Defining Verbal Synonyms: between Syntax and Semantics

Zdeňka Urešová, Eva Fučíková, Jan Hajič, Eva Hajičová

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic

{uresova,fucikova,hajic,hajicova}@ufal.mff.cuni.cz

ABSTRACT

While studying verbal synonymy, we have investigated the relation between syntax and semantics in hope that the exploration of this relationship will help us to get more insight into the question of synonymy as the relationship relating (similar) meanings between different lexemes. Most synonym lexicons (Wordnets and similar thesauri) are based on an intuition about the similarity of word meanings, or on notions like "semantic roles." In some cases, syntax is also taken into account, but we have found no annotation and/or evaluation experiment to see how strongly can syntax contribute to synonym specification. We have prepared an annotation experiment for which we have used two treebanks (Czech and English) from the Universal Dependencies (UD) set of parallel corpora (PUDs) in order to see how strong correlation exists between syntax and the assignment of verbs in context to pre-determined (bilingual) classes of synonyms. The resulting statistics confirmed that while syntax does support decisions about synonymy, such support is not strong enough and that more semantic criteria are indeed necessary. The results of the annotation will also help to further improve rules and specifications for creating synonymous classes. Moreover, we have collected evidence that the annotation setup that we have used can identify synonym classes to be merged, and the resulting data (which we plan to publish openly) can possibly serve for the evaluation of automatic methods used in this area.

KEYWORDS: Synonyms, lexical resource, parallel corpus, annotation, interannotator agreement, syntax, semantics, universal dependencies, valency.

1 Introduction and motivation

While current NLP systems using modern machine learning methods, such as Deep Learning, minimize the amount of linguistic information necessary to build many successful applications, we believe that semantically-based lexical resources are necessary for linking linguistic content to real world knowledge. At the same time, building such resources – so far often largely manual effort – is time-consuming and could certainly be automated, at least to a certain degree, especially when linguistically annotated data (corpora) are available: "High quality linguistic annotation is the vehicle that has, to a large degree, enabled current successful Natural Language Processing systems." (Palmer et al., 2017).

We study verbal synonymy, as defined in existing lexicons. For the purpose of this paper, we have focused on CzEngClass (Urešová et al., 2018a,c), a bilingual verbal synonym lexicon. In CzEngClass, synonymy is specified by lexical meaning of the synonym class members related to their syntactic behavior by using "contextual constraints," which map semantic roles to valency (and its morphosyntactic realization), to constrain the otherwise intuitive process of deciding which verbs (verb senses) can be considered synonymous. As shown in (Urešová et al., 2018c), the inter-annotator agreement, however, is less than satisfactory, and thus a question arises what else could potentially be used to help the lexicon annotator / creator to decide the synonymy question.

To start with something relatively simple, in this paper we explore the influence of syntax (as realized in the Universal Dependencies (UD)[1] (Nivre et al., 2016) annotation standard) on the decisions made by annotators when assigning a verb in textual context (i.e., in a corpus) to a particular synonym class. We are aware that both the intuition of the annotator about the actual meaning of the verb (in textual context) will certainly play a more important role than a purely syntactic phenomena. One could even argue that especially in a bilingual (or multilingual) context, it is clear from the onset that pure syntax cannot be shared among the synonymous verbs. However, even if it is hard to easily turn around the observation that verbs that fall into the same class exhibit similar syntactic behavior (Levin, 1993) (meant, of course, in the monolingual setting), there could be a correlation that might be explored. In addition, there has not been a situation yet when two or more languages share the same annotation specification, as UD now allows. Thus, if a correlation between (dependency) syntax – namely, between the presence of what UD calls "core dependents" (Zeman, 2017) – and the agreement on assigning the synonym class (made more robust by using bilingual Czech-English material, corresponding to the way CzEngClass is being built) will be found, the findings could possibly be used in e.g., extension of the lexicon with additional (semi-)automatic means. Moreover, it could possibly also bring some insight to the problem of synonymy definition and detection itself. We approach this question by means of a parallel corpus-based verb sense annotation, followed by investigating a correlation between such annotation and the verb's immediate dependency syntax context (as represented by its core dependents). We realize that UD core dependents are only a subset of all possible syntactic features, but we believe that (due to their "coreness", which includes some oblique dependents as well, see Sect. 5.1) they do represent the most of those features that can be reasonably generalized. On the other hand, morphological features of verb dependents are, based on our experience with several past NLP tasks, too language specific to be taken into account in the context of synonymy investigation (even though with more data, it would certainly be appropriate to test them as well).

