ZDEŇKA UREŠOVÁ
Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic
uresova@ufal.mff.cuni.cz
+420 951 554 364

A CROSS-LINGUAL SYNONYM CLASSES LEXICON

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

Keywords: valency, syntax, semantics, lexicon, bilingual corpus

Klíčová slova: valence, syntax, sémantika, slovník, bilingvní korpus

1. Introduction

Lexical resources have a long tradition in computational linguistics, mainly as resources used in Natural Language Processing (NLP) to develop various tools for basic tasks and applications. Even though recent progress in machine learning for these NLP tasks and applications somewhat decreased the need for lower-level complexity lexicons (such as for morphology), at the more advanced levels, such as semantics and the syntax-semantic interface, there is still need to develop, extend and link together new and existing resources. Our cross-lingual synonym class lexicon aims to be an interconnected lexicon capturing semantic information about verbs for more languages. We started with Czech and English, building a synonym class lexicon called CzEngClass.

While there are both synonym lexicons such as WordNet (Miller, 1995; Fellbaum, 1998; Pala and Smrž, 2004) and verbal lexicons containing semantic information, such as VerbNet (Kipper at al., 2000; Kipper et al. 2008) and FrameNet (Baker et al., 1998, Fillmore et al., 2003), our approach is different since it focuses on

the valency behavior of verbs and on mapping the valency information to Semantic Roles (SRs) comparable to the roles used in FrameNet (and in some cases, in VerbNet). At the same time, our goal is to make the resource linked to other lexicons – in particular those mentioned above – to be able to do comparative studies as well as provide an electronic resource for various Natural Language Processing experiments.

Last but not least, we are aiming at a multilingual resource, somewhat in the lines of the Multilingual FrameNet (Boas, 2005) project, but more tightly integrated, with the synonym classes being multilingual from the start. We have started with only two languages, Czech and English, since we have the essential resources for them ready, but it is our goal to enable easy extension to other languages which have at least some basic lexical and corpus resources.

The article is organized as follows: it starts in Section 2 with the description of the valency theory in the Functional Generative Description (Sgall et al., 1986). Section 3 describes the resources used as evidence for creating the synonymous classes in CzEngClass. The use of valency in our definition of synonym classes is presented in Section 4. The structure of the CzEngClass lexicon is described in Section 5. Finally, we conclude with some observations from the work done so far (Section 6).

2. Valency in the Functional Generative Description

Functional Generative Description (FGD, Sgall et al., 1986) is a formal description of "language meaning" based on dependency-based syntactic framework. The representation of a sentence in the multilayered approach of the FGD is called "tectogrammatical representation" of a sentence, and it can be described also as "deep" dependency syntactic representation, where the word "deep" means that elements of semantics are present in the formal representation, such as semantic relations. Integral part of the description of syntactic and semantic structure of a sentence is the description of basic units of such a structure (content words, represented as "tectogrammatical lemmas"), some of which display predicate behavior (mostly verbs). This behavior of verbs (and other parts of speech with predicative meaning) is captured in FGD's Valency Theory. In practice, all verbs are captured in a valency lexicon, in which each entry (roughly) corresponds to one sense of the headword verb. Each entry

also contains a list of arguments associated with that sense, together with additional information such as obligatoriness of a valency "slot" and its morphosyntactic realization (morphological and/or prepositional case, lexical realization for idiomatic expressions and light verb constructions, particles used etc.). All this information is called a "valency frame".

Arguments are further classified. The core ones ale called "inner participants", and the FGD recognizes five of them:

- ACT for Actor, typically the deep subject (John.ACT slept.)
- PAT for Patient, typically the deep object (Mary.ACT wrote a book.PAT.)
- ADDR for Addresse (John.ACT gave the phone.PAT to Mary.ADDR.)
- EFF for Effect (Mary.ACT said [that John is leaving tomorrow.].EFF
- ORIG for "Originated from" (Mary.ACT made a chair.PAT from wood.ORIG.)

