or almost identical classes; therefore these classes have then been merged, reducing the overall number of classes. Such global operations, as well as the resolution of any disagreement between the annotators, have been done by a single adjudicator, who might also have modified or better exemplified the issues in the annotation guidelines.

Also, such modifications might further concern the Roleset, mappings of arguments to semantic roles, or the links to the other lexical resources.

The decision which roles to select for a given synonym class cannot be inferred easily from any single Class Member. Often, it is only during the annotation of all additional potential Class Members (as added in Steps 2 and 3) that the semantic character of each Class (and its Class Members) becomes clear, and some amendment has to be done in order to keep consistency at the level of "semantic depth" of the semantic roles (SRs) used. While the SRs are mostly taken from FrameNet, sometimes their labels have to be modified to fit the properties of the synonym class. Currently, there are 123 SRs taken from FrameNet and 49 SRs have a specific SynSemClass label; we are also using some of the ideas from previous work on comparison of semantic roles, such as (Aguilar et al., 2014).

In addition, the Roleset composition and annotation is very closely related to the mapping of valency slots of the valency frame for each Class Member to the individual semantic roles (Sect. 2, Sect. 4.3). It could even change the decisions made during the Class Membership inclusion/exclusion (Sect. 4), since if no mapping between the Roleset and the valency frame can be established, the verb must be excluded from the class.[6]

4.2. Determining Class Membership

4.2.1. Interannotator Agreement

In (Urešová et al., 2018d), the first interannotator agreement experiment was described where 8 annotators were determining Class Membership of the automatically preselected list of English verbal translation equivalents (total of 544 English verbs, as inserted by the automatic phase of Step 1) for 60 of the Czech seed verb senses (as selected in Step 0). In that experiment, the annotators had to prune the English Class Member candidates by checking the corpus examples through the CzEngVallex valency argument alignments, i.e., their usage in context. They could select not only Yes or No, but also a tentative Yes or No (Rather Yes, Rather No) and also a special value "Delete", which was used to signify a total mismatch (wrong underlying corpus alignment, in most cases); they were thus using sort of a 5-point Likert scale, mapped back to binary decisions (with

[6] In the actual annotation process, the annotators only report such a case and the final determination of deletion is being done by the adjudicator.

Yes and Rather Yes taken as a positive answer) for the IAA computation.

The agreement for the Class Membership, as described in (Urešová et al., 2018d), was as follows: Of the 544 datapoints the annotators (4 annotators for each decision, in our case) fully agreed in 358 cases (65.8%), which gives a good idea of the adjudication effort needed.[7]

For a pairwise comparison, we have used Cohen's kappa (Eq. 2), see also (Urešová et al., 2018d); macroaveraged over all annotator pairs, $\kappa = 0.44$.

The agreement as measured over all the annotators using Fleiss' kappa was $\kappa = 0.45$. While both the averaged Cohen's kappa as well as the Fleiss' kappa values are low, (Urešová et al., 2018d) also measured deviation from an average value on the full Likert scale used, and that value was surprisingly low (0.36 when the Yes-Rather Yes-...-Delete 5-point scale has been converted to values 4 to 0).

In Steps 2 and 3 (see intro to Sect. 4), we have concentrated on pairwise comparison, limited to three annotators (two of them continuing work from Step 1 and one new annotator). For those classes annotated by two annotators, we have computed mutual F_1 score (Eq. 1) (Jardine and van Rijsbergen, 1971) and Cohen's kappa κ (Eq. 2) (Cohen, 1960):[8]

$$F_1 = \frac{2PR}{P+R}, \tag{1}$$

where P is precision and R recall, and

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

where p_o is the observed and p_e the expected probability, as estimated from the annotated data of the pair of annotators. We provide the numbers microaveraged over the n classes each annotator pair worked with, and then macroaveraged over the three pairs of annotators.

In Step 2, three annotators have been pruning the automatically preselected Czech verbs, using the same scale as in Step 1 (except now for Czech verbs). At least seven classes (498 verbs min.) have been double annotated by each pair of annotators, to measure the IAA. The pairwise results are summarized in Tab. 1; the macroaveraged F_1 score is 0.95, and the macroaveraged kappa value is 0.94.