[1] Universal Dependencies is a project that is developing cross-linguistically consistent treebank annotation for many languages (`http://universaldependencies.org/introduction.html`).

The structure of the paper is as follows. In Sect. 2, the main direction of the paper is motivated by arguing that both semantics and syntax necessarily play an important role in the definition of synonymy, and then the specification of the synonym classes in CzEngClass, which is used as the main dataset for the subsequent experiments, is briefly described. In Sect. 3, the parallel, UD-annotated corpus used in the present experiment is introduced. The annotation part of the experiment is described in Sect. 4, together with basic quantitative evaluation of inter-annotator agreement. Next, we describe annotation post-processing to arrive at a dataset needed for exploring the role of syntax in verb assignment into classes. This exploration and its results are discussed in Sect. 5 and the side effect of the present annotation experiment in Sect. 6. We conclude in Sect. 7, where also some possible next steps and future investigation directions are identified.

2 Synonymy

2.1 Syntax and Semantics in the Definition of Synonymy

Functionally-oriented linguistics understands syntax as an inseparable part of semantics and pragmatics (Wu, 2017). Both within and across languages, there are strong regularities in the relationship between thematic roles and syntactic functions (Bowerman, 2002). Those rules that link thematic roles such as agent and patient to syntactic functions such as subject and direct object are called “linking rules,” cf. e.g. (Alishahi and Stevenson, 2010). These correspondences has been extensively studied, e.g., (Levin, 1993; Levin and Hovav, 2005; Kettnerová et al., 2012; Čech et al., 2015), but there are still differences in opinions on the nature of these rules (Hartshorne et al., 2010).

In our opinion, to describe the interplay of syntax and semantic relations in a sentence by linking syntactic (predicate-argument structure) and semantic (semantic roles) phenomena is essential in order to understand the meaning of a sentence. In this context, we believe, in line with (Urešová et al., 2018a), that it is inadequate to analyze just one structure - whether syntactic or semantic, cf. (Pustejovsky, 1991). However, in order to study the relative importance of these two components, we will focus on how much the knowledge of (purely) syntactic structure can help to correctly assign a synonym class to an occurrence of a verb in running text (corpus).

2.2 The CzEngClass synonym lexicon

We are using CzEngClass (Urešová et al., 2018a,c,d) as a lexicon containing verb synonyms. The grouping of verb senses in this lexicon is based on context expressed by a common set of semantic roles assigned to each class and their ability to map these roles to valency arguments of the individual verbs in that class. In other words, this lexicon acknowledges the role of syntax and semantics (in its treatment of synonymy) as described in Sect. 2.1. The lexicon is currently bilingual, meaning that each class contains both Czech and English verbs (synonymous verb senses), presumably mapable into the same ontological concept in both languages. The entries also contain examples and links to external lexical resources in which the referenced verbs are described from various other perspectives (FrameNet (Baker et al., 1998; Fillmore et al., 2003), VerbNet (Palmer, 2009), PropBank (Kingsbury and Palmer, 2002), WordNet (Fellbaum, 1998), and English and Czech valency lexicons (Cinková, 2006; Hajič et al., 2003)). CzEngClass currently contains 200 classes grouping approx. 3500 verb senses. Part of the lexicon (60 classes) has been tested for inter-annotator agreement (Urešová et al., 2018c), which appears to be low by the standard metrics (Cohen’s and Fleiss’ kappa), but some other metrics developed specifically for lexicon annotation show more positive figures (Sect. 7.4 of (Urešová et al., 2018c)).

Table 1 illustrates a (simplified) class and the context (semantic role to valency argument mappings) for each member, as taken from (Urešová et al., 2018b). While some entries in Table 1 show very regular mapping from valency to the class' semantic roles (one English verb, *protect*, and two Czech ones *chránit, ochraňovat* have the same valency - ACT, PAT, EFF), others show that different syntactic structures can have the same semantic roles (i.e., the same meaning), and presumably also vice versa. Please note that the "deep syntactic" valency labels, as opposed to the UD annotation of core dependents, already represent certain "normalization", for example for passivization and other diatheses etc.).

	Roles		
	Agent	Asset_Patient	Harmful_situation
protect	ACT	PAT	EFF
conserve	ACT	PAT	#sth
insulate	ACT	PAT	ORIG
...	...	...	...
chránit	ACT	PAT	EFF
bránit se	ACT	ACT	PAT
zaclonit	ACT	PAT	ADDR
ochraňovat	ACT	PAT	EFF
zastávat se	ACT	PAT	#sb,REG
...	...	...	...