For practical reasons, there are two more relations that have a standard head-to-dependent relation in the tectogrammatical representation, even though they could be often considered one "concept node" in the structure, namely

- DPHR (Dependent PHRaseme) for idioms (kick the bucket.DPHR), and
- CPHR (Compound PHRaseme) for (mostly) light verb constructions (make an adjustment.CPHR).

All other relations between a verb and its dependents are called "free modification", and there is a list of 38 of them. These can become an argument if they are semantically obligatory for a given verb sense, and therefore will also be included in its valency frame in the lexicon. The free modifiers used with verbs as a head of the dependency relation are listed below (from Urešová, 2011) in Tables 1 to 6, organized into major groups.

Relation Label	Relation full name / definition	Example
TFHL	Temp: For How Long	She left for a week.TFHL
TFRWH	Temp: FRom WHen	It comes from the last year.TRFWH
THL	Temp: How Long	She studied for long.THL

THO	Temp: How Often	They were meeting daily.THO
TOWH	Temp: TO WHen	Xmas eve fell on Monday.TOWH
TPAR	Temp: PARallel	He left during the lecture.TPAR
TSIN	Temp: SINce when	He is working since summer.TSIN
TTILL	Temp: TILL when	I will wait till spring. TTILL
TWHEN	Temp: WHEN	I will come tomorrow.TWHEN

Table 1. Time relations

Relation Label	Relation full name / definition	Example
LOC	LOCative (where?)	She appeared at home.LOC
DIR1	DIRectional: from where	She came from Prague.DIR1
DIR2	DIRectional: which way, through where	He walks through the forest.DIR2
DIR3	DIRectional: to where	Go to the parking lot.DIR3

Table 2. Spatial relations

Relation Label	Relation full name / definition	Example
ACMP	ACcoMPaniment	He left with his son.ACMP
CPR	ComPaRison	He is taller than his friend.CPR
CRIT	CRITerion	Acording to statistics.CRIT, ...
DIFF	DIFFerence	Temperature rose by 2 degrees.DIFF
EXT	EXTent	It limits us only partially.EXT
MANN	MANN (also fallback)	She acted spontaneously.MANN
MEANS	MEANS	Attach it by the rope.MEANS
REG	REGard	According to him.REG, she left.
RESL	RESuLt	She suburned brown.RESL
RESTR	RESTRiction (exception)	Except for Ike.RESTR, no one knew.

Table 3. Manner-type relations

Relation Label	Relation full name / definition	Example
AIM	AIM	He studies to excel.AIM
CAUS	CAUSe (reason)	He died of hunger.CAUS
CNCS	CoNCeSsion	Despite all efforts.CNCS they lost.

COND	CONDition	Call if you do not succeed.COND
INTT	INTenT	Cows are herding to return.INTT

Table 4. Causal and similar relations

Relation Label	Relation full name / definition	Example
ATT	ATTitude	It is unfortunately.ATT too late.
INTF	INTensiFication	It.INTF was Charles who plays.
MOD	MODality	Apparently.MOD he is not behaving.
RHEM	RHEMatizer	It is already.RHEM known that …
PREC	PRECeding text reference	But.PREC chances are nevertheless.PREC minimal.

Table 5. Rhematizers, referential, modal and inter-sentential relations

Relation Label	Relation full name / definition	Example
BEN	BENefactor	He worked for children.BEN
CONTRD	CONTRaDiction	Her daughter joined, while.CONTRD her son declined.
HER	HERitage	Jean was named after her aunt.HER
SUBS	SUBStitution	He arranged it instead of him.SUBS
COMPL	COMPLement (attributive predication)	Mark was leaving exhausted.COMPL

Table 6. Other relations

Each valency frame contains the verb headword and the labels for all arguments, i.e., all obligatory slots, and also for optional inner participants, if any. For each slot, a list of possible morphosyntactic realization of that particular argument on the surface (in the resulting readable text) is included. Usually some comments, synonyms and/or examples are also included, to facilitate the use of the valency frame e.g., in manual annotation (to distinguish between individual senses of the verb, which might not be clear from the argument list alone). A few examples of Czech valency frames (slightly simplified) follow.[1]