Annotator	A_4	A_5	A_9
A_4	-	0.93 / 0.91	1.0 / 1.0
A_5		-	0.91 / 0.90
A_9			-

Table 1: F_1-measure / Cohen's kappa κ for Class Membership annotation in Step 2 (Czech verbs)

[7]The total agreement has been measured using all labels; i.e., if 3 annotators assigned No and 1 assigned Rather No, it did not count as full agreement.

[8]While Cohen's kappa is routinely used in IAA computation, we were curious how it differs from the F_1 measure, which is used for many tasks in NLP. Please note that there is (naturally) no true gold standard when computing IAA; the F_1 is symmetrical between the two annotators, with Precision and Recall swapped when computed in the opposite direction.

In Step 3, similarly to Step 2, several classes have been selected to double-annotate them to assess IAA, in this case on English verbs (Tab. 2). At least seven classes (169 verbs min.) have been annotated by a pair of annotators. One of the annotators for the multiple annotation experiment was a native speaker.[9] The macroaveraged F_1 score over the three pairs of annotators is 0.82, and the macroaveraged kappa value is 0.52. This is better than in Step 1 (which was also concerned with English verbs).

Annotator	A_4	A_6	A_9
A_4	-	0.79 / 0.52	0.83 / 0.49
A_6		-	0.84 / 0.56
A_9			-

Table 2: F_1-measure / Cohen's kappa κ for Class Membership annotation in Step 3 (English verbs)

It can be seen from these numbers that apparently determining Class Membership for English verbs is harder (results of Steps 1 and 3) than for the Czech verbs (Step 2). A natural explanation would be that Czech native speakers would be better aligned for determining Class Membership for the Czech verbs, but the numbers from Step 3, where one of the annotators was a native speaker of English, suggest that this might not necessarily be the case.

4.3. Mapping Semantic Roles to Valency

As described in Sect. 2, an important part of each synonym class is the Roleset, set of semantic roles that are shared among Class Members. To make sure that these SRs are applicable to all of them, there must exist a mapping between the valency frame slots of each Class Member and the SRs of that class' Roleset. This mapping was also done fully manually, for all the 3515 Class Members in the current version of SynSemClass. For some mapping examples, see Fig. 1 - e.g., the verb "grumble" (more precisely, the valency frame ev-w1502f1 for grumble) in the class "complain–stěžovat si" maps ACT to Complainer, ADDR to Addressee and PAT to Complaint.

4.3.1. Interannotator Agreement

For IAA in the Roleset-to-valency-slot mappings, we have computed a full match between an annotator pair (microaveraged over all verbs in the classes that were annotated by the two annotators), and then also matches for the individual valency slot labels, or functors (ACT, PAT, ADDR, EFF, ORIG, and "other" (all remaining) used in the valency frames). Only fully equivalent mapping of all (valency slot functor $\leftrightarrow$ semantic role) pairs counts as a correct complete match. Only agreement accuracy is computed, as the ratio of a number of complete matches between the two annotators to the number of Class Members considered as valid Class Members by both annotators (valid means that they annotated the Class Member as Yes or Rather Yes when

[9]This annotator was new to the project and had to learn the objectives, principles and concrete rules of annotation from the project documentation, but there has been enough learning period to consider the experience on par with the other annotators.

determining its Class Membership).[10] The results are in Tab. 3 and Tab. 4, for Step 2 and Step 3 mappings, respectively.

Annot. pair	A_4-A_5	A_4-A_9	A_5-A_9	Avg.
# of pairs	70	124	84	-
Accuracy	31.4%	78.2%	45.2%	51.6%

Table 3: Accuracy of a full manual match between slot to role mappings, for 3 annotator pairs in Step 2 (Czech verbs)

Annot. pair	A_4-A_6	A_4-A_9	A_6-A_9	Avg.
# of pairs	87	311	91	-
Accuracy	67.8%	87.5%	83.5%	79.6%

Table 4: Accuracy of a full manual match between slot to role mappings, for 3 annotator pairs in Step 3 (English verbs)

The low agreement numbers in Tab. 3 are mainly due to the fact that during Step 2 (when the first set of Czech verbs has been added) not only the mappings, but also the Roleset as created in Step 1 for the initial, mostly English verb set has been often modified, causing a mismatch between the two annotators.[11] Examples include adding or deleting a role, or a partial swap etc. In Step 3 (i.e., after adding another set of English verbs to each class), the Roleset has hardly been ever changed, and only the mapping to valency slots was the cause of mismatches. Thus in this case, we believe that the difference in average agreement (51.6% vs. 79.6%) is not a language issue.