Table 1: Role-to-valency mappings for the PROTECT class

This leads us to the formulation of the main question, whether there is any correlation between the syntactic behavior of the verbs in the synonym classes and the agreement in assigning these classes to verb occurrences in a running text (i.e., when effectively doing the word sense disambiguation task). Should there be a strong correlation, perhaps the synonymy relation for grouping verb senses in classes could be defined in simpler terms - using just dependency syntax instead of the complex semantic-roles-to-valency mapping, as used in CzEngClass.

3 The Corpus and Text Selection for the Experiment

For the annotation experiments described below, we use the Czech and English parallel treebanks ("PUD") from the UD collection, version 2.2 (Nivre et al., 2018). This corpus has been selected since according to (Urešová et al., 2018c) it has not been used in the semi-automatic CzEngClass version 0.2 (Urešová et al., 2018b) creation (avoiding thus the "training data bias"); at the same time, it is an aligned parallel corpus that is annotated for UD dependency syntax, allowing the type of experiment we are aiming at. We have selected aligned pairs of sentences in which at least one verb has been found in at least two of the 60 synonym classes which are fully annotated for synonym class membership in CzEngClass, version 0.2. Such verbs are deemed to have at least two different senses (Urešová et al., 2018a). Together with its aligned verb in the other language they form a pair, and together with the sentences in which they have been found they form an *annotation item* (more precisely, a pair of annotation items).

For each annotation item, classes have been pre-selected based on the verb in question. Each annotation item consists of between two and five possible classes to be assigned by the annotators, as described below in Sect. 4. However, the verbs in the opposite language have been removed from the set of class members presented to the annotators; i.e., the annotators have not seen the annotation item from the other language. On each side, an "Other" class has been added to allow for annotating cases when no suitable pre-selected class was deemed to be adequate for the meaning of the annotated verb in the context of the sentence in which it occurs.

4 The Annotation Experiment

4.1 The Task

We have hired ten annotators, and divided the annotation items in such a way that each annotation item has been annotated by three of them. As mentioned above, the annotation items have been monolingual (but we did keep the alignment information between the pairs, even though hidden to the annotators), i.e., an annotator could only see the Czech sentence and Czech class members of the classes offered, and similarly for English annotation items (Fig. 1).

The annotation task was to determine the word sense of the marked verb occurrence in the text. More precisely, the annotators have been asked to check one or more of the verb classes offered, or to check the "Other" option, which has been always present. They were not allowed to check "Other" *and at the same time* one or more of the classes for any of the annotation items (i.e., sentences with a particular verb occurrence) being annotated.

The five annotators were given 449 English verbs (139 different) used in 358 different sentences and another five annotators were given 448 Czech verbs (187 different) used in 358 translated sentences.[2] The annotators have not been given the definition of synonymy as presented in CzEngClass. Instead, they have been instructed to determine the correct class intuitively by extracting the "meaning" of the presented tentative class from all the members (although they can be ambiguous just by seeing the lemma, or base form as recorded in the class), much in the way the "meaning" of a WordNet synset can be understood from all the synset members.

On the other hand, several classes may actually have the same meaning, although a few verbs might only be in one or the other, with a majority of verbs being shared between the classes. This happens due to the fact that the CzEngClass lexicon is still under development, as described in (Urešová et al., 2018a), and two classes might still be merged later (we will discuss the side effect of the present annotation experiment in determining which classes should probably be merged in Sect. 6). For these reasons, annotators were allowed to assign more than one class to the verb occurrence in text. Conversely, they could use "Other" when no other class seemed to adequately fit the meaning of the verb in question.

The annotators marked the class(es) they believed to fit the meaning of the verb in question (marked by the double square brackets), or selected the "Other" class by simply putting and 'x' after the colon delimiting the class ID (or the "Other" label).

Since each annotation item has been annotated by three annotators, we have automatically determined the final annotation by taking a majority (of the three) as the final assignment of the verb to its class(es). This majority voting has been done per class offered at each annotation item and language separately.

After this step, each annotation pair has been marked for a *complete match* (when all classes marked agreed between the two languages, or "Other" was selected on both sides) or for a

[2] Some sentences contained two or more verbs, sometimes linked to the same verb in the paired sentence, thus the difference in the number of annotation items. In fact, there were 450 aligned annotation items altogether.

```
ID: n01020004#7
Lemma:  see
Sentence:  Previously the jets had only been [[seen]] by bloggers.
vec00032:  assume, believe, consider, feel, figure, see, think
vec00092:  anticipate, assume, believe, call, envisage, expect, figure,
foresee, predict, presume, project, see, suppose
vec00115:  examine, eye, follow, follow_up, look, monitor, observe, pursue,
see, study, trace, track, track_down, watch
vec00192:  analyze, determine, discover, find, find_out, learn, see
other:
```

Figure 1: Example annotation item as presented to the annotators

partial match (when at least one of the classes was the same for both languages). Any complete match was also considered a partial match, but indeed not vice versa.

