Mapping Czech Verbal Valency to PropBank Argument Labels

Jan Hajič, Eva Fučíková, Markéta Lopatková, Zdeňka Urešová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské nám. 2/25, 118 00 Prague 1, Czechia
{hajic,fucikova,lopatkova,uresova}@ufal.mff.cuni.cz

Abstract

For many years, there have been attempts to compare predicate-argument labeling schemas between formalisms, typically under the dependency assumptions (even if the annotation by these schemas could have been performed on constituent-based specifications). Given the growing number of resources that link various lexical resources to one another and thanks to parallel annotated corpora (with or without annotation), it is now possible to do more in-depth studies of those correspondences. We present here a high-coverage pilot study of mapping the labeling system used in PropBank (for English) to Czech, which has so far used mainly valency lexicons (in several closely related forms) for annotation projects, under different levels of specification and different theoretical assumptions. The purpose of this study is both theoretical (comparing the argument labeling schemes) and practical (to be able to annotate Czech under the standard UMR specifications).

Keywords: predicate-argument structure, valency, syntax, semantic, semantic roles, PropBank, Prague Dependency Treebank, SynSemClass, Unified Verb Index

1. Introduction

PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005), as it is usually referred to, is an English treebank (usually meant to be the Penn Treebank, Marcus et al., 1993, in its many [later] versions) annotated with a predicate-argument structure, and in addition, with semantic roles, as defined in (Palmer et al., 2005). The individual verbs (further sub-divided into verb senses, and assembled in so-called "PropBank Frame Files"), form a lexicon containing proper set of argument labels for each of their senses. The corpus and the lexicon (frame files), with the annotation guidelines as the main source of the underlying theoretical description, form a specification of a fundamental view of the predicate-argument structure as applied to English. The success of such predicate-argument annotation has led to the creation of several treebanks in other languages annotated in the Penn-Treebank-style, and "propbanked" using the specification for English, such as Arabic, Basque, Chinese, Finnish, Hindi, Persian, Portuguese, Turkish and Urdu, as well as the multilingual collection of the IBM Universal Proposition Banks for 23 languages (Jindal et al., 2022).[1] The PropBank scheme of labeling predicate-arguments has been also used for the Abstract Meaning Representation (AMR) annotation (Banarescu et al., 2013) (with some specific predicates added) and recently, also for the Uniform Meaning Representation (UMR) annotation (Van Gysel et al., 2021; Wein and Bonn, 2023) (with even more abstract predicates added). Both AMR and UMR guidelines[2] call, in principle, for the same predicate-argument labeling scheme as in the original PropBank.

The Czech language valency scheme, essentially also a predicate-argument labeling scheme, is however based on the Functional Generative Description (dependency) theory (Sgall et al., 1986), which treats especially the first two verb arguments differently than PropBank and uses a different specification and labeling style for the remaining arguments. It is used in the main Czech valency dictionaries (Urešová et al., 2014; Lopatková et al., 2022): VALLEX (Žabokrtský and Lopatková, 2007; Lopatková et al., 2016) and PDT-Vallex (Hajič et al., 2003). The latter has been used in the Prague Dependency TreeBank (PDT) in the annotation of the four PDT-C (Hajič et al., 2020) subcorpora annotated on the so-called Tectogrammatical Layer, or Tectogrammatical Representation (TR), which is "deeper" than the traditional dependency syntax used in the Analytical Layer (surface syntax) of the PDT(-C) (Hajič et al., 2020) or in the Universal Dependencies annotation scheme.

At the same time, lexical semantic resources[3] have been increasingly available in an interlinked form. That covers both linking across such lexicons, and/or linking them across languages. An example

[1] The UP 2.0 project creates the resulting annotated files, at least for some languages, by an automatic conversion from the UD-style annotation. This includes all the Czech UD treebanks as well, which means that the resulting labeling depends almost purely on the UD syntactic scheme.

[2] https://umr4nlp.github.io/web/guidelines.html

[3] We are interested primarily in verbal lexical resources, but other resources are being linked together too, e.g., in the Linguistic Linked Open Data project https://pret-a-llod.github.io.

5th Workshop on Designing Meaning Representations (DMR 2024) @LREC-COLING-2024, pages 88–100

of such linking[4] is the Unified Verb Index[5] (Palmer et al., 2014; Stowe et al., 2021) and the multilingual SynSemClass ontology and lexicon[6] (Uresova et al., 2020), which has a rich set of links to PropBank, FrameNet, VerbNet, WordNet for English, and to the VALLEX and PDT-Vallex lexicons for Czech. In addition, the CzEngVallex lexicon[7] (Urešová et al., 2015a) links Czech and English verb entries,[8] using the PDT scheme for Czech. It is also important to note that the EngVallex lexicon, used as a basis for the bilingual CzEngVallex, was built upon PropBank - albeit it also uses the PDT argument labeling scheme - and contains (some) links back to the original PropBank frame files (Cinková, 2006).

The goal of this paper is to describe a recent attempt at a large-scale, large-coverage mapping of the predicate-argument labeling schemas: the PDT-based valency approach and the PropBank approach, applied to Czech. Mapping means to try to capture the same predicate-argument relations (as found in the Czech valency dictionaries) using the PropBank specifications (by mapping the labels of predicate-argument relations). The results will help to see the theoretical differences, and will perhaps also lead to an easier annotation of Czech within the UMR scheme (which also uses the PropBank argument labeling).[9] While there are several theoretical questions to answer, there are also more practical issues and open questions (and benefits if the differences can be explicitly and formally described):

- How is the PropBank approach different from the semantic point of view, especially in the labeling of the first two arguments?
- Can an algorithm be designed to convert, for a particular verb sense, its PDT-based valency structure into the PropBank predicate-argument labeling scheme?
- If yes, what are the biggest differences that cause complications or lead to the impossibility of mapping to the PropBank scheme exactly?
- What information from the richly annotated lexical resources, such as SynSemClass and PropBank, and the associated bilingual corpora between Czech and English can be used?