Tables 5 and 6 show the breakdown of the accuracy of the valency slot to role mappings. Only the core argument slots are listed individually, while all other (incl. the special #any, #sb and #sth slots) are grouped together. As expected, the agreement accuracy measure is higher than for the complete match for the whole valency frame, and quite high in general, except for the mix of other non-core valency slots for the Czech verbs added in Step 2 and for the EFF valency slot for English verbs added in Step 3. Since the EFF valency slot corresponds to the third, fourth or fifth argument, as the case may be, we can only speculate that perhaps the EngVallex valency slot labeling might not be consistent enough to allow the annotators understand well its relation to the semantic roles in the given class, and they then therefore differ in their judgment.

4.4. Mapping to Other Lexical Resources

In this section, we evaluate the interannotator agreement in linking the individual class members to external resources, as described in Sect. 3.2, i.e., to VALLEX on the Czech side (after Step 2), and to FrameNet, WordNet, OntoNotes, VerbNet and PropBank on the English side (after Step 3).

The annotators could assign none, one, or more links to an entry in the external resource. Multiple links have been allowed in case they believed that both (or all) such links relate well to the given class member, i.e., in cases where the granularity of the external resource has been finer than the granularity of the PDT-Vallex or EngVallex entries, respectively, in terms of sense distinctions. In the opposite case, when the granularity of PDT-Vallex (or EngVallex) is finer than the external resource entry(ies), the annotators have been asked to simply assign the link to such a more coarse-grained entry, without any special notes or markup. After the lexicon is completed, it will be possible to extract such asymmetric cases by reverting the links.

When comparing the links assigned by two annotators, only a full match (when *all* links agreed, for each external resource individually) counted as agreement, including cases when multiple (or no) links have been assigned by any of the two.

[10] Due to the high number of combinations of valency slots and semantic roles (e.g., for 3 slots and 3 roles, allowing for combined assignment of more slots to one role, or the possibility to leave out any of them, plus to assign any of #sb, #sth or #any, the number of combinations is (3!)x8=48), kappa value comes out very high due to p_e being very low, and is thus not telling much in terms of the agreement.

[11] However, we did take into account simple label renaming, which has not been considered a mismatch.

Annot. pair	A_4-A_5	A_4-A_9	A_5-A_9	Avg.
# of pairs	70	124	84	-
ACT	100%	100%	98.7%	99.6%
PAT	98.3%	98.9%	95.5%	97.6%
ADDR	100%	100%	100%	100%
EFF	100%	100%	86.7%	95.6%
ORIG	N/A	N/A	N/A	N/A
other	86.7%	100%	83.3%	90.0%

Table 5: Accuracy for each valency slot mapping to a semantic role, for 3 annotator pairs in Step 2 (Czech verbs)

Annot. pair	A_4-A_6	A_4-A_9	A_6-A_9	Avg.
# of pairs	87	311	91	-
ACT	100%	99.7%	98.9%	99.5%
PAT	97.5%	99.7%	98.8%	98.7%
ADDR	100%	95.0%	100%	98.3%
EFF	66.7%	33.3%	83.3%	61.1%
ORIG	100%	100%	100%	100%
other	75.0%	96.9%	100%	90.6%

Table 6: Accuracy for each valency slot mapping to a semantic role, for 3 annotator pairs in Step 3 (English verbs)

4.4.1. Interannotator Agreement

Interannotator agreement on linking to the external resources has been measured again as a simple agreement rate (mutual accuracy), taken as the ratio on agreed upon links to the total number of class members annotated by a given pair of annotators. External links have only been annotated for valid class members, i.e., those retained after the manual pruning of automatically preselected class members which is always performed first in each Step (Step 2 for Czech and Step 3 for English verbs in this case).