4.2 Interannotator Agreement

Having three annotation per annotated item, we computed both Cohen's kappa (pairwise) and Fleiss' kappa (for all annotations).

Evaluation has been done not on the class level, but on the individual membership question annotation level. In other words, the evaluated *annotation decisions* have been just yes/no for each class offered (or the "Other" label). For each annotation item, there have been at least three such decisions. The example in Fig. 1 shows five annotation decisions, four for the classes `vec00032, vec00092, vec00115, vec00192` or one for `other`. We have separately evaluated three types of agreement: on all decisions points (1762 yes/no decisions in Czech, 1763 in English), on decisions for classes offered to the annotators (i.e., excluding the `other` choices; 1314 decisions) and for an estimate on non-news data, we limited the evaluation to classes excluding "verbs of saying" (which we noticed there have been many in the PUD corpus used, and these are apparently easy to do, padding the numbers for the better too much; excluding further 400 decisions points for those, this evaluation has been performed for 914 annotation decisions).

4.2.1 Interannotator agreement: Cohen's kappa

Cohen's kappa, used for comparing annotations pairwise, is defined as follows:

$$\kappa = (p_0 - p_e)/(1 - p_e) \tag{1}$$

where p_0 is the observed agreement, and p_e the expected agreement based on the two annotations. Results are summarized in Table 2.

From these figures, it is clear that agreement on the individual annotation decisions is low, especially for English the group of annotators labeled A_3 vs. the other two annotator groups. Further inspection has shown that in many cases, the decisions have not been "off" completely, but using only one sentence context (and intuition only based on the composition of the class, and without access to the often defining words in the other language) is difficult. The way the PUD sentences and the classes offered have been selected also contributed to the difficulty (only ambiguous classes have been used). Nevertheless, for the purpose of the syntax correlation

Czech Annotator group	A_2	A_3	A_{random}
A_1	0.630 (0.624, 0.587)	0.650 (0.663, 0.603)	0.034
A_2		0.641 (0.660, 0.572)	0.056
A_3			0.032

English Annotator group	A_2	A_3	A_{random}
A_1	0.600 (0.564, 0.605)	0.514 (0.527, 0.480)	-0.007
A_2		0.461 (0.452, 0.427)	-0.006
A_3			0.053

Table 2: Interannotator agreement: Cohen's kappa, pairwise; the three figures for each pair of annotators correspond to all (non-other, non-other and non-speech) decision points. The last column shows Cohen's kappa computed against a random baseline, for each annotator group (for all annotation items), run as a sanity check.

experiment, this seemed sufficient due to the majority voting mechanism as described in Sect. 4; the comparison against the random baseline also shows that the annotators do contribute significantly their langauge and world knowledge.

4.2.2 Interannotator agreement: Fleiss' kappa

Fleiss' kappa can be used in a multiannotator setup with n (three in our case) annotators and N datapoints (1762/1763, 1314 and 914 datapoints for the three evaluations, as defined in Sect. 4.2.1):

$$\kappa = (\bar{P} - \bar{P}_e)/(1 - \bar{P}_e) \tag{2}$$

where

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n(n-1)}(n_{iYes}(n_{iYes}-1) + n_{iNo}(n_{iNo}-1)), \tag{3}$$

$\bar{P}_e = {p_Y}^2 + {p_N}^2$, $p_Y = \frac{1}{Nn}\sum_{i=1}^{N} n_{iYes}$ and $p_N = \frac{1}{Nn}\sum_{i=1}^{N} n_{iNo}$. The n_{iYes} (n_{iNo}) counts are defined as the number of times the i-th annotation decision has been annotated Y and N, respectively (i.e., sum of n_{iYes} and n_{iNo} is n). The Ys correspond to an 'x' being marked by an annotator next to the class they deemed as being the one corresponding to the meaning of the verb in question in the annotated sentence (Fig. 1), and an 'N' means the class (incl. the `other` one) has *not* been marked by the 'x'.

Table 3 shows the values of Fleiss' kappa for the three agreement evaluations performed.