2. Related Work

Mapping the (English) PropBank scheme to other languages has been researched previously. The PropBanks mentioned in the Introduction have used some form of mapping. For example, Xue et al. (2002) describes a mapping for Chinese. The first comparative study on English and Czech valency draws a comparison between PropBank, LCS Database, and PDT (Hajičová and Kučerová, 2002). Further, for English, the relations between the PropBank arguments and the valency slots as defined in the PDT scheme have been described by Cinková (2006). The resulting EngVallex lexicon has then been used for the tectogrammatical annotation of English in the Prague Czech-English parallel Dependency Treebank (PCEDT,[10] Hajič et al., 2012).

Studies on English-Czech valency using treebank examples or treebank token alignment are described in (Šindlerová and Bojar, 2009; Bojar and Šindlerová, 2010) and resulted in the creation of a bilingual Czech-English valency lexicon - CzEngVallex - described in (Urešová et al., 2015b, 2016). Detailed studies on aligning English and Czech arguments also exist, such as (Šindlerová et al., 2015). However, all these studies compare the valency solely under the PDT labeling scheme.

A comprehensive description comparing Czech PDT-based valency and the English PropBank labeling schema is presented in the papers (Urešová et al., 2014) and (Xue et al., 2014). They provide a detailed inspection of argument labeling differences between Czech and English annotation within the AMR scheme. As the study (Urešová et al., 2014) reveals, the by far most frequent mismatch is caused by different argument labeling. While there is a complete match for most purely transitive verbs, there is a discrepancy for most other verbs since PropBank continues to number arguments of corresponding verbs consecutively but PDT-Vallex attempts the semanticization of argument labels: `ADDR` (addressee), `EFF` (effect) and `ORIG` (origin). These two studies have been made on a very small subset of verb frames: Xue et al. (2014) use only 100 sentences and verbs found in them.

Finally, a detailed study of mappings between the structures used in AMR and those used in UMR are presented in (Bonn et al., 2023). However, here the Czech AMR annotation uses the Czech PDT-Vallex valency lexicon labels, while the English AMR uses the standard English PropBank Roleset Lexicon (Frame Files).

[4] Among others, such as BabelNet (Navigli and Ponzetto, 2012), Predicate Matrix (Lopez de Lacalle et al., 2016), LLOD etc.

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

[6] https://lindat.mff.cuni.cz/services/SynSemClass50, http://hdl.handle.net/11234/1-5230

[7] http://hdl.handle.net/11234/1-1512

[8] https://lindat.mff.cuni.cz/services/CzEngVallex

[9] While UMR does not strictly require the PropBank approach, it is understood that having a unified argument labeling scheme is an advantage.

[10] http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

3. Data Sources

The datasets used for pre-assigning the PropBank-defined arguments to the PDT-based valency frames and their individual slots have been the following:

- **PropBank Frame Files** taken from the current github version of PropBank;[11] see Sect. 3.1,
- **CzEngVallex bilingual valency lexicon** available in the LINDAT/CLARIAH-CZ repository[12] (Urešová et al., 2016), see Sect. 3.2,
- **SynSemClass ontology 5.0**[13] (Urešová et al., 2023), see Sect. 3.3.

In the following sections, we will present the basic structure of these resources stressing the predicate-argument labeling scheme and properties.

3.1. PropBank and PropBank Frame Files scheme

The original Proposition Bank project (Palmer et al., 2005) "took a practical approach to semantic representation, adding a layer of predicate-argument information, or semantic role labels, to the syntactic structures of the Penn Treebank" (Marcus et al., 1999). In fact, one of the original motivations was to define **semantic roles** for the annotation for each verb used in the corpus, with the alterations appearing in the corpus being one of the important points. It was clearly stated that syntax alone is not sufficient to generalize (or, better to say, uniformly annotate) over various forms of expressions to represent the same meaning in relation to the verb arguments. This approach can be demonstrated on the verb *break* appearing in two syntactically distinguished constructions: *John broke the window.* and *The window broke.* In both cases, the affected object is the window, syntactically expressed as an Object in one case and Subject in the other. There are many verbs behaving similarly, such as *play* (*The sergeant played taps.* vs. *Taps played quietly in the background.*)[14] or *load* (*He loaded the truck with hay.* vs. *He loaded hay onto the truck.*).

As a result, PropBank uses an approach (at the top-level abstraction) similar to that of the PDT (Sect. 3.2), i.e., using a list of arguments specific for each verb. However (as opposed to the PDT), PropBank, while using numbered argument roles, defines Arg0 for a prototypical Agent and Arg1 for a prototypical Patient (or Theme), following (Dowty, 1991). I.e., in the aforementioned example of the two uses of *break*, the *window* argument will always be marked as Arg1 to signal the same semantic "position" relative to the verb *break*, regardless of the syntactic structure; as a consequence, in the case of the second example sentence, the verb *break* will have no argument labeled Arg0. Furthermore, each sense of the verb lemma has a separate **roleset** (denoted by an ordinal number attached to the lemma, such as *kick.02*), and they are collected in one frame file for a given lemma.

