CZENGCLASS – TOWARDS A LEXICON OF VERB SYNONYMS WITH VALENCY LINKED TO SEMANTIC ROLES

ZDEŇKA UREŠOVÁ – EVA FUČÍKOVÁ – EVA HAJIČOVÁ
Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic

UREŠOVÁ, Zdeňka – FUČÍKOVÁ, Eva – HAJIČOVÁ, Eva: CzEngClass – Towards a Lexicon of Verb Synonyms with Valency Linked to Semantic Roles. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 364 – 371.

Abstract: In this paper, we introduce our ongoing project about synonymy in bilingual context. This project aims at exploring semantic 'equivalence' of verb senses of generally different verbal lexemes in a bilingual (Czech-English) setting. Specifically, it focuses on their valency behavior within such equivalence groups. We believe that using bilingual context (translation) as an important factor in the delimitation of classes of synonymous lexical units (verbs, in our case) may help to specify the verb senses, also with regard to the (semantic) roles relation to other verb senses and roles of their arguments more precisely than when using monolingual corpora. In our project, we work "bottom-up", i.e., from an evidence as recorded in our corpora and not "top-down", from a predefined set of semantic classes.

Keywords: lexical resources, valency, synonymy, semantic roles, dependency corpus, multilingual

1 INTRODUCTION

It is widely accepted that verbs play a crucial role in a sentence structure – they form its core, relate other elements of the sentence to each other. Verbs can describe many events and states depending on the collocates they appear with, which in turn leads to the problem of ambiguity of verbs related to their meanings (senses). In addition, the same verb with no obvious meaning ambiguity can get translated into two or more different verbs in the target language, yet forming a perfect translation conveying the same meaning as in the source language. Take the verb "widen" in English, seen 32 times in the Penn Treebank [21] – in its Czech translation, 14 different verbs have been found: not only the most direct translation "rozšířit", but also "prohloubit" (lit. "deepen"), "rozrůst se" (lit. "grow [oneself]"), "stoupnout" (lit. "rise"), "zvětšit se" (lit. "enlarge"), "zvyšovat" (lit. "raise," "get higher") etc. Immediately, questions arise primarily about synonymy, but also about concrete vs. abstract distinction, relation to valency and argument structure, and more.

Different meanings of the same verb, or verb senses, are recorded and described – usually rather implicitly and informally – in both monolingual and bilingual dictionaries and we as humans can understand the sense distinctions well. However, our aim should be to describe verb senses precisely and explicitly. How do we know what is the explicit set of senses for any particular verb? Which senses (of different verb lexemes) are synonymous or near synonymous [7], [31] in the broader context

DOI 10.1515/jazcas-2017-0046

of use? It has been shown that if we let different people determine this, even on the same set of examples (i.e., using the same corpus), they inevitably come up with a different set. More precisely, the inter-annotator agreement [1] will be low, regardless of the level of linguistic expertise the annotators might have. Some researchers even go so far as to declare that they "do not believe in word senses" (legendary quote by the lexicographer Sue Atkins [2], explained by an article with the same title by Kilgarriff [16] that it should be interpreted as not believing in pre-determined, fixed set of word senses). Others try to find a sweet spot between a hard-to-agree-on, fine-grained set, represented e.g., by WordNet [6], and a coarse(r)-grained set, which does not provide enough detail–such as VerbNet lexicon [24], [14]. FrameNet [3], [8], [9] an English lexical resource which adds roles and uses semantic frames to group verbs and provide examples of use (based on attested corpus examples) is another well-known resource.

Regarding other languages, only some non-English WordNets link to the original English WordNet "synsets". FrameNet covers several languages, but it is not created systematically from parallel corpora. VerbNet is English-only. Moreover, these lexicons do not contain detailed morphosyntactic description of verb argument behavior (perhaps due to the selection of the original languages, which are in general not inflectional). There are no bilingual (or multilingual) resources describing verbs and their senses together with their semantic and morphosyntactic behavior in a bilingual setting. To fill this gap, our project will focus primarily on synonymy in bilingual context.

We believe that using the existing resources (mostly bilingual) based on the Functional Generative Description theory (FGD; [29]) will help us proceed in that direction. We are using two manually treebanked corpora: PDT (`http://ufal.mff.cuni.cz/pdt2.0`) and PCEDT [11], and the valency lexicons linked to these treebanks: PDT-Vallex [32], EngVallex [5], and a parallel valency lexicon CzEngVallex [33], [34]. We also take advantage of another FGD-based lexicon VALLEX [20], [19], [15] and other available resources, such as VerbNet [24], [14], FrameNet [3], English [6] and Czech WordNet [22], [23].