Tab. 7 shows the agreement for linking to VALLEX, the alternative Czech valency lexicon, which is not used in the annotation of the Czech corpora but developed independently (Sect. 3.2). VALLEX, however, uses almost the same principles as PDT-Vallex for sense distinctions as well as for slot labels, the two features most important here. Only a full match (valency frame ID, or a set of valency frame IDs in case of multiple links) counts as "correct".

Annot. pair	A_4-A_5	A_4-A_9	A_5-A_9	Avg.
# of pairs	70	124	84	-
VALLEX	65.7%	69.4%	71.4%	68.8%

Table 7: Agreement ratio for linking the Czech verbs to the VALLEX lexicon, for 3 annotator pairs in Step 2

While the three pairs of annotators do not differ much in the agreement ratio, it is interesting to observe the relatively low agreement on assigning links to a very closely related resource, possibly caused by the fact that every entry in the VALLEX lexicon has been throughly and widely researched and all possible senses of a given lexeme added, making it both more fine-grained than PDT-Vallex as well as having more senses for each lexeme, including those not found in the underlying corpora. Such a richness made the task of the annotators apparently quite hard.

For the English verbs (as added in Step 3), the results are in Tab. 8.

Annot. pair	A_4-A_6	A_4-A_9	A_6-A_9	Avg.
# of pairs	87	311	91	-
FrameNet	59.8%	60.1%	72.5%	64.1%
WordNet	41.4%	34.4%	27.5%	34.4%
VerbNet	54.0%	57.2%	49.5%	53.6%
PropBank	88.5%	72.0%	80.2%	80.2%
OntoNotes	79.3%	80.4%	91.2%	83.6%

Table 8: Agreement ratio for linking the English verbs to the external lexicons, for 3 annotator pairs in Step 3

As the results show, the agreements ratios vary widely. The relatively high agreement for PropBank and OntoNotes is undoubtedly due to the fact that the EngVallex verb senses come from the same corpus (at least in part), namely the WSJ part of the Penn Treebank, despite the differences in creating the argument structure in ProbBank and the OntoNotes groupings, vs. the valency frames in EngVallex.

FrameNet frame assignment agreements, due to the relatively broad nature of FrameNet frames, are somewhat low.

VerbNet, even though its classes are definitely broader than the synonym sets in SynSemClass, displays very low agreement, which might be caused by mismatches in the assignment to the single class or subclass in the VerbNet hierarchy.

WordNet links display an extremely low agreement, caused, in our opinion, by the very fine-grained distinctions in WordNet verb senses, which often caused multiple WordNet senses being assigned to a single SynSemClass class member. This leads easily to a disagreement between the annotators due to the fact that an agreement is counted as correct only if all links agree (i.e., linking to WordNet senses #1 and #2 by one annotator and to only sense #2 by the other annotator is a mismatch).

5. Converting to Linked Data

Linked Data is a widespread effort to make data available "in context", i.e., to link them to other data, in our case to lexical resources, in order to use the "knowledge" these links add mutually to the individual resources. Our motivation is to be compatible with such lexicons, e.g., resources available in the ELEXIS[12] project, as describe e.g., in (Declerck et al., 2015).

In the work described so far, we have concentrated on the content creation, including the various links, especially to the existing external lexicons. In order to design the structure for providing the data in Linked Data form, we have used the OntoLex (lemon) format[13]. Content-wise, we have been mostly inspired by (McCrae et al., 2014), since in the treatment of word (verb) senses and the view of ontology it is closest to our approach. Similarly, (Corcoglioniti et al., 2016) has a set of modules for PropBank, FrameNet, VerbNet and NomBank, which we will use as well.

For verbs, as already mentioned in the Introduction section, there is no ontology as we can find, e.g., in the medical domain (e.g., the ICD, or various other classifications schemes in MESH), or biology or other domains. In fact, the idea behind SynSemClass is to build an ontology substitute that could be used for a sort of grounding (at least at the event type level) in data (text) annotation. Thus, we treat a class in SynSemClass as substitute for an ontology unit, similar to the treatment of WordNet synset in (McCrae et al., 2014). Each member of the class is a *sense*, denoted by a concatenation of the verb lemma (usually, infinitive form or a concatenated infinitive form of a MWE in case of e.g., phrasal verbs) and the valency frame ID, which is unique in the whole linked dataset, including across languages (cf. the valency frame ID prefix), for example `confirm-ev-w649f1`, while a *word* (LexicalEntry), even though redundant (because reachable through the link to the valency lexicons) is represented by its lemma concatenated by the word ID, e.g., `confirm-ev-w649`, see Fig. 3.