The original definition of a roleset in the frame files required a description associated with each argument, such as "sayer" for Arg0 of the verb (sense) *suggest.01* or "utterance (suggestion)" for its Arg1, or "chart-maker" for Arg0 of *chart.01* or "thing being charted" for its Arg1.[15] However, these descriptions are not formally defined, so they are unique for each roleset, and not related (much) to the same description at a different roleset.[16] Also, they do not generalize over "content" synonymy (as in *buy* and *sell*, as the original FrameNet did by putting them to a single frame labeled COMMERCE)[17] - the description of Arg0 for *sell* is "seller" while the same description is used for Arg2 of *buy*. Similarly, PropBank does not group what would be called synonyms, e.g., in WordNet (Fellbaum, 1998) - it keeps each lemma (and word sense) separately. However, thanks to mappings from PropBank to VerbNet (Schuler and Palmer, 2005), available in PropBank v3.4[18] or in the UVI index,[19] at least the broadly defined semantic classes as represented in VerbNet can be determined.

3.2. CzEngVallex: Parallel Czech-English Valency Lexicon and the PDT Valency Scheme

CzEngVallex (also CEV) is a bilingual Czech-English verbal valency lexicon (Urešová et al., 2015). It includes 20,835 aligned valency frame pairs[20] and their aligned arguments. This lexicon uses data from the PCEDT corpus and also takes advantage of the existing valency lexicons for both

[11] http://propbank.github.io/v3.4.0/frames/index.html
[12] http://hdl.handle.net/11234/1-1512
[13] http://hdl.handle.net/11234/1-5230
[14] Examples from the (Palmer et al., 2005) paper.
[15] https://github.com/propbank/propbank-frames/blob/main/frames/chart.xml
[16] This is similar to the approach of FrameNet, which also declares that a semantic role defined or used in two different frames should not be taken to mean the same. See SynSemClass (Sect. 3.3) for a different approach.
[17] Currently, FrameNet v2 uses two separate frames, Commerce_buy and Commerce_sell, corresponding to the PropBank approach.
[18] http://propbank.github.io/v3.4.0/frames/index.html
[19] https://uvi.colorado.edu/uvi_search
[20] Each valency frame in the PDT-based valency approach essentially corresponds to one verb sense, therefore, the term "verb sense" and the term "valency frame" are used interchangeably (simplifying the matter somewhat given that there are some cases where the difference matters).

languages (PDT-Vallex and EngVallex).

FGD valency theory. The PDT-Vallex and EngVallex lexicons, and subsequently the CzEngVallex, are built upon the valency theory developed within the Functional Generative Description approach (FGD). As described in detail in (Urešová et al., 2016; Lopatková et al., 2016), in this dependency approach, valency is seen as a property of (some) lexical items, mainly the verb being the core of the sentence, to select for certain complementations in order to form larger units of meaning (sentence, phrase, etc.). The valency characteristics (i.e., the number or arguments and morphosyntactic surface realization of the selected dependent elements constituting the valency structure) are represented in the form of (PDT-)valency frames; these frames are listed in valency lexicons.

The basic characteristics of the FGD valency theory can be found in (Panevová, 1994): it combines the syntactic and semantic approach for distinguishing valency elements. The relation between the governor (primarily verb) and its dependent is characterized by so-called *functors* at the tectogrammatical layer: a functor is a label representing the semantic values of a syntactic dependency relation.[21] There are two axes of classifying the valency modifications in the FGD valency theory: the first axis distinguishes inner participants (arguments) and free modifications (adjuncts), and the other axis distinguishes between obligatory and optional complementations.

There are five "inner participants" (arguments): Actor/Bearer (functor ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF). Out of the five argument types, FGD states that the first two are connected with no specific semantics, contrary to the remaining three ones. The first argument is always the Actor (ACT), the second one is always the Patient (PAT). The Addressee (ADDR) is the semantic counterpart of an indirect object that serves as a recipient or simply an "addressee" of the event described by the verb. Effect (EFF) is the semantic counterpart of the second indirect object describing typically the result of the event (or the contents of an indirect speech, for example, or a state as described by a verbal attribute – the complement). Origin (ORIG) also comes as the second (or third or fourth) indirect object, describing, not surprisingly, the origin of the event (in the "creation" sense, such as to build from metal sheets, not in the directional sense).

FGD valency theory has further adopted the concept of shifting of "cognitive roles". According to this special rule, semantic Effect, semantic Addressee and/or semantic Origin are being shifted to the Patient (PAT) position in case the verb has only two arguments.[22]

In addition to the inner participants, FGD distinguishes about 50 types of semantically determined adjuncts (free modifications), such as temporal, locative or causal. Due to the "free nature" of adjuncts, only the presence of arguments (obligatory or optional) and obligatory adjuncts is recorded in verbal valency frames.

The FGD-based valency lexicons (PDT-Vallex, EngVallex, and CzEngVallex). CzEngVallex (CEV) has been developed together with the PCEDT corpus[23] (Hajič et al., 2012), i.e., a sentence-parallel treebank based on the sentences of the Wall Street Journal section of Penn treebank[24] and their manual translations. This annotation includes also verb sense annotation by links to valency frames in PDT-Vallex (for Czech) and EngVallex (for English).

PDT-Vallex (Urešová, 2011) has been developed as a resource for valency annotation in the PDT. This lexicon is publicly available as a part of the PDT version 2 published by the Linguistic Data Consortium and also separately.[25] The version of PDT-Vallex used for CzEngVallex contains 11,933 valency frames for 7,121 verbs.

EngVallex (Cinková, 2006) was built within the same FGD valency theory and makes use of PropBank, from which it was automatically preconverted and subsequently manually refined and used for the tectogrammatical annotation of the Wall Street Journal section of the Penn Treebank. EngVallex contains 7,148 valency frames for 4,337 verbs.