2 RESEARCH QUESTIONS

We will take advantage of the aforementioned lexical key resources as well as of large monolingual corpora, other parallel corpora such as Intercorp [28] and the NLP tools available for both languages, to work towards answering the following research questions:

– Do (verb) classes of synonyms based on monolingual and bilingual contexts differ, and if yes, in which respects? How are they related to structural representations (FGD, Abstract Meaning Representation (AMR, [4])?

– Crucially though, can the classes based on bilingual context be still kept disjoint (as the synsets in WordNet are)? Which consequences would overlapping classes have on the underlying theoretical approach(es)? Should any of the verb senses, as defined in the available dictionaries previously, be split or merged, based on the bilingual usage evidence?

– Which properties of a verb sense and the corresponding valency frame are relevant for grouping such verb senses into classes of synonyms, again in a bilingual vs. a monolingual context? Are they supported by corpus evidence?

– Conversely, what have the verb senses grouped in one class in common in terms of valency?

– Can a common set of verb "roles" (inspired, e.g., by FrameNet's Frame Elements [3] and by VerbNet's Thematic Roles [24]) be associated with one class, and how are these roles expressed in terms of valency (arguments, morphosyntactic expression)?

Our ultimate goal is in fact even a step deeper than to look at these questions in isolation: we hope to use the answers to these questions to confirm our hypothesis that using translation (i.e., bi-/multilingual context) as an important factor in determining the composition of such verb classes helps to define verb senses and their (primarily equivalent) relation to other verb senses and roles of their arguments more precisely than when using monolingual corpora, even if they follow Kilgarriff's postulate of giving substantial weight to individual occurrences in a corpus. We will create a lexicon of such synonymous verb pairs around representatively selected "seed" verbs; such similarity will be tested primarily against the translational equivalents in context, as found in the parallel corpora. We will compare the results with the approach of [18] as embodied, e.g., in the VerbNet [24], [14], as a representative of classes of semantically and syntactically similar verbs based on monolingual resources and research on one language (English). Last, but certainly not least, we will compare the resulting classes and their properties to the VALLEX lexicon [20], [19] and VerbaLex [13] on the Czech side. Results will be analyzed from the point of view of the Functional Generative Description theory [29] and its approach to valency [26], [27], [12] and the relation of form and meaning, and possibly generalized across the two languages we will work with.

3 PROJECT WORKFLOW

3.1 Preparatory Part

In the preparatory part of the project, we have been analyzing the existing Prague Czech-English Dependency Treebank (PCEDT), as well as the related valency lexicons: PDT-Vallex [32], EngVallex [5] and CzEngVallex [34]. We have been also studying the methodology of the VerbNet class-based verbal lexicon [17] and FrameNet [3]. We have performed a detailed analysis of the verbs contained in the CzEngVallex lexicon, in order to create classes of synonyms similar to those of VerbNet, but–importantly–in a bilingual setting, which needs the support of the PCEDT to see the use of such verbs in the parallel corpus, i.e., in the context of real usage. Next, we have selected a representative sample of verbs (about 50 classes centered around "seeds" from the sample selected), along the dimensions of frequency and richness of sense inventory and translation equivalents. Simultaneously, we have been preparing technical tools (software) allowing to manually (re-)group and refine verb senses, build classes of synonyms and assign them an appropriate semantic frame and roles.