External links are represented as links to the LD versions of WordNet (McCrae et al., 2014) and FrameNet (Bryl et al., 2012), while the links to VerbNet, FrameNet, PropBank frames and OntoNotes sense groupings are represented as URLs (URIs) to their web presence, if it exists in the Unified Verb Index, or as the customary ID with an appropriate lexicon-unique prefix if they do not. While the SynSemClass lexicon does not have a hyponym/hyperonym hierarchy (yet), it will be represented by the *broader* relation as found in SKOS. Grammatical properties (i.e., the mapping between the valency arguments and semantic roles, as a property of each class, will be represented as standard *properties*.

Fig. 3 shows a linked representation of one SynSemClass entry, or more precisely, one sense of the verb *confirm*, identified by its EngVallex reference (`ev-w649f1`). The lower half of the scheme shows the links to external resources, as described above. Please note that each link to each external lexicon can appear multiple times; for example, to simplify the picture, we have left

[12] `https://elex.is/`

[13] `https://www.w3.org/2019/09/lexicog`

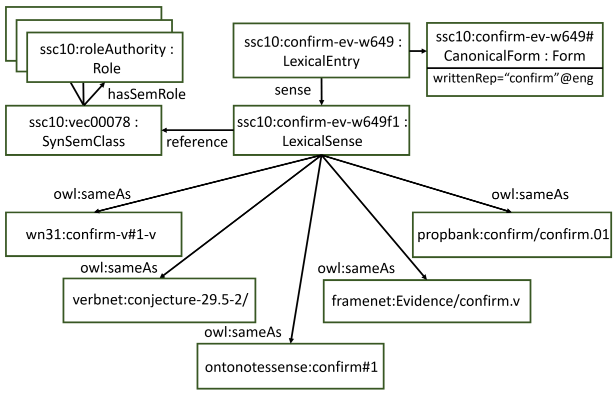


Figure 3: The Linked Data scheme of the SynSemClass lexicon entry (example entry confirm, simplified)

links to WordNet sense `confirm#2` and to two additional FrameNet frames (`Statement` and `Verification`), as identified by the annotators of that entry, and one more VerbNet class (`indicate-78-1-1`). All entries in other languages (Czech, for the moment) will have their own entries, but they will share the reference to the SynSemClass (`ssc10:vec00078` in this case).

6. Related work

We have presented the first version of SynSemClass, a bilingual verbal synonym lexicon linked to several external resources, as an initial step in developing multilingual verbal ontological resource that can link to existing lexical-semantic resources (and resources bordering already on ontologies, such as FrameNet). We are aware of several such projects (or similar ones), such as the Predicate Matrix project (Lacalle et al., 2014), VerbAtlas (Di Fabio et al., 2019) and especially the SemLink work lead by M. Palmer and colleagues (Palmer, 2009; Bonial et al., 2013; Bonial et al., 2012). Our contribution here is the inclusion of valency and its mapping to semantic roles as a major criterion of including a verb (or better, its sense) in a synonym class, while adding the fully manually assigned mapping (linking) to these other resources.
In terms of Linked Data, there has been previous projects, especially for FrameNet, as comprehensively described in (Ide, 2014). WordNet has also several conversion to Linked Data (more precisely, to RDF/OWL), and there is also a description of the model(s) and outstanding issues;[14] previous work on WordNet conversion to Linked Data can be found in `http://xmlns.com/2001/08/wordnet` as well as in (McCrae et al., 2014) (which, among other features, also links to VerbNet).