3.3. SynSemClass Ontology

SynSemClass (SSC) (Urešová et al., 2023) is an event-type ontology for multiple languages. It includes Czech, English (Urešová et al., 2019a)), German (Urešová et al., 2021) and Spanish words and definitions (Fernández-Alcaina et al., 2022). In SynSemClass, contextually-based synonymous verbs in various languages are classified into event-type concepts, or **multilingual synonym classes**

[21] For a full list of all PDT dependency relations and their labels (functors), see (Mikulová et al., 2005).

[22] This can be illustrated on the sentence *The teacher asked the pupil* where the semantic Addressee (*the pupil*) is shifted to the Patient position and thus gets the PAT functor. This rule, when viewed from the annotation point of view, helps to keep consistency at the expense of lower "semantic adequacy".

[23] http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

[24] https://catalog.ldc.upenn.edu/LDC99T42

[25] https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-4338-F

according to the semantic and syntactic properties they display. To have empirical evidence for such classification, SynSemClass is being developed in a "bottom-up" fashion: The candidate verbs for synonym classes are taken from actual examples from parallel English-Czech, English-German, or English-Spanish corpora.

As described in detail in (Urešová et al., 2020, 2019b, 2018a,c,b), SSC synonym classes are characterized by the following main features:

- The **name** of each **class** stands for a single concept (e.g., of *eating*)[26] and corresponds to the verb that represents the prototypical sense, in each of the languages included.
- Each class is provided with a brief general **class definition** in each language included, which characterizes the meaning (concept) of the class.
- For each class, SynSemClass also provides a fixed set of "situational participants", labeled with SSC **semantic roles** common for all the class members in that class. The roles are mapped to the predicate-argument (valency) structure of the individual class members. Thus, they are characterized both meaning-wise (semantic roles) and structurally-wise (valency arguments). When mapping the roles from a given set of participants, each one must be realised as "something" taken from the valency frame of a verb in that class.
- Each verb (sense) included in a given SSC synonym class is linked to one or more lexical resources for the given language. In SSC, there are links to e.g., VALLEX[27] for Czech, FrameNet[28] and VerbNet[29] for English, E-VALBU[30] for German, AnCora[31] for Spanish, among others.

  Further, each verb is exemplified by instances of real texts extracted from translated or parallel corpora. Specifically, data is extracted from the Prague Czech-English Dependency Corpus (PCEDT)[32] for Czech-English, from the Paracrawl corpus[33] for German-English and from the X-SRL dataset[34] for Spanish-English.

[26] This is different from the commonly used term of "semantic classes of verbs" as represented, for example, in VerbNet, where the class is defined much more broadly – such as for all verbs of movement.

[27] https://hdl.handle.net/11234/1-3524

[28] http://framenet.icsi.berkeley.edu/

[29] http://verbs.colorado.edu/verbnet/index.html

[30] https://grammis.ids-mannheim.de/verbvalenz

[31] http://clic.ub.edu/corpus/es/ancoraverb_es

[32] https://ufal.mff.cuni.cz/pcedt2.0/en/index.html

[33] https://paracrawl.eu

[34] https://catalog.ldc.upenn.edu/LDC2021T09

4. Mapping PDT to PropBank

Given the plethora of richly interlinked resources, as described in Sect. 3, one might wonder why the mapping between arguments of verbs in these resources is ever worth investigating and not just a simple technical problem. The reason is first of all the richness of the language itself and its ambiguity as well as redundancy. In addition, the usual problems related to manually annotated and curated resources are present, too: ambiguous guidelines, (low) inter-annotator agreement, not enough details in lexical resource descriptions, evolving resources over time with changing approaches and annotator teams, and their insufficient coverage (Fučíková et al., 2024).

As an example, let's take the Czech word *sídlit* (lit. *to reside*).[35] It has two core arguments (in the base PDT-Vallex resource): `ACT` for the thing residing somewhere, and `LOC` for the location. In the CzEngVallex resource (Sect. 3.2), the entry for *sídlit* is linked to seven different English verbs (anchor, base, be, ensconce, house, locate, and reside), as collected (and manually filtered and annotated) from the annotated parallel PCEDT corpus. The conflicting argument mappings, when traced from CzEngVallex to PCEDT to PropBank, are shown in Fig. 1.[36]

In the SynSemClass ontology (Sect. 3.3), the Czech verb *sídlit* appears in the class named (in English) `locate`. On top of the aforementioned English equivalents coming from CzEngVallex, there are also some additional English verbs (lie, settle, spread, sit) presumably bearing the same meaning as *sídlit*, with yet another set of mappings traced from the original PDT-like ones to PropBank arguments.

The natural question is whether these mappings can be (automatically) consolidated somehow to serve as a basis for a (manually-based) filtering and editing process to arrive at such a set of PropBank-style arguments for *sídlit* that would respect the PropBank guidelines as much as possible. An algorithm that tries to do exactly that and which makes use of the input resources (as presented in Sect. 3) plus the of parallel Czech-English annotated corpus PCEDT for getting corpus-based preferences, is described in the next section.

[35] *sídlit* is relatively simple example, given that from the Czech language perspective, there is only one meaning; cf. the old Dictionary of Standard Czech Language, or SSJČ, at https://ssjc.ujc.cas.cz.

[36] The verb "be" has been left out, since it was treated as auxiliary in the corpus.