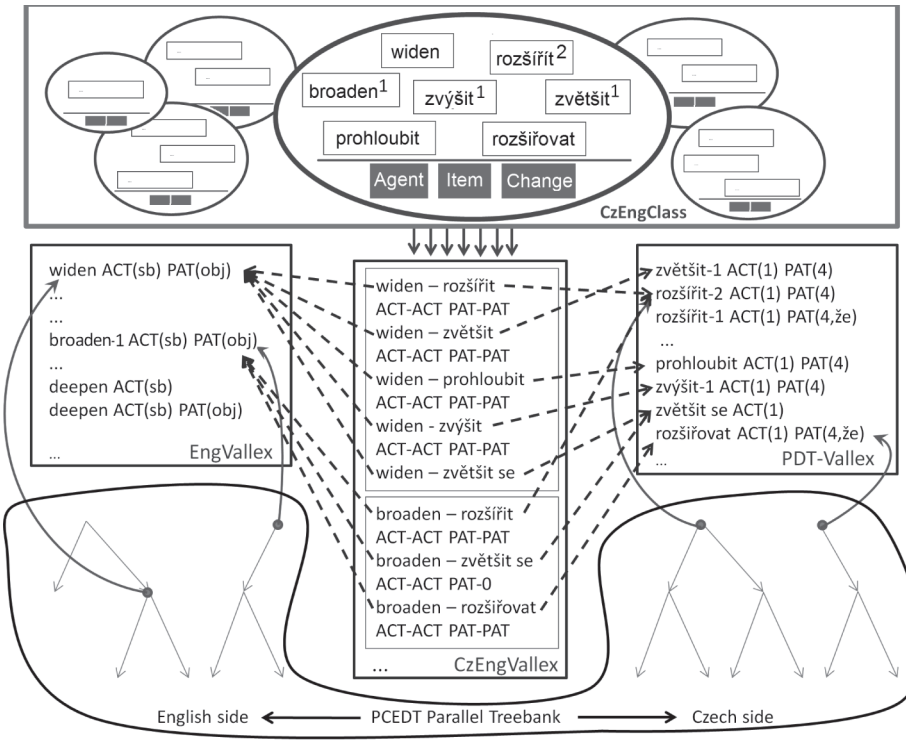

Fig. 1. Overall scheme of CzEngClass

3.2 Data Extraction

In the data extraction part, verb sense (~ verb valency frame) pairs selected from CzEngVallex have been grouped into classes corresponding to their semantic similarity (i.e., synonyms or near synonyms). While this is more complex in the bilingual setting, the translation context from PCEDT provided very strong, empirical evidence for their equivalence in context, as opposed to mere similarity in argument types or in their surface realization (and one's often unreliable intuition). The resulting "database" (Fig. 1, working name "CzEngClass") has a relatively simple form – it groups together frame pairs captured in CzEngVallex into classes, which represent synonym or near-synonym pairs of verbs (verb senses). However, every pair in every class is also linked to CzEngVallex, PDT-Vallex and EngVallex, and the PCEDT, allowing for relation-based search by computational tools in the analysis part.

Part ot the data extraction process will keep links to external resources as well. The following resources will be used: FrameNet, VerbNet, PropBank [25] and WordNet for English, and Czech WordNet for Czech. Most of these resources are accessible through the Unified Verb Index (`https://verbs.colorado.edu/verb-index/`). However, it will be necessary to find the right correspondences; for example, the senses as recorded in VerbNet "Groupings" have to be linked to EngVallex senses (frames), and of course the verb arguments, e.g, from PropBank are structured differently than in EngVallex.

3.3 Data Analysis

This part is the core part of the project. Here we plan to analyze the complex set of relations between meaning and form for the synonym classes of verb senses (as represented by their valency frames) created in the data extraction part. We will study the classes as a whole as well as its members individually in terms of arguments, their types, their surface morphosyntactic realization, and also all anomalies and deviations which we encounter either in the valency lexicons PDT-Vallex and EngVallex or in the PCEDT parallel data. We believe that such findings will lead to the description of bilingual-corpus-based semantically defined classes of synonyms or near synonyms. In this analysis, the external resources will also be consulted to get more insight into semantic role labeling, semantic classes etc.

4 PROJECT OUTPUT

The output of the project will be CzEngVallex, a lexicon of synonym classes, where each verb (verb sense), Czech and English alike, will be assigned to one class, and it will be linked to the other available resources for reference and to support other follow-up studies. In addition, each class will be also characterized by a set of semantic roles which will be shared about the class members, and verb arguments will be mapped to these roles. The data will be openly and freely available.

Verb lexemes	Closest FrameNet frame	Roles:			
		Cognizer	Means/ Instrument	Phenomenon	Source
dozvědět se[1]	Becoming_aware	ACT		PAT	ORIG
get[1]	Becoming_aware	ACT		PAT	ORIG
hear[1,2]	Becoming_aware	ACT		PAT	ORIG
know[1,3]	Becoming_aware	ACT		PAT/EFF	
learn[1]	Becoming_aware	ACT		PAT	ORIG
tell[3]	Becoming_aware	ADDR	ACT?	PAT/EFF	ACT?

Tab. 1. Example set for "learn" ("dozvědět se") with (initial) argument mappings

An example of preliminary synonym set with equally preliminary mappings from verb argument labels to a set of roles initially identified for each class are in Table 1. It is clear that there are immediate problems to solve:

- what (FGD-)based roles should be used to map the candidate verb arguments to the Means/Instrument and the Source semantic roles?
- why the translation uses the word "know" as an translation equivalent of "dozvědět se", given that "know" is more of a state-type of verb, while "dozvědět se" is describing the process of "getting to know", albeit it is in perfective voice?
- is "tell" really a good synonym (even in the loose, contextually-based sense), given that the ACTor could well be assigned to both the Source and the Means semantic roles?

We also expect that the existing valency lexicons will be amended, since inconsistencies in the previous annotation may be found. The corrected lexicons PDT-Vallex and EngVallex will thus be also published openly.

The overall structure of the lexicon with the basic referencing (from CzEngClass to the two valency lexicons and the parallel corpus, but not to the external resources) are depicted schematically on Fig. 1. So far, an XML scheme for the lexicon has already been designed and a work on an editor is in progress (cf. also Sect. 3.1.) and it will be described in the final version of the paper.