[1] Morphosyntactic properties of the argument are shown in parenthesis following the argument slot label, where morphological cases are coded as numbers, i.e., 1 for nominative, 2 for genitive, etc.; prepositional cases use the preposition itself + case number. Alternative morphosyntactic relations are

(1) spravit (‚repair‘) ACT(1) PAT(4) ex.: spravit kolo (‚to repair a bicycle‘)

(2) dát[5] (‘give’) ACT(1) PAT(4) ADDR(3) ex.: dali jim dárky (‘they gave them presents’

(3) probudit_se[1] (‘wake up’) ACT(1) ?PAT(z+2) ex.: Tomáš se probudil ze sna (‘Tom woke up from a dream’)

(4) ucítit[3] (‚feel‘) ACT(1) CPHR({nutkání,potřeba,…}.4) ex.: ucítil potřebu odejít (‚he felt the urge to leave’)

(5) udat[5] ACT(1) DPHR(tón.S4) ?PAT(3) ex.: To může udat tón obchodování na burze. (‘It could set the tone for trading on the stock market.’)

(6) uslyšet[3] (‚hear‘) ACT(1) PAT(4;že;jak-2;jestli;zda;.c) ex.: uslyšel, co si povídají (‚he heard what they are talking about‘)

For brevity, we are leaving out certain other considerations for creating valency frames, such as shifting of argument labels (in short, the first two arguments are always an ACT and PAT, even though semantically, the PAT could be e.g., more of an addressee semantically; but when no other deep object is present, it will be „shifted“ to PAT). For more details, please see (Urešová, 2011) or for the theoretical background see (Panevová, 1974; Panevová, 1975).

3. Lexical and Corpus Resources

As mentioned in the Introduction section, the goal is not only to create a new resource, but to explicitly link it to other resources: primarily to the valency lexicons that have been created previously and that provide all of the valency information that we use, and additionally to existing English and Czech resources, since such interconnections multiply the usefulness of such a lexical resource. In addition, we use a richly annotated bilingual parallel corpus for usage evidence.

For the core valency information, we use the PDT-Vallex[2] lexicon for Czech (Hajič et al., 2003, Urešová, 2011) and the EngVallex[3] lexicon (Cinková, 2006) for English; previous work has resulted in a Czech-English parallel valency lexicon called

separated by a semicolon. Upper index at the verb headword signifies sense number. A question mark preceding the argument slot label denotes and optional argument (inner participant).

[2] http://lindat.mff.cuni.cz/services/PDT-Vallex, http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

[3] http://lindat.mff.cuni.cz/services/EngVallex, http://hdl.handle.net/11858/00-097C-0000-0023-4337-2

CzEngVallex (Urešová et al., 2015; Urešová et al., 2016), which we also draw on substantially in the work presented here.

The external lexicons used are FrameNet[4] (Fillmore, 1976; Baker et al., 1998), VerbNet[5] (Kipper at al., 2000; Kipper et al. 2008; Levin, 1993), PropBank[6] (Palmer et al., 2005), VALLEX[7] (Lopatková et al., 2008; Lopatková et al., 2017), English WordNet[8] (Fellbaum, 1998) and Czech Wordnet[9] (Pala and Smrž, 2004; Pala et al., 2011).

The Prague Czech-English Dependency Treebank[10] (Hajič et al., 2011; Hajič et al., 2012), which is already interlinked with all the three valency lexicons (PDT-Vallex, EngVallex and CzEngVallex), serves as the main textual evidence and for examples of annotation.

4. Synonymy Definition

Defining synonyms is surprisingly intuitive endeavor. When creating the currently most popular synonym resource, the WordNet (Miller, 1995) dictionary, G. Miller in fact was interested in semantic relation and no proper definition of synonymy was adopted: "Miller wondered whether a semantic network could in fact be built for the bulk of the English lexicon. In the mid-80s, he recruited a group (…) and, *without much further instruction*, asked them to cluster nouns, verbs, and adjectives into "synsets" that could be interrelated with a handful of semantic relations. Relying on conventional lexical resources and *intuition*, the WordNet team created tens of thousands of entries (…)" (Fellbaum, 2013; emphasis by author).