7. Summary and Future Extensions

Our main contribution is the linking of the (currently) bilingual verbal synonym lexicon in two directions: (a) to deep syntactic information for each verb included (i.e., to the Czech and English valency lexicons), and (b) to existing "popular" lexical resources (i.e., VerbNet, PropBank, OntoNotes groupings, FrameNet and WordNet). We have also presented a suggested mapping to the Linked Data scheme, and shown that all the necessary components are there; this conversion will be physically made once we increase the coverage of SynSemClass and check its consistency (for the moment, the lexicon is available in XML format as an Open Resource).[15]
In the short term, we plan to extend the resource by using both automatic and manual methods and annotation. As shown in this paper, the manual effort involved is relatively large for getting all the components of the lexicon together, and agreement among annotators is not that high, even though it has been improving. Thus the plan is to involve distributional methods (in part using deep learning based on this initial version, e.g., to find more precisely additional synonym candidates in both parallel as well as monolingual texts, including also languages other than Czech and English). We have also started to cooperate on linking the resources from the Unified Verb Index[16] to SynSemClass and vice versa, sharing data in order to minimize the annotation effort needed to enrich both resources.
In the long term, we would like to add connection (including entry to entry links) to additional and newly appearing resources, such as VerbAtlas (as being worked on within the BabelNet project) (Di Fabio et al., 2019), which is in fact very close in goals to the project presented here. This implies adding verbal and event nominals and provide the linking for them, too.
We believe that SynSemClass can be used, already in the current state and coverage, as an evaluation resource for any automated methods and tools for annotation of all three areas: synonym class membership, valency to semantic role mapping, and mapping to external resources.
We also plan to create a textual resource (preferably, a treebank, and ideally, a parallel one or ones) that would be annotated by the classes (and semantic roles associated with these classes) from SynSemClass. Such a resource can then be used to train various NLP tools, from verb sense disambiguation to information extraction to full grounding that would include both entities as well as events/states. In this area, we plan to cooperate with other projects and initiatives that tackle universal or uniform semantic representations, such as the UMR project (Pustejovsky et al., 2019), or the semantic representations that have been used in the CoNLL 2019 MRP Shared Task (Oepen et al., 2019).

8. Acknowledgements

This work has been supported by the GAČR (Grant Agency of the Czech Republic) grants No. GX20-16819X and GA17-07313S. It uses resources hosted by the LINDAT/CLARIN (LINDAT/CLARIAH-CZ) Research Infrastructure, projects No. LM2015071 and LM2018101, supported by the Ministry of Education and Youth of the Czech Republic. The work described herein would not at all be possible without the annotators who have actually made it happen–we would thus like to thank them all: Kathryn S. Conger, Darja Klambotskaya, Vlaďka Krajcsovicsová, Jakub Sláma, Barbora Bulantová, Kateřina Truksasová, Martin Babička, and Alena Straková.

[14] `https://www.w3.org/TR/wordnet-rdf`

[15] `http://hdl.handle.net/11234/1-3125`

[16] `https://uvi.colorado.edu`

9. Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

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.

Bonial, C., Feely, W., Hwang, J. D., and Palmer, M. (2012). Empirically Validating VerbNet Using SemLink. In *Seventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, Istanbul, Turkey, May.

Bonial, C., Stowe, K., and Palmer, M. (2013). Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17, Pisa, Italy, September. Association for Computational Linguistics.

Bryl, V., Tonelli, S., Giuliano, C., and Serafini, L. (2012). A novel framenet-based resource for the semantic web. *Proceedings of the ACM Symposium on Applied Computing*, 03.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Corcoglioniti, F., Rospocher, M., Aprosio, A. P., and Tonelli, S. (2016). PreMOn: a lemon extension for exposing predicate models as linked data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 877–884, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Declerck, T., Wandl-Vogt, E., Krek, S., and Tiberius, C. (2015). Towards multilingual elexicography by means of linked (open) data. In Jorge Gracia, et al., editors, *Proceedings of the Fourth Workshop on the Multilingual Semantic Web (MSW4) co-located with 12th Extended Semantic Web Conference (ESWC 2015), Portorož, Slovenia, June 1, 2015*, volume 1532 of *CEUR Workshop Proceedings*, pages 51–58. CEUR-WS.org.

Di Fabio, A., Conia, S., and Navigli, R. (2019). VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China, November. Association for Computational Linguistics.

Fillmore, C. J., Johnson, C. R., and L.Petruck, M. R. (2003). Background to FrameNet: FrameNet and Frame Semantics. *International Journal of Lexicography*, 16(3):235–250.