The Fleiss' kappa figures show that the annotator agreement is lower for English, which is explained by the annotator group A_3 disagreement with A_1 and A_2. They also show that contrary to expectations, the exclusion of the `other` label did not lower the agreement dramatically; for

Czech

# of decisions	p_Y	p_N	$\bar{P}$	$\bar{P}_e$	κ
1762	0.353	0.647	0.836	0.543	**0.640**
1314	0.389	0.611	0.833	0.524	**0.649**
914	0.332	0.668	0.817	0.556	**0.587**

English

# of decisions	p_Y	p_N	$\bar{P}$	$\bar{P}_e$	κ
1763	0.352	0.648	0.783	0.544	**0.524**
1314	0.362	0.638	0.774	0.538	**0.511**
914	0.308	0.692	0.788	0.574	**0.502**

Table 3: Interannotator agreement: Fleiss' kappa; the three figures for each pair of annotators correspond to all (non-other, non-other and non-speech) decision points

English it did but for Czech it increases, even if insignificantly. The additional exclusion the frequent "verbs of saying" in direct and indirect speech constructions did lower the kappa, but only marginally and more in Czech than in English.

The effect of the A_3 group has been minimized when preparing the final dataset for the experiments with syntax correlation by simply using majority voting (the two or three of the 3 decisions that agreed ('Y' vs. 'N', or 'x' vs. nothing) have been selected and used as "gold data".

5 Correlation of Syntax and the Synonym Class Annotation

5.1 The Data

The data have been prepared as described in Sect. 4. For each annotated item (a verb in a sentence context), for both languages (Czech and English), we knew the "gold" class(es) annotated, and therefore also knew if the Czech and English class(es) are the same or not. Due to the possibility of having more than one class per verb, we have established two variants of class agreement: complete match (all classes selected for the Czech verb match all the classes selected for the aligned English verb), and partial match (at least one class matches across the languages). We have computed the correlation to syntactic properties, as described below, always for both.

The syntactic features used to check if there is any correlation between them and the agreement (or disagreement) between the two languages have been the presence or absence of the verb's "core arguments" as defined in the UD specification (Zeman, 2017).[3] The following dependency relations (DEPRELs), as used in the Czech and English UD annotation, are considered core arguments: `nsubj, csubj, obj, iobj, ccomp, xcomp, expl`, including cases when they have a language specific extension (e.g., `nsubj:pass or obj:caus`). In addition, `obl:arg and obl:agent` are also considered core arguments.[4]

[3] `http://universaldependencies.org`

[4] For detailed UD syntactic relations see http://universaldependencies.org/u/dep/, where `nsubj` is a nominal which is the syntactic subject, `csubj` is a clausal syntactic subject, `obj` is an object, `iobj` is an indirect object (mostly in the dative case), `ccomp` (clausal complement) is a dependent clause which functions like an object of the verb, `xcomp` (open clausal complement) is a predicative or clausal complement without its own subject, `expl` (expletive) captures expletive or pleonastic nominals that appear in an argument position of a predicate but which do not themselves satisfy any of the semantic roles of the predicate, `obl:arg` is used for prepositional objects, and `obl:agent` for agents of passive verbs.

Example of the resulting data (four annotation items) is in Table 4.

			match	
item	cs core args – en core args	cs ID – en ID	complete	partial
...	...	...	...	...
139	ccomp nsubj – nsubj	n01145028#11 – n01145028#15	1	1
140	ccomp nsubj – nsubj obj	n01059054#19 – n01059054#24	0	0
141	ccomp nsubj – nsubj xcomp	n01108003#2 – n01108003#3	0	1
142	ccomp nsubj – nsubj xcomp	n01136006#4 – n01136006#4	0	0
...	...	...	...	...

Table 4: Section of the correlation dataset (ordered by core arguments)

Complete match always implies partial match, as in item 139. Items 140 and 142 did not exhibit even partial match between the class assigned to the Czech and English verb.

Such data can be used in several ways. We have looked at them from two points of view, to get some insight into the following two research questions:

- if we know the core argument pattern, or perhaps only some of its features, can we predict whether there will be an agreement in assigning the Czech and English class to a given occurrence of a verb (even if annotated independently of each other, as described in Sect. 4)?
- if we know the core argument pattern, or perhaps only some of its features, can we predict the difficulty of predicting a match?

While answering the first question might give us the possibility to extract a set of argument patterns and/or their features, for which we can get better predictions and thus need less human annotation (because we can then use just a single class for both Czech and English annotation in a parallel corpus, achieving higher accuracy upfront), answering the second question in the positive would mean that we can extract a set of core argument patterns that are difficult to predict even from the point of view of agreement (regardless whether there is one or not), and those will definitely need human intervention and investigation.