Other definitions of synonymy are rather theoretical, both for intra-language and inter-language cases (Lyons, 1968, Cruse, 1986). We side with the general opinion that absolute synonymy essentially does not exist, and that context must be taken into account (Zeng, 2007; Palmer, 1981). For verbs, we consider valency the most

[4] https://framenet.icsi.berkeley.edu. Used also for inspiration with the selection of semantic roles.
[5] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[6] https://verbs.colorado.edu/~mpalmer/projects/ace.html
[7] Czech Valency lexicon based also on the FGD theory.
[8] http://wordnetweb.princeton.edu/perl/webwn
[9] http://hdl.handle.net/11858/00-097C-0000-0001-4880-3
[10] https://ufal.mff.cuni.cz/pcedt2.0, https://catalog.ldc.upenn.edu/LDC2004T25

appropriate "context" of their use, and therefore define verb synonymy using valency and its relation to semantic roles. Similarly to other approaches (including WordNet), we use word (verb) sense as the unit for which synonymy can be defined (as opposed to the word (lemma) itself, which can have several senses with very different meanings).

For assigning verb senses to the appropriate verb classes (which correspond to synsets in the WordNet terminology) containing synonymous verb senses, the following definition of context (Urešová, 2018a) is used to determine synonymy: "*context* is defined as the set of SRs that the given verb, as a member of a bilingual synonym class, expresses by its arguments and/or adjuncts, or which are implicitly present, possibly with additional structural or semantic restrictions. Each class has an associated, single (common) set of SRs while such a set is shared by all its members, even if each SR can be expressed (mapped to) by a different argument (or by an adjunct, or implicitly or explicitly in the verb's dependent substructure) for different verbs as members of that class. Conversely, such mapping must exist at least for all obligatory valency slots as defined in the two corresponding valency lexicons."

Clearly, valency (and the class-wide set of Semantic Roles common for any synonym class, i.e., its Roleset) is the most important phenomenon that governs the structure of the classes in the CzEngClass lexicon. While the Semantic Roles are unique for each synonym class, the only requirement for the valency frame of each member of the synonym class is that its slots are mappable to the Semantic Roles in the assigned Roleset for the class in question, but not necessarily identical. In other words, not only the surface form of arguments but also the valency slots can differ for a verb (sense) in a class, and the two verbs can still be considered synonymous (provided a mapping of their valency arguments to the class-wide set of Semantic Roles can be established). However, it is not strictly necessary for all the valency slots be present in the valency frame, since if an adjunct (free modification in the FGD valency terminology) can be uniquely identified to map to a particular Semantic Role from the given Roleset for the class in question, such a verb can also be a member of that class.

For example, the verbs *defend*, *prevent* and *insulate* are members of one synonym class (provisionally called "*defend*"), despite having the following, different valency frames:

(7) defend ACT() PAT() EFF(); ex.: she.ACT defended Jane.PAT against the gang.EFF

(8) prevent ACT() PAT() ADDR(); ex.: he.ACT prevented Tom.ADDR from accepting.PAT that offer

(9) insulate ACT() PAT() ORIG(); ex.: John.ACT insulated his business.PAT from currency fluctuations.ORIG

In this case, it has been determined (by manual inspection of many occurrences of potential synonyms in a corpus, in our case a parallel Czech and English corpus) that the Roleset for this class contains the following Semantic Roles: Agent (the person, thing, event or action that is defending [someone/something against someone/something]), Asset_Patient (the thing or person who is being defended) and Harmful_situation (the danger, event, thing or person that the Asset_Patient is being defended against). The mapping of the valency slots to these Semantic Roles is as shown in Table 7.