Figure 1: Source mapping for the arguments of the Czech verb *sídlit* to its English counterparts

sídlit (PDT-based arguments)	reside	anchor	base locate	house ensconce
ACT	→Arg0	→Arg1	→Arg1	→Arg1
LOC	→Arg1	no mapping	→ArgM-LOC	→Arg2

5. The Mapping Algorithm

The automatic mapping is created in two steps, for all Czech verb senses (valency frames) as found in the PDT-Vallex lexicon:

- collecting, for each frame, all possible mappings by tracing the available resources for *each argument separately* to get its possible PropBank argument label(s), together with frequencies of these mappings in available corpora (Sect. 5.1), and
- consolidating and creating the new, complete PropBank-style rolesets for Czech verb senses, with the right number of arguments and their labels (Sect. 5.2).

For cases where the final roleset cannot be determined unambiguously, we collect statistics from the parallel PCEDT corpus,[37] which is annotated by both the Czech and English valency frames and their arguments. These numbers of corpus occurrences are then used in determining the preferred mapping when displaying it to the annotator for making the final decisions.

5.1. Collecting Instances of Argument Mappings from Existing Resources

There are two main sources where the traces leading from the Czech PDT-based valency frame and its individually taken arguments to the PropBank ones come from: CzEngVallex (CEV, Sect. 3.2) and SynSemClass (SSC, Sect. 3.3).

CEV mapping. The mapping(s) of the PDT-style labels (functors), as listed for each PDT-Vallex valency frame, to the PropBank argument labels are collected from CzEngVallex entries, together with the PCEDT corpus frequencies. For example, the Czech verb *asistovat* (one sense only, with two arguments: ACT and PAT) is linked to two single-sense EngVallex entries *assist* and *support* in CEV.[38] From these two entries and their occurrences on the English side of the PCEDT, the following PropBank arguments[39] have been identified (numbers in parentheses indicate occurrences of these mappings in the English part of the PCEDT):

asistovat	assist	support
ACT	→Arg0 (16x)	→Arg0 (102x)
PAT	→Arg1 (20x) →Arg2 (3x)	→Arg1 (149x)

Thus by performing three "hops" - from PDT-Vallex to CzEngVallex to PCEDT to PropBank - we are getting, for *asistovat*, an unambiguous mapping from ACT to Arg0 (attested 118 times in the corpus) and an ambiguous mapping from PAT to both Arg1 (169 times in data) and Arg2 (three times).

SSC mapping. While using CEV gives us technically simple means to arrive at (unambiguous, or ambiguous (frequency-annotated)) mappings to PropBank argument labels, it only covers the PCEDT data. To exploit another highly relevant resource, we are using SSC to collect mappings for more verbs (verb senses/frames) from PDT-Vallex to PropBank argument labels. Instead of using the direct frame-to-frame mappings available in CEV, we use one of the major SSC features, namely the mapping between the semantic roles (common for each multilingual class, and thus shared by the verbal lexical units in several languages, including Czech and English) and the original verb arguments, taken from PDT-Vallex and EngVallex. From these mappings, we can extract direct functor-to-functor mapping (as if from CEV) and consequently the PropBank argument labels based on the links in EngVallex. Given that the SSC classes are much broader than the direct bilingual verbal links in PCEDT, we can get bigger coverage, but also more ambiguity.

Let's start with the SSC class "commit / dopustit_se" in which all verbs (including the English verbs *blunder* and *commit*) share two semantic roles, "Perpetrator" and "Deed". For *blunder*, SSC maps these roles to ACT and PAT, respectively, and

[37] https://ufal.mff.cuni.cz/pcedt2.0/publications/eng_pb_links.txt - actually, only from the English side as the target side of each of the possible mappings, since the Czech frequencies are irrelevant for this task, given we go through all of the verbs found in the lexicon.

[38] ... because *asistovat* had these two translational counterparts in PCEDT.

[39] Please recall that the English side of the PCEDT is in fact the original WSJ portion of the Penn Treebank with PropBank annotation on top of it, see Sect. 3.2.

afterwards EngVallex traces them to PropBank's Arg0 (2 instances) and Arg1 (1×), respectively; for *commit*, "Perpetrator"→ACT and "Deed"→CPHR are traced to Arg0 (9x) and Arg1 (12x), respectively. In this case (since no ambiguity arises), the Args can then be easily mapped to the arguments of all the Czech verbs from the same class, namely *dopustit_se, dopouštět_se, páchat and spáchat*, because we know which of their valency frame functors correspond to "Perpetrator" and which to "Deed".

However, it is common that the resulting mappings are (even highly) ambiguous. Fig. 2 illustrates the case for the Czech verb "líbit se", which is one of the Czech verbs in the class "appeal / líbit_se" (meaning to "like" or "be pleased by" something).

5.2. Mapping PDT Valency Frames to PropBank Rolesets

The final steps are to suggest the mapping for the whole valency frame to the PropBank-style roleset, incorporating the procedure described above for the individual arguments. Since these are the last steps before the manual pass of obtaining a PropBank-style Czech rolesets, we are describing them more technically by referring to the actual worksheet (table)[40] that will be used by the annotators.

Mappings for individual functors. Each verb record consists of several rows – one identifying the verb sense (roleset) and then one for each (original) functor / PDT argument, followed by an empty row. A description of how the rows and columns are filled is described below.

1. For each verb sense (frame) from PDT-Vallex, create its PropBank ID (column A, `UMR ID`).
   Example: *spolknout* "swallow; eat_up": PDT-Vallex spolknout (v-w6385f1) –> *spolknout-001*.