5.2 Prediction of a Match

For the investigation of the first question, we use the simple conditional probability of a match between the Czech and English annotation given the core argument structure (or some projection into simpler features).

We have investigated systematically several such distributions, conditioned on the following:

- the whole pattern
- one of the arguments, or some of its features (passivization)
- a combination of two or more arguments, either on Czech side, the English side, or both

Some of the above conditionings split the data into many parts, some of them very small (with a long tail of singletons, as usual). We have thus excluded any split that displayed less than 5 datapoints.

	configuration	probability
core argument(s) configuration	probability	of a complete match
exact cs–en: `ccomp - nsubj`	0.018	0.875
cs side contains: `csubj`	0.016	0.860
en side contains: `iobj`	0.016	0.860
exact cs–en: `- nsubj obj`	0.013	0.833
exact cs–en: `- nsubj`	0.018	0.750
>2 arguments on both sides	0.024	0.727

	configuration	probability
core argument(s) configuration	probability of	a partial match
exact cs–en: `nsubj obj iobj - nsubj obj xcomp`	0.011	1.000
cs side contains: `csubj`	0.016	1.000
en side contains: `iobj`	0.016	1.000
exact cs–en: `ccomp - nsubj`	0.018	1.000
exact cs–en: `nsubj ccomp - nsubj`	0.051	1.000
exact cs–en: `nsubj ccomp - nsubj ccomp`	0.076	1.000
cs side contains: `nsubj ccomp`	0.156	0.971
en side contains: `nsubj ccomp`	0.136	0.934
cs side contains: `ccomp`	0.249	0.911
>2 arguments on both sides	0.024	0.909
en side contains: `ccomp`	0.153	0.899
>2 arguments on cs side:	0.084	0.895
exact cs–en: `nsubj obj -`	0.016	0.857
2 arguments on both sides	0.251	0.832

Table 5: Best predictors of a match in terms of core arguments and/or core argument features (complete match in the upper part, partial match in the lower part)

Core argument combinations with high probability (over 70%) of a match between Czech and English synonym class when annotating a given verb in a textual context are listed in Table 5.

It is clear from this table that while some of the core argument features predict a match quite well, they are very infrequent (and thus do not help much overall) due their specificity, e.g., “no argument configuration” on the Czech side (for `- nsubj obj` and `- nsubj`), or the rare `csubj` on the Czech side, etc. Only when we resort to partial matches (hoping that after the synonym dictionary is cleaned and merger candidates actually merged, they will become complete matches), there are some more frequent situations which result in high probability of a partial match, e.g., if `ccomp` appears on the Czech side (which happens in 24.9% of the annotated sentences), or the situation when there are exactly 2 arguments on both sides (which happens in 25.1% of them). Closer inspection has shown, however, that most of the cases where core arguments contain `nsubj` and/or `ccomp` on either the Czech or English side are, not surprisingly, the “verbs of saying”, which are easy to classify.

None of the other combinations of core argument configurations and/or their features (such as passives) showed more than 70% prediction of a complete match (80% in the case of a partial match), excluding those occurring less than five times, as stated above.

5.3 Correlation (Mutual Information)

In the previous section, we have used the conditional probability of a match given a core argument configuration as a measure to find “good” predictors. This measure is intuitive, but it does not find *globally* “good” predictors. This is better accomplished by using correlation measures. We have used both the Pearson correlation and Mutual Information (Cover and Thomas, 2006). Since the results are very similar, we will resort to Mutual Information only to determine which core argument configurations predict the (complete or partial) match better.

Mutual Information (MI) is defined as the entropy reduction when a variable X is used to predict Y (as opposed to not using the variable X). It can thus be used to compare the contribution of various variables X_k to the prediction of Y. Using probability distribution(s), it is defined as

$$MI(X_k, Y) = \sum_{i=1}^{|X_k|} \sum_{j=1}^{|Y|} p(x_{k_i}, y_i) log_2 \frac{p(x_{k_i}, y_i)}{p(x_{k_i}) p(y_i)} \tag{4}$$

MI is a “global” measure, thus it helps to determine which X_k works best over all data, since it includes the proper distribution of weights, equal to the joint probability of the various configurations ($= X_k$s) and the prediction (0 or 1, in our case). MI has always a positive value. In our case, the higher it is, the more the tested core argument configuration of features as represented by the X_k variable helps to predict the match.

We have computed Mutual Information between the same large set of possible core argument configurations as in Sect. 5.2, using the presence or absence of the given configuration or feature (represented as 0 or 1) as the predictor, and also computed the Mutual Information of the set of full individual core argument configurations. The results are in Table 6.