5 SUMMARY

We have described (based on the grant No. GA17-07313S proposal, of which the authors of this article are participants) a project which is just starting and which is supposed to lead to an interconnected synonym bilingual lexicon based on parallel corpus and existing lexicons. Entries in this lexicon will share, for each class, a set of semantic roles mapped to arguments in the valency lexicons. The lexemes will also be linked to the Universal Verb Index to keep relations to the widely used verb lexicon resources, such as FrameNet, VerbNet or PropBank, whenever possible. An indispensable resource, which is directing the research, is the Czech-English parallel richly annotated corpus which brings a new view on cross-lingual (and multilingual) contextual synonymy.

All the new resources and the linking will be made public as open data.

ACKNOWLEDGMENTS

This work has been supported by the grant No. GA17-07313S of the Grant Agency of the Czech Republic, and it uses resources hosted by the LINDAT/CLARIN Research Infrastructure, project No. LM2015071, supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

[2] Atkins, S. (1993). Tools for computer-aided corpus lexicography: the Hector project. In *Papers in Computational Lexicography: Complex'93*, pages 1–60, Budapest, Hungary.

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

[4] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th LAW Workshop*, pages 178–186, ACL, Sophia, Bulgaria.

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

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

[7] Filipec, J. (1961). *Česká synonyma z hlediska stylistiky a lexikologie.* Nakladatelství Československé akademie věd, Praha.

[8] Fillmore, Ch. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280:20–32. Accessible at: doi: 10.1111/j.1749-6632.1976.tb25467.x.
[9] Fillmore, Ch. J., Johnson, Ch., and Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.
[10] Fučíková, E., Hajič, J., Šindlerová, J., and Urešová, Z. (2015). Czech-English Bilingual Valency Lexicon Online. In *Proceedings of the 14th TLT 2015*, pages 61–71, IPIPAN, Warszawa, Poland.
[11] 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., and Žabokrtský, Z. (2011). Prague Czech-English Dependency Treebank 2.0. Data/software, UFAL MFF UK, Prague, Czech Republic. Accessible at: `http://ufal.mff.cuni.cz/pcedt2.0` (23. 3. 2015).
[12] Hajičová, E. (1983). Remarks on the meanings of cases. *Prague Studies in Mathematical Linguistics*, 8:149–157.
[13] Hlaváčková, D., Horák, A., and Kadlec, V. (2006). Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In *Proceedings of 9th International Conference on Text, Speech, and Dialogue (TSD 2006)*, pages 85–92, Springer, Berlin – Heidelberg, Germany.
[14] Kawahara, D., Peterson, D., Popescu, O., and Palmer, M. (2014). Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2014*, pages 58–67, Gothenburg, Sweden.
[15] Kettnerová, V. (2014). *Lexikálně-sémantické konverze ve valenčním slovníku*. Karolinum, Prague.
[16] Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.
[17] Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42(1):21–40.
[18] Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, Chicago.
[19] Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.
[20] Lopatková, M., Žabokrtský, Z., Kettnerová, V., Skwarska, K., Bejček, E., Hrstková, K., Nová, M., and Tichý, M. (2008). *Valenční slovník českých sloves*. Karolinum, Praha.
[21] Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building A Large Annotated Corpus of English: The Penn Treebank. *Computational Linguisics*, 19(2):313–330.
[22] Pala, K. and Smrž, P. (2004). Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7(2–3):79–88.
[23] Pala, K. and Všianský, J. (1994). *Slovník českých synonym*. 1. vyd. Nakladatelství Lidové Noviny, Praha.
[24] Palmer, M., Hwang, J. D., Brown, S. W., Kipper, S. K., and Lanfranchi, A. (2009). Leveraging lexical resources for the detection of event relations. In *Proceedings of the AAAI 2009 Spring Symposium on Learning by Reading*, pages 81–87, Stanford, CA.
[25] Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.
[26] Panevová, J. (1974). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40.
[27] Panevová, J. (1975). On verbal frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 23:17–52.
[28] Rosen, A. and Vavřín, M. (2015). Korpus InterCorp, verze 8 z 4. 6. 2015. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.
[29] Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht – Academia, Prague.
[30] Skoumalová, H. (2001). *Czech Syntactic Lexicon*. Charles University in Prague, Prague.
[31] Sparck, J. K. (1986). Synonymy and semantic classification. Edinburgh University Press (Edinburgh information technology series 1).
[32] Urešová, Z. (2011). *Valenční slovník Pražského závislostního korpusu (PDT-Vallex). Studies in Computational and Theoretical Linguistics 9*, UK Praha.

[33] Urešová, Z., Dušek, O., Fučíková, E., Hajič, J., and 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)*, pages 124–128, ACL, Stroudsburg, PA, USA.
[34] Urešová, Z., Fučíková, E., and Šindlerová, J. (2016). CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.