2. Copy the PDT-Vallex ID and its frame members (functors) to column B (`PDT frame`), with the verb lemma and frame ID on the same line as the PropBank ID, and the argument functors immediately under that.
   Example: *spolknout-001* "swallow; eat_up" gets link to the PDT-Vallex spolknout (v-w6385f1), its two functors are indicated in separate lines, `ACT` (in nominative) and `PAT` (in accusative).

3. If the verb sense occurs in some SSC class(es), put its class ID and its semantic roles to the appropriate rows corresponding to the role-to-argument mapping as recorded in the SSC (column C, `Role_mapping`). Each mapping has the form functor→role, e.g., `PAT→Deed` for *dopustit se* from the SSC class "commit / dopustit_se", see above.
   If the verb belongs to more SSC classes, create one record for each class (see, e.g., the *bouchnout-002* records with the `ACT→Agent` & `PAT→Instrument` mapping for the "bang / prašťit" SSC class and the `ACT→Assailant` & `PAT→Target` mapping for the "hit / třísknout" SSC class).

4. Copy the mappings retrieved via CEV (Sect. 5.1) to column L (`mapping via CzEngVallex`), with aggregated PCEDT occurrences for each Argx.
   Example: for the verb *asistovat*, "assist; support" this column indicates the `ACT→Arg0` mapping with 118 occurrences (102 "inherited" from the verb *support* and 16 from *assist*); further, two mapping options for `PAT` are identified, 169 cases of `PAT→Arg1` (149 from *support* and 20 cases from *assist*) and 3 occurences of `PAT→Arg2` (from *assist*), see Sect. 5.1.

5. Copy the mappings retrieved via SSC (Sect. 5.1) to column N (`mapping via SynSemClass5.0`), with aggregated PCEDT occurrences for each Argx.
   Example: for *asistovat* "assist; support", this column indicates ambiguous mappings of both `ACT` and `PAT` functors:

asistovat	
`ACT→Protagonist`	→Arg0 (166x) →Arg1 (128x) →Arg2 (1x)
`PAT→Event`	→Arg1 (53x) →Arg2 (295x)

Based on the retrieved mappings, the algorithm tries to resolve ambiguities:

6. If SSC and/or CEV provide an unambiguous mapping of individual PDT functors to PropBank arguments, put it to column G, `Unambiguous mapping - SSC and/or CEV`. This is, e.g., the case of the verb *svolat* "assemble" and its `PAT` functor where both CEV and SSC suggest the `PAT→Arg1` mapping (with 197 occurrences collected in CEV and 418 in SSC, the later via "Event" semantic role).

7. If the mappings offered by the SSC and/or CEV lexicons are ambiguous but some prevail (based on PCEDT counts), show them in column H (`Prevailing mapping - SSC and/or CEV`; multiple suggestions are separated by `#`) and report `ambiguity` in column J.

[40] http://hdl.handle.net/11372/LRT-5480

Figure 2: Mapping PDT functors via the SSC roles for the class "appeal / líbit se" to PropBank arguments

appeal / **líbit se**	"Experiencer" →ACT	"Stimulus" →PAT
appeal	→PAT→Arg1 (19)	→ACT→Arg0 (12)
displease	→ACT→Arg0 (1)	→PAT→Arg1 (2)
sit	no PB mapping	no PB mapping
like	→ACT→Arg0 (57)	→PAT→Arg1 (61)
please	→PAT→Arg1 (14)	→ACT→Arg0 (3), Arg2 (4)
		→MEANS→Arg0 (1), Arg2 (5)
Summary for **líbit se**:	ACT→ Arg0 (58), Arg1 (33)	PAT→ Arg0 (16), Arg1 (63), Arg2 (9)

The mapping of a functor is "prevailing" whenever the number of PCEDT instances of the respective mapping is within 10 percentage points of the immediately more frequent suggestion, starting from the highest count.
Example: for *svolat* "assemble", there are two possible mappings for ACT (both corresponding to the "Host" semantic role), namely 310 occurrences of Arg0 and just 1 occurrence of Arg1; the prevailing mapping ACT→Arg0 is suggested as the relevant mapping in column H.

8. If SSC offers an unambiguous mapping for at least some of the functors that differs from the mapping suggested by CEV, the SSC mappings go into column I (Unambiguous SSC mapping (other than CEV)) as SSC is considered more relevant due to its more "semantic" nature. If the SSC mapping is ambiguous, no suggestion is made and disagree is noted in column J.)
Example: with *střetnout_se* "compete" the mapping ACT→Arg0 unambiguously suggested in SSC with 72 occurrences in PCEDT (with the "Competitor" role) is considered as the relevant mapping and copied to column I, disregarding Arg1 mapping suggested in CEV (6 cases).

Final mappings for the whole rolesets. After the above rather bookkeeping steps (providing, at the same time, relevant background information for the annotators), the algorithm continues by deciding which suggestions to actually make to the annotators.

The suggested mapping is a union of those individual argument mappings inserted in the above steps to columns G, H and I (unambiguous, prevailing, and SSC-only mappings), fulfilling these additional "well-formed roleset" criteria:

- The indices of automatically proposed argument labels must be continuous, starting with Arg0 or Arg1 (per PropBank rules); e.g., the sequence Arg0, Arg2, and Arg3 is not a valid roleset (in such a case, the discontinuous Args note is put in column J).
Example: the valency frame corresponding to *hnát-001* "drive; force" consists of three functors, ACT for "Stimulus", PAT for "Affected" and DIR3 for "State_final"). The mapping retrieved from the relevant SSC class "bring / dovést" suggests their correspondence to Arg0, Arg1, and Arg3, respectively, which is not considered "well-formed roleset"; thus, no final mapping is suggested. However, the annotators get highly useful information about prevailing ACT→Arg0 mapping, unambiguous PAT→Arg1 mapping, and possible mappings of DIR3 to (already taken) Arg1 (attested 19x for the given SSC class in PCEDT), Arg2 (attested 22x), and (inapt) Arg2 (attested 37x).