Verb	Agent	Asset_Patient	Harmful_situation
defend	ACT	PAT	EFF
prevent	ACT	ADDR	PAT
insulate	ACT	PAT	ORIG

Table 7. Mapping valency slots to SRs (class '*defend*', English verbs)

Since we aim at a multilingual lexicon, i.e., having verbs form multiple languages in one synonym class, we have to determine consistent application of the valency mapping principle across languages. As could be expected, the Roleset assigned to a synonym class is the key: a verb from a different language is included in the class if its arguments map to this Roleset. Since for the time being we work with Czech and English only, the situation is relatively simple thanks to the existence of valency dictionaries built on the same principles for both languages. In the previous example of the class '*defend*', we can also add the following Czech verbs, among others:

(10) *bránit* (‘defend’) ACT() PAT() EFF(); ex.: *bránit někoho*.PAT *před něčím*.EFF (‘defend someone.PAT against something.EFF’)

(11) *zaclonit* (‘shield’) ACT() PAT() ADDR(); ex.: *zaclonit detektory*.PAT *před zářením*.ADDR (‘shield the detectors.PAT from rays.ADDR’)

with the valency-to-SR mapping shown in Table 8.

Verb	Agent	Asset_Patient	Harmful_situation
bránit	ACT	PAT	EFF
zaclonit	ACT	PAT	ADDR

Table 8. Mapping valency slots to SRs (class ‘*defend*’, Czech verbs)

In certain cases, a „general argument“ can be mapped to a given Semantic Role, if no specific argument can be found in the valency slot of otherwise intuitively clearly synonymous verb. For example, for the aforementioned class ‘*defend*’, we also include the verb ‘*conserve*’, even though it has only two valency slots in the English valency dictionary:

(12) conserve ACT() PAT(); ex.: owners.ACT conserve energy.PAT

In this case, the “Harmful_situation” is not present in the valency frame, since the verb ‘conserve (energy)’ implicitly points to ‘wasting (energy)’ as the situation against which the ‘defense’ is aimed.

If a slot cannot be filled by an argument, but it is clear which adjunct (free modification, such as MANNer, MEANS etc.) would map to the Semantic Role with a missing valency slot counterpart, then such an adjunct can be mapped to. For example, in the synonym class ‚*view*‘ (in the sense ‚something in a certain way‘), which has been associated with three SRs (Perceiver, Item, Manner), the last role (Manner), which is sometimes mapped to EFF (e.g., for the English verb ‚view‘ itself), can only be mapped to MANNer for other verbs, such as the Czech verb ‚nahlížet’ (‘view’), which contains only two arguments, ACT and PAT, mapped to Perceiver and Item, respectively.

The mapping scheme between valency slots and arguments also allows to bring together alternations (Kettnerová, 2014) to one class. For example, the following two

valency frames for the verb 'load,' which - based on the valency principles in FGD – constitute two different entries in the valency lexicon, namely

(13) load ACT() PAT() ?MEANS(); ex.: load the truck.PAT with hay.MEANS

(14) load ACT() PAT() DIR3(); ex.: load the bags.PAT onto the truck.DIR3

since their valency structure differs, but they can be clearly mapped to a set of the following Semantic Roles associated with the synonym class for verbs like load, fill, put (in), etc.: Agent (~ Mover), Entity_Moved and Container.

5. Creating the CzEngClass Lexicon

The CzEngClass lexicon contains all the Czech and English bilingual synonym classes, grouped manually based on the corpus and lexical resources listed in Section 3. The manual process is helped by automatic preprocessing enabled by the existing PCEDT corpus, which is linked to the valency lexicons associated with it – PDT-Vallex for Czech, EngVallex for English and CzEngVallex for aligned verb senses as found in the PCEDT. The preprocessing started with a selection of 200 Czech verbs (their senses as represented by their valency frames) of various frequencies in the corpus to represent a broad spectrum of verbs. Through the CzEngVallex lexicon, all aligned and previously manually checked English verb sense counterparts from EngVallex have been preselected. These have been pruned manually, and a reverse extraction took place, this time for more Czech verbs aligned in CzEngVallex with all the English verbs as pruned in the first step. Needless to say, there still was a lot of noise, especially for the verbs with higher frequency, and another pruning had to be run (with multiple annotators trained to do so).

For 60 of these classes, Semantic Roles have been associated with the synonym class, and valency slots for each member of that class have been mapped to these SRs as described in Section 4. In addition, for these 60 classes, all the class members have been linked individually to appropriate entries in all the external resources (lexicons): VALLEX for Czech[11] and FrameNet, OntoNotes senses (groupings), PropBank,

[11] Czech WordNet will be added later.