Ide, N. (2014). FrameNet and linked data. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 18–21, Baltimore, MD, USA, June. Association for Computational Linguistics.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240.

Lacalle, M. L. D., Laparra, E., and Rigau, G. (2014). Predicate Matrix: extending SemLink through WordNet mappings. In *Proceedings of the 9th conference on International Language Resources and Evaluation (LREC'14*.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Nakladatelství Karolinum, Praha, Czechia.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics, colocated with LREC 2014*, Reykjavik, Iceland.

Oepen, S., Abend, O., Hajic, J., Hershcovich, D., Kuhlmann, M., O'Gorman, T., Xue, N., Chun, J., Straka, M., and Uresova, Z. (2019). MRP 2019: Cross-framework meaning representation parsing. In *Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning*, pages 1–27, Hong Kong, November. Association for Computational Linguistics.

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

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9–15.

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

Pradhan, S. S. and Xue, N. (2009). 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*, pages 11–12, Boulder, Colorado, May. Association for Computational Linguistics.

Pustejovsky, J., Lai, K., and Xue, N. (2019). Modeling Quantification and Scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 28–33, Florence, Italy, August. Association for Computational Linguistics.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Com-*

prehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht.

Urešová, Z., Fučíková, E., and Šindlerová, J. (2016). CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

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

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018b). A Cross-lingual synonym classes lexicon. *Prace Filologiczne*, LXXII:405–418.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018c). Defining Verbal Synonyms: between Syntax and Semantics. In Dag Haug, et al., editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, number 155, pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018d). Synonymy in Bilingual Context: The CzEngClass Lexicon. In *Proceedings of The 27th International Conference on Computational Linguistics*, pages 2456–2469, Sheffield, GB. ICCL, ICCL.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018e). Tools for Building an Interlinked Synonym Lexicon Network. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2019a). Meaning and Semantic Roles in CzEngClass Lexicon. *Jazykovedný časopis / Journal of Linguistics*, 70(2):403–411.

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2019b). Parallel Dependency Treebank Annotated with Interlinked Verbal Synonym Classes and Roles. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 38–50, Paris, France. Université Paris Sorbonne Nouvelle, Association for Computational Linguistics.

Urešová, Z. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

10. Language Resource References

Cinková, Silvie and Fučíková, Eva and Šindlerová, Jana and Hajič, Jan. (2014). *EngVallex - English Valency Lexicon*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, `http://hdl.handle.net/11858/00-097C-0000-0023-4337-2`.

Jan Hajič and Eva Hajičová and Jarmila Panevová and Petr Sgall and Silvie Cinková and Eva Fučíková and Marie Mikulová and Petr Pajas and Jan Popelka and Jiří Semecký and Jana Šindlerová and Jan Štěpánek and Josef Toman and Zdeňka Urešová and Zdeněk Žabokrtský. (2012). *Prague Czech-English Dependency Treebank 2.0*, `https://catalog.ldc.upenn.edu/LDC2004T25`. Also at LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

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

Jan Hajič and Jarmila Panevová and Eva Hajičová and Petr Sgall and Petr Pajas and Jan Štěpánek and Jiří Havelka and Marie Mikulová and Zdeněk Žabokrtský and Magda Ševčíková-Razímová and Zdeňka Urešová. (2006). *Prague Dependency Treebank 2.0*. LDC, number LDC2006T01, `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

Urešová, Zdeňka and Štěpánek, Jan and Hajič, Jan and Panevová, Jarmila and Mikulová, Marie. (2014). *PDT-Vallex: Czech Valency lexicon linked to treebanks*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, `http://hdl.handle.net/11858/00-097C-0000-0023-4338-F`.

Zdeňka Urešová and Eva Fučíková and Jan Hajič and Jana Šindlerová. (2015). *CzEngVallex - Czech English Valency Lexicon*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, `http://hdl.handle.net/11234/1-1512`.

Urešová, Zdeňka and Fučíková, Eva and Hajič, Jan and Hajičová, Eva. (2019). *SynSemClass 1.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, `http://hdl.handle.net/11234/1-3125`.