- The PDT-based valency frame as a whole (i.e., all its functors) must be mapped onto arguments (if not, the partial note is put in column J).
Example: with *donášet-003* "inform; snitch", only for one functor (out of 4), possible ACT→Arg0 mapping is suggested in CEV (with 3 occurrences); no roleset is proposed.

- Argument labels do not repeat; e.g., the roleset (Arg0, Arg1, Arg1) is not a valid one (reported as repeated in column J).
Example: the valency frame of the verb *donést-002* "carry" consists of three functors, ACT for "Transporter" semantic role, PAT for "Transported" and DIR3 for "Area 2". While in SSC, ACT is unambiguously mapped to Arg0 (40x in PCEDT) and the PAT→Arg1 mapping prevails (80x), the only suggestion for DIR3 comes from CEV, repeating Arg1. Thus no final roleset is proposed (and information on partial mappings is provided to the annotators).

Mappings that satisfy these criteria are copied to the AUTOMATIC MAPPING column (column D; the SSC-only mappings are preceded with ?). Column

K (source) contains the source of the suggested mapping (czengvallex, ssc or both).[41]

To summarize, the final output has a form of a simple table identifying, for each Czech verb sense from the PDT-Vallex lexicon (columns A, B), its functors/arguments (column B), its SSC class and semantic roles for individual functors (column C), and their automatic mapping to PropBank arguments, whenever such mapping has been considered as reliable enough (column D, with column K substantiating the decision).

Finally, columns E (CORRECTION) and F (COMMENTS) serve as the editable columns for the annotators to eventually fill in. The other columns store the source information from CEV and SSC (whenever available) plus information why it was not possible to suggest the reliable mapping automatically (where relevant, in column J).

6. Statistics And Limitations

While we cannot yet report on the amount of manual work necessary to fill in the gaps caused by the missing, ambiguous or otherwise unusable data, we present here overall statistics about the major cases, especially those mappings where the level of certainty of producing the correct mapping automatically is high.

For the individual functors, as found in the source valency lexicon, PDT-Vallex, and regardless in which valency frame they occur, the following results have been obtained:

	unambiguous	preferred	unmapped	total
functors	9,465	8,579	24,072	42,116
percent	23	20	57	100

The above table shows that about 43 percent of arguments was possible to map to a PropBank argument label automatically with certainty (or as a preferred variant based on corpus usage statistics).

From the full roleset point of view, the situation is less favorable, albeit expected since for a valency frame to be fully mapped to a PropBank roleset, all arguments must be reliably mapped (with an avg. of 2,69 arguments per valency frame):

	auto-suggested	unassigned	total
rolesets	5,085	10,569	15,654
percent	32	68	100

It is however important to note that most of the unassigned rolesets are simply due to missing source-side mappings (in CEV and SSC). When some mapping was available, then the problematic cases have only been a few: 117 ambiguous mappings for a functor to Argx link, 328 for non-continuous numbering or Agrxs in the roleset, 354 for repeated Argx in a roleset, and 1,123 for only partially mapped frames.

Limitation. There is an important limitation for the approach to argument mapping as described in this paper: it needs the richly linked resources as described in the paper, in order to have reliable indications for what frames can be mapped automatically and which can only be proposed as preferred mappings, with the preferences coming from a corpus annotated by the very valency frames that have been used as a starting point.

However, the limitation might be relieved by using only one input resource, which however must at least be linked to PropBank, such as the SynSemClass one. While it can produce ambiguous or partial rolesets, and given the lack of checks against another resource, less reliable results, it can still be considered a good starting point as demonstrated by the fact that slightly more of the extracted mappings came from the SSC than from CEV (by about 200, or 1.3/3.9% from all/auto-suggested rolesets).

7. Conclusions

We have demonstrated that a carefully designed preprocessing for finding automatic mappings from a Czech valency dictionary which is based on a different theoretical approach can still produce many reliable PropBank-style rolesets (32 percent of the original full frames) to be included in a PropBank frame files for Czech. Additionally, the preprocessing produces a table (spreadsheet) with the necessary valency / predicate-argument information and clickable links for the annotators to finish the work manually in an efficient manner. In the future, the resulting Czech PropBank frame files will be used for Czech UMR annotation that follows the original guidelines requiring PropBank-style argument labels. In addition, it will also allow for more direct, large-scale comparison between the two approaches to predicate-argument labeling.

Acknowledgements

The work described herein has been supported by the grant *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. 20-16819X). It has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

[41] For technical reasons, some valency frames recently edited, the older version of which should rather be deleted in the resulting roleset list (greyed rows), are marked as copy in column K.

8. Bibliographical References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Ondřej Bojar and Jana Šindlerová. 2010. Building a Bilingual ValLex Using Treebank Token Alignment: First Observations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 304–309, Valletta, Malta. ELRA.

Julia Bonn, Skatje Myers, Jens van Gysel, Lukas Denk, Meagan Vigus, Jin Zhou, Andrew Cowell, William Croft, Jan Hajič, James Martin, Alexis Palmer, Martha Palmer, James Pustejovsky, Zdeňka Urešová, Rosa Vallejos, and Nianwen Xue. 2023. Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories*, pages 74–95, Washington, D.C., USA. Association for Computational Linguistics, Association for Computational Linguistics.