VerbNet and WordNet for English (cf. Section 3 for details and references for these external resources).

As expected, interannotator agreement is not very high for such a complex resource. We have measured pairwise agreement using Cohen's kappa (κ) formula, which was between 0.19 and 0.68 for eight annotators; the value of the multiannotator-based Fleiss' kappa was 0.45 for the task of determining whether a particular preselected class member should stay in the class or not (Urešová et al., 2018a). However, when measuring a deviation from a simple consensus (averaged over all annotators who annotated a particular word), this deviation came to 0.09 when normalized to decisions simplified to 0 and 1.

The current version of the resource,[12] including the multiannotator annotation database for comparing annotators and their approach, has been released in the LINDAT/CLARIN repository.[13] In addition, we have developed an editor for manual pruning of the synonym classes, associating the Semantic Roles with classes, mapping valency to them and for inserting the external links to every class member. This editor is called SynEd (for more details, see Urešová et al., 2018b).

6. Conclusions and Future Work

While building the CzEngClass lexicon is still work in progress, the first version released gives us confidence we can finish the lexicon with a reasonable coverage for both Czech and English. That way, we will demonstrate that it is possible to build such a resource with tight multilinguality built in, and that there will be no principled obstacles to adding more languages in the future. Any manually created resource is necessarily subjective, but we believe that by using valency-to-SRs mapping as the "contextual" criterion helps to keep consistency among the resulting synonym classes,[14] even though intuition about verb senses still plays an important role.

[12] The lexicon itself is a single XML file with external references to the lexical resources used.
[13] http://hdl.handle.net/11234/1-2824 available under CC BY-NC-SA 4.0.
[14] For detailed analysis of interannotator agreement, see (Urešová et al, 2018a).

In the nearest future, we plan to extend the full mapping of valency slots to Semantic Roles for the remaining 140 classes, and add the external links to all of the class members, and publish the first version, which will cover about 60% of running texts in Czech and English. As a next step, we plan experiments with automatic preprocessing and addition of other class members to the 200 existing classes, as well as identifying additional classes and adding them to the CzEngClass lexicon.

7. Acknowledgements

This project 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 as projects No. LM2015071 and CZ.02.1.01/0.0/0.0/16_013/0001781.

8. Abbreviations

FGD - Functional Generative Description
NLP - Natural Language Processing
SR - Semantic Role
SynEd – Synonym Editor
PCEDT - Prague Czech-English Dependency Treebank

9. References

Baker, C. F., Fillmore, C. J., Lowe, J. B. (1998). *The Berkeley FrameNet project*. In: Proceedings of the COLING-ACL, Montreal, Canada. 86-90.

Boas, H. C. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography*. 18(4). 445-478. DOI: 10.1093/ijl/eci043

Cinková, S. (2006). *From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description.* In: Proceedings LREC 2006, Genova, 2170-2175.

Cruse, D. A. 1986. *Lexical Semantics.* Cambridge University Press, UK.

Fellbaum, Ch. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fellbaum, Ch. (2013). George A. Miller – Obituary. *Computational Linguistics* 39(1). 1-3. DOI: 10.1162/COLI_a_00131

Fillmore, Ch. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280: 20–32. DOI: 10.1111/j.1749-6632.1976.tb25467.x

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

Fučíková, E., Hajič, J., Šindlerová, J., Urešová, Z. (2015). *Czech-English Bilingual Valency Lexicon Online*. In: Proceedings of the 14th TLT 2015, IPIPAN, Warszawa, PL, 61-71.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z. (2011). *Prague Czech-English Dependency Treebank 2.0*. UFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/pcedt2.0 (23.3.2015)

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z. (2012). *Announcing Prague Czech-English Dependency Treebank 2.0*. In Proceedings of the European Association for Computational Linguistics Conference (EACL). Istanbul, Turkey. 3153-3160

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., Pajas, P. (2003). *PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation.* In Proceedings of the 2nd TLT conference, Vaxjo, Sweden. Vol. 9. 57-68.