Using all possible configurations reduces the prediction entropy the most, as expected. There are 128 different cs – en configurations in the data. Unfortunately, only 20 of them occur more than in 1% of the data, and the easily predicted (in)direct speech verbs occur in more than 30%, coinciding with the more frequent core argument configuration (typically containing or solely consisting of `nsubj` and `ccomp`). So while these predict the “training data” well, problems will arise on previously unseen data. This is normally alleviated by using the more robust binary features that only look at the presence of absence of individual core arguments on one or the other side (cs or en). However, as Table 6 shows, these have very low values of MI, meaning that they only marginally contribute to the prediction certainty. In addition, they mostly apply again to the (in)direct speech verbs, with only a few exceptions, such as the presence of the `obl:agent` relation in Czech or of an expletive relation (Bouma et al., 2018) on the English side.

To summarize, we consider these to be negative results, essentially confirming that syntax alone (as represented by the UD core arguments) cannot be used to tell the easy from the difficult cases of assigning a common class to an aligned pair of verbs in English-Czech translation.

6 Side Effect: Classes to be Merged

Since the CzEngClass version 0.2 lexicon used (Urešová et al., 2018b) is still under development and the classes are being created independently of each other, based on independently determined seed words (Urešová et al., 2018a), some pairs of classes are very similar.

core argument(s) configuration	MI between configuration and a complete match
all configurations	0.336
cs side contains: `nsubj ccomp`	0.010
cs side contains: `obl:agent`	0.008
cs does not contain passive, en side does (`:pass`)	0.006
exact cs–en: `nsubj ccomp - nsubj ccomp`	0.005
en side contains passive (`:pass`)	0.005

core argument(s) configuration	MI between configuration and a partial match
all configurations	0.285
cs side contains: `nsubj ccomp`	0.043
cs side contains: `ccomp`	0.032
exact cs–en: `nsubj ccomp - nsubj ccomp`	0.030
cs side contains: `nsubj ccomp`	0.022
en side contains: `ccomp`	0.015
en side contains: `expl`	0.010
cs side contains: `iobj`	0.007
>2 arguments on cs side	0.007
2 arguments on both sides	0.006
cs side contains: `csubj`	0.006
cs side contains: `iobj`	0.006
en side contains: `nsubj`	0.005

Table 6: Best global predictors (based on MI) of a match in terms of core arguments and/or core argument features (complete match in the upper part, partial match in the lower part)

The annotation as set up for the purpose of this paper (Sect. 4) can be used to identify those classes that are hard to distinguish, thanks to the fact that annotators may check multiple classes for each annotated item.

Statistics about the multiple selections suggest to merge several pairs of classes. These include the verbs introducing direct or indirect speech, which are frequent in our corpus (and which we have mentioned a few times already) and currently in two classes (they include verbs like *say, announce, declare, disclose, report, post, advise,* etc.). Other pairs (and one triple) with high annotator agreement include (only English verbs from these classes shown):

- allow, enable, permit, ... vs. approve, acknowledge, reaffirm, ...
- anticipate, forecast, foresee, ... vs. expect, await, pend, ...
- avoid,[5] bypass, circumvent, get around, ... vs. avoid, discourage, preclude, prevent
- (triple) encourage, galvanize, inspire, ... vs. spark, launch, set off, ... vs. aid, bolster, encourage, facilitate, ...

[5] An interesting example is the "avoid" example, which could go unnoticed if only the lexicon classes themselves are investigated, since it looks at the same event from different perspectives (with distinct different syntactic realization). Annotation on the PUD corpus has however revealed its closeness if not identity from the semantic point of view.

These results confirm that synonym classes might be impossible to define categorically, or mathematically speaking, as equivalence classes; it points rather to something like “soft class membership” function or other means to relate related words which we call, perhaps incorrectly, “synonyms”, and it suggests to depart from thinking of every two words as strictly either being synonyms or not. This naturally leads to considerations of “weighted” synonymy, or to introduce a distance between every two (senses of) a verb (or any other word) which reflects their synonym(it)y; there are plenty of measures to experiment with, starting with cosine similarity to modified “soft” Brown classes (Brown et al., 1992) to distance measures over various types of embeddings, such as (Mikolov et al., 2013); this is however outside of the scope of this paper and will be our future work on this topic.