Silvie Cinková. 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2170–2175, Genova, Italy. ELRA.

S. 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*.

D. Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA and London.

Cristina Fernández-Alcaina, Eva Fučíková, and Zdeňka Urešová. 2022. Annotation guidelines for Spanish verbal synonyms in the SynSemClass lexicon. Technical Report 72, UFAL MFF UK.

Eva Fučíková, Cristina Fernández-Alcaina, Jan Hajič, and Zdeňka Urešová. 2024. Textual coverage of eventive entries in lexical semantic resources. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy. European Language Resources Association/ICCL (to appear).

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague dependency treebank - consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

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 Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).

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 *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.

Eva Hajičová and Ivona Kučerová. 2002. Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 846—851. ELRA.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. 2016. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

Maddalen Lopez de Lacalle, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2016. A multilingual predicate matrix. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2662–2668, Portorož, Slovenia. European Language Resources Association (ELRA).

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, and Lucie Kučová. 2005. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Martha Palmer, Claire Bonial, and Diana McCarthy. 2014. SemLink+: FrameNet, VerbNet and event ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 13–17, Baltimore, MD, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Jarmila Panevová. 1994. Valency frames and the meaning of the sentence. *The Prague School of Structural and Functional Linguistics*, 41:223–243.

Karin Kipper Schuler and Martha S. Palmer. 2005. *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, USA. AAI3179808.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht.

J. Šindlerová and O. Bojar. 2009. Towards English-Czech Parallel Valency Lexicon via Treebank Examples. In *Eighth International Workshop on Treebanks and Linguistic Theories*, pages 185–195.

Jana Šindlerová, Eva Fučíková, and Zdeňka Urešová. 2015. Zero alignment of verb arguments in a parallel treebank. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 330–339, Uppsala, Sweden. Uppsala University, Uppsala University.

Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, James Gung, and Martha Palmer. 2021. SemLink 2.0: Chasing lexical resources. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 222–227, Groningen, The Netherlands (online). Association for Computational Linguistics.

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

Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Jana Šindlerová. 2015a. CzEngVallez – Czech–English Valency Lexicon.

Zdeňka Urešová, Eva Fučíková, Jan Hajič, and Karolina Zaczynska. 2021. Annotation guidelines for german verbal synonyms included in synsemclass lexicon. Technical Report TR-2021-70, ÚFAL MFF UK.

Zdeňka Urešová, Eva Fučíková, and Eva Hajičová. 2019a. Czengclass: Contextually-based synonymy and valency of verbs in a bilingual setting. Technical Report 62, ÚFAL MFF UK, Prague, Czechia.

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. European Language Resources Association (ELRA).

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. A Cross-lingual synonym classes lexicon. *Prace Filologiczne*, LXXII:405–418.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018c. Defining verbal synonyms: between syntax and semantics. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), Vol. 155*, Linköping Electronic Conference Proceedings, pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2019b. Meaning and Semantic Roles in CzEngClass Lexicon. *Jazykovedný časopis / Journal of Linguistics*, 70(2):403–411.

Zdenka Uresova, Eva Fucikova, Eva Hajicova, and Jan Hajic. 2020. SynSemClass linked lexicon: Mapping synonymy between languages. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 10–19, Marseille, France. European Language Resources Association.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020. SynSemClass Linked Lexicon: Mapping Synonymy between Languages. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography (LREC 2020)*, pages 10–19, Marseille, France. European Language Resources Association.

Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2015b. Czengvallex: Mapping valency between languages. Technical Report TR-2015-58, ÚFAL MFF UK.

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, Jan Hajič, and Ondřej Bojar. 2014. Comparing czech and english AMRs. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014, at Coling 2014)*, pages 55–64, Dublin, Ireland. Dublin City University, Association for Computational Linguistics and Dublin City University.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI - Künstliche Intelligenz*, 35(0):343–360.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 23–33, Nancy, France. Association for Computational Linguistics.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of english AMRs to chinese and czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1765–1772, Reykjavík, Iceland. European Language Resources Association.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *Proceedings of COLING 2002*, volume 2, pages 1100–1106, Taipei, Taiwan.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, 2007(87):41–60.

9. Language Resource References

Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Fučíková, Eva and Hajičová, Eva and Havelka, Jiří and Hlaváčová, Jaroslava and Homola, Petr and Ircing, Pavel and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mareček, David and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Novák, Michal and Pajas, Petr and Panevová, Jarmila and Peterek, Nino and Poláková, Lucie and Popel, Martin and Popelka, Jan and Romportl, Jan and Rysová, Magdaléna and Semecký, Jiří and Sgall, Petr and Spoustová, Johanka and Straka, Milan and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Šindlerová, Jana and Štěpánek, Jan and Štěpánková, Barbora and Toman, Josef and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. 2020. *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3185.

Lopatková, Markéta and Kettnerová, Václava and Mírovský, Jiří and Vernerová, Anna and Bejček, Eduard and Žabokrtský, Zdeněk. 2022. *VALLEX 4.5*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University http://hdl.handle.net/11234/1-4756.

Mitchell P. Marcus and Beatrice Santorini and Mary Ann Marcinkiewicz and Ann Taylor. 1999. *Penn Treebank-3 (LDC99T42)*. Linguistic Data Consortium, Philadelphia, PA, USA, ISLRN 141-282-691-413-2.

Urešová, Zdeňka and Alcaina, Cristina Fernández and Bourgonje, Peter and Fučíková, Eva and Hajič, Jan and Hajičová, Eva and Rehm, Georg and Rysová, Kateřina and Zaczynska, Karolina. 2023. *SynSemClass 5.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-5230.

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 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.