Kettnerová, V. (2014). *Lexikálně-sémantické konverze ve valenčním slovníku*. Karolinum, Prague, Czech Republic, 280 p.

Kipper, K., Hoa Trang Dang, Palmer, M. (2000). *Class-Based Construction of a Verb Lexicon.* AAAI-2000Seventeenth National Conference on Artificial Intelligence, Austin, TX, July 30 - August 3.

Kipper, K., Korhonen, A., Ryant, N., Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal* 42(1).21–40.

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

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., Žabokrtský, Z. (2017). *Valenční slovník českých sloves VALLEX*. Karolinum. Praha. 698 p.

Lopatková, M., Žabokrtský, Z., Kettnerová, V., Skwarska, K., Bejček, E., Hrstková, K., Nová, M., Tichý, M. (2008). *Valenční slovník českých sloves*. Karolinum. 381p.

Lyons, J. 1968. *Introduction to Theoretical Linguistics*. Cambridge University Press.

Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Pala, K., Čapek, T., Zajíčková, B., Bartůšková, D., Kulková, K., Hoffmannová, P., Bejček, E., Straňák, P., Hajič, J. (2011). *Czech WordNet 1.9 PDT*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.

Pala, K., Smrž, P. (2004). Building Czech Wordnet. *Romanian Journal of Information Science and Technology*. Budapest, 79-88.

Palmer, F. R. 1981. *Semantics.* 2nd ed. Cambridge University Press, UK.

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

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

Panevová, J. (1975). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 23. 17–52.

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

Urešová, Z. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics 9, UK Praha, 375

Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., Šindlerová, J. (2015). *Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus.* In: Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015), ACL, Stroudsburg, PA, USA, 124-128.

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

Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J. (2018a). *Synonymy in Bilingual Context: The CzEngClass Lexicon*. In Proceedings of Coling 2018. Santa Fe, New Mexico, USA.

Urešová, Z., Fučíková, E., Hajičová, E., Hajič, J. (2018b). *Tools for Building an Interlinked Multilingual Synonym Lexicon Network*. In Proceedings of LREC 2018, May 7-12, 2018. ELRA. Miyazaki, Japan.

Zeng, X. 2007. Semantic relationships between contextual synonyms. *US-China Education Review*, 4(9). 33–37.

Summary:

In the present article, we focus on valency and synonymy of verbs in a bilingual, Czech-English setting. Our research of semantic equivalence of verbs is based on the FGD theory on the syntactic side (including valency), and gets main inspiration from FrameNet and VerbNet on the semantic side. As the main source of evidence, we use the Prague Czech-English Dependency Treebank 2.0. We consider this “bottom-up”

approach a novel and appropriate approach to study verbal synonymy. Synonymous Czech and English verbs are being grouped into cross-lingual synonym classes and captured in the new CzEngClass lexicon. This lexicon contains not only mappings of valency arguments to semantic roles for each member of the synonym group, but also links them to individual verb entries in FrameNet, VerbNet, Vallex(es) and Czech and English WordNets, making CzEngClass also a richly interconnected lexicon.

V článku se zaměřujeme na valency a synonymii sloves v bilingvním česko-anglickém kontextu. Náš výzkum sémantické ekvivalence sloves je z pohledu syntaktického (včetně valence) založen na FGP a z pohledu sémantického je inspirován především FrameNetem a VerbNetem. Jako hlavní zdroj korpusových dokladů používáme Pražský česko-anglický závislostní korpus 2.0. Postup výzkumu "od spoda nahoru" považujeme za nový a adekvátní přístup pro stadium slovesné synonymie. Synonymní česká a anglická slovesa jsou uskupeny do mezijazykových synonymních tříd v novém slovníku CzEngClass. Tento slovník pro každé synonymní sloveso dané třídy obsahuje jak mapování valenčních argumentů na sémantické role, tak propojení s jednotlivými slovesnými významy ve FrameNetu, VerbNetu, ve Vallexech a v anglickém WordNetu, což ze slovníku CzEngClass zároveň vytváří bohatě propojený slovník.

Submitted on July 31, 2018