7 Conclusions and Outlook

We have performed an experiment using manual annotation followed by sort of “data mining” to determine whether the presence or absence of (syntactic) core arguments (as defined in the UD specification) can be used in determining the difficult cases of assignment of a synonym class (defined in semantic terms). While expecting that syntax alone cannot be used for such a prediction, we were surprised by a correlation so low.[6]

As a positive side effect outcome of the annotation experiment performed, we have clearly identified candidate class pairs (and one triple) for a merger. It remains to be seen if the small dataset that we have created could be of any use when developing automatic methods for such mergers. We believe that it can at least serve as a test corpus. This part of the experiment has also shown that it might be better (or even necessary) to introduce “soft” (or “fuzzy”) membership in classes.

It has been confirmed that using bilingual data (parallel corpus) provides important information; in our experimental setting we could not automatically identify “matches” (to be used as “gold truth” for the syntax correlation experiments) without using them.

The main conclusion is that for the actual assignment of classes (i.e., not only the prediction of the *difficulty* of such assignment, as demonstrated in this paper), which is certainly a more difficult task, it will not be sufficient to look at (dependency) syntax only, and that semantically-based methods will have to be used, such as the use of automatically derived classes and their bilingual pruning, or the use of embeddings trained on large corpora.

Acknowledgments

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic. The data used in this work have been created and are maintained in the LINDAT/CLARIN digital repository (`http://lindat.cz`), supported by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS) as projects No. LM2015071 and CZ.02.1.01/0.0/0.0/16_013/0001781. In addition, some of the data used has been the result of the “T4ME” project funded by the EC as project No. ICT-2009-4-249119 (MEYS No. 7E11040).

[6] The only mild exception are “verbs of saying” (*say, announce, add, report, ...*), relatively frequent in the news texts that the PUD corpora represent, since they are easily recognizable by their syntactic structure (usually `nsubj` and `ccomp` only).

References

Alishahi, A. and Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, 25(1):50–93.

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.

Bouma, G., Hajič, J., Nivre, J., Solberg, P., and Ovrelid, L. (2018). Expletives in universal dependency treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26, Bruxelles, Belgium. Association for Computational Linguistics.

Bowerman, M. (2002). Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? In *Mouton Classics: From Syntax to Cognition, from Phonology to Text, Vol. 2. (Original in Linguistics, 1990, 28, 1253-1289.)*. Mouton de Gruyter.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

Čech, R., Mačutek, J., and Koščová, M. (2015). On the relation between verb full valency and synonymy. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 68–73. Uppsala University, Uppsala, Sweden.

Cinková, S. (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.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edition*. Wiley.

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

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.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Nivre, Joakim//Hinrichs, E., editor, *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57—68, Vaxjo, Sweden. Vaxjo University Press.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., and Snedeker, J. (2010). Linking meaning to language: Linguistic universals and variation. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, pages 1186–1191, Austin, TX. Cognitive Science Society, Department of Linguistics and Scandinavian Studies, University of Oslo.

Kettnerová, V., Lopatková, M., and Bejček, E. (2012). Mapping semantic information from FrameNet onto VALLEX. *The Prague Bulletin of Mathematical Linguistics*, 97:23–41.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In *Proceedings of the LREC*, Canary Islands, Spain.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago and London.

Levin, B. and Hovav, M. R. (2005). *Argument realization*. Cambridge Univ. Press, Cambridge.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bellato, S., Bengoetxea, K., Bhat, R. A., Biagetti, E., Bick, E., Blokland, R., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kahane, S., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Peng, S., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., Roșca, V., Rudina, O., Sadde, S., Saleh, S., Samardžić, T., Samson, S., Sanguinetti, M., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Takahashi, Y., Tanaka, T., Tellier, I., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Vincze, V., Wallin, L., Washington, J. N., Williams, S., Wirén, M., Woldemariam, T., Wong, T.-s., Yan, C., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu,

H. (2018). Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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

Palmer, M., Gung, J., Bonial, C., Choi, J., Hargraves, O., Palmer, D., and Stowe, K. (2017). The pitfalls of shortcuts: Tales from the word sense tagging trenches. *To appear in: Lexical Semantics and Computational Lexicography*.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

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. European Language Resources Association (ELRA).

Urešová, Z., Fučíková, E., Hajičová, E., and Hajič, J. (2018b). CzEngClass 0.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, `http://hdl.handle.net/11234/1-2824`, Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

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

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

Wu, Y. (2017). The interfaces of Chinese syntax with semantics and pragmatics (Routledge Studies in Chinese Linguistics). *Journal of Linguistics*, 54(1):222–227.

Zeman, D. (2017). Core arguments in universal dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296. Linköping University Electronic Press.