Textual Coverage of Eventive Entries in Lexical Semantic Resources

Eva Fučíková, Cristina Fernández-Alcaina, Jan Hajič, and Zdeňka Urešová
Institute of Formal and Applied Linguistics
Charles University, Faculty of Mathematics and Physics, Computer Science School
Malostranské nám. 25, Prague 1, Czech Republic
{fucikova,alcaina,hajic,uresova}@ufal.mff.cuni.cz

Abstract

This short paper focuses on the coverage of eventive entries of some well-known lexical semantic resources when applied to random running texts taken from the internet. In order to get the widest coverage, only verbs have been chosen for the comparison, to get as many resources as possible (even though some of the resources cover other parts of speech as well). While coverage gaps are often reported for manually created lexicons (which is the case of most semantically-oriented lexical ones), it was our aim to quantify these gaps, cross-lingually, on a new purely textual resource set produced by the HPLT Project from crawled internet data. Several English, German, Spanish and Czech lexical semantic resources have been selected for this experiment. We also describe the challenges related to the fact that these resources are (to a varying extent) semantically oriented, meaning that the texts have to be preprocessed to obtain lemmas (base forms) and some types of MWEs before the coverage can be reasonably evaluated, and thus the results are necessarily only approximate. The coverage of these resources, with some exclusions as described in the paper, range from 41.00% to 97.33%, confirming the need to expand at least some (even well-known) resources to cover the prevailing source of today's textual resources with regard to lexical units describing events or states (or possibly other eventive mentions).

Keywords: language resource, lexical semantics, event types, ontology, text corpora, plain text, textual coverage

1. Introduction

Lexical semantic resources and ontologies, together with their syntactic counterparts, play an important role in today's NLP, even in the age of powerful, but often factually incorrect LLMs like ChatGPT or similar. Their (obvious) disadvantage is however that due to the fact that they are overwhelmingly manually curated, they are always more or less incomplete. We are thus interested in their coverage on running texts, i.e., measuring how many occurrences of words (tokens) in some text actually do appear in the lexical resource.[1] In order to make the comparison as broad as possible, we have only included verbs from the resources being compared. Polysemy has not been considered due to the absence of reliable (and comparable across languages and/or resources) word sense disambiguation tools capable of accommodating the diversity of the resources. While this approach introduces errors (by increasing coverage because of the inevitable inclusion of non-matching senses), we still believe that when comparing the resources on relative basis, the coverage figures are useful even if they cannot be taken as fully accurate in absolute terms.

There are many papers describing methods and processes to increase coverage, both type-based and token-based, using various approaches, from manual (e.g., (Sio and Morgado da Costa, 2022)) to semi-automatic to fully automatic (with the expected increase in noise inversely proportionate to the manual effort put in), e.g., (Feely et al., 2012; Gábor et al., 2012; Samvelian et al., 2014; Nimb et al., 2021). Increased coverage can also be obtained indirectly via linking of resources where each of them covers different areas of the language, as in SemLink (Stowe et al., 2021), SynSemClass (Urešová et al., 2020, 2022) or BabelNet (Navigli and Ponzetto, 2010).

However, we could not find comparable figures regarding the coverage of the existing resources on large texts, especially those taken from the internet, available in large quantities. This paper thus tries to fill this gap for languages that have several such lexical resources available.

The paper is structured as follows: Sect. 2 describes the data used (both the textual and lexical resources), Sect. 3 describes the data preprocessing necessary to match the lexical resources' entries to text tokens, and Sect. 4 tabulates and discusses the results. Finally, we conclude and draw future plans in Sect. 5. The data on which this paper builds and the full outputs are available for verification and reproducibility purposes at `http://hdl.handle.net/11372/LRT-5444`.

[1] In this paper, we do not cover [pun intended] lexical coverage, i.e., the percentage of types which appear in the lexicon, since even if it might be an interesting figure, it is not much relevant when processing data.

LREC-COLING 2024, pages 15835–15841

2. Data

2.1. The Corpora Used

For this study, we have chosen data recently produced by the project called High Performance Language Technologies (HPLT), which aims at collecting large plain text data in 80+ languages and then high-performance computing to build powerful and efficient language and translation models.[2] For our purposes, we have used monolingual corpora formatted as JSONL files which are compiled from large web crawls provided by the Internet Archive project[3] and CommonCrawl.[4] The HPLT project has released its first dataset in September 2023;[5] this is the data we have used, even though a new (cleaner, but smaller) version 1.2 of the HPLT data exists at the time of the final submission.[6]

For each of the languages there is a list ("map") of files containing the data.[7] We have chosen, for all our languages (English, Spanish, German, and Czech) one sample called (`3.jsonl.zst`).[8] From each of these files, the first 125,000 entries have been kept, and the "`text`" field extracted from each JSONL entry. Each such text string contains a complete document as downloaded and processed by the HPLT project to get a "clean" text. These limits have been set to keep the sample text corpus for each language around 100 million tokens. The exact sizes of the samples are:

Language	Token count
English	104,408,596
German	98,956,434
Spanish	117,477,816
Czech	101,075,477

2.2. Lexical Resources Tested

For our coverage evaluation against the text corpora as described above, we have chosen the following lexical resources:

- FrameNet (Baker et al., 1998) (English),[9] (Ziem, 2020) (German)[10]
- WordNet (Fellbaum, 1998) (English)[11], Czech WordNet 1.9 PDT (Pala et al., 2011)[12]
- SynSemClass (Urešová et al., 2023) (English, German, Spanish and Czech)[13]
- VerbNet (Kipper et al., 2006) (English)[14]
- EngVallex (Cinková et al., 2014) (English)[15]
- PropBank (Palmer et al., 2005) (English)[16]
- German Universal Propositions (Akbik et al., 2015) (German)[17]
- E-VALBU (Kubczak, 2014) (German)[18]
- Spanish Verbal SenSem Lexicon (Fernández et al., 2004) (Spanish)[19]
- AnCora (Taulé et al., 2008) (Spanish)[20]
- PDT-Vallex (Urešová et al., 2014) (Czech)[21]
- VALLEX 4.0 (Lopatková et al., 2020) (Czech).[22]

Most of these 17[23] resources are semantic in nature, except for PropBank, EngVallex (which

[2] https://hplt-project.org
[3] https://archive.org
[4] https://commoncrawl.org
[5] https://hplt-project.org/datasets/v1
[6] https://hplt-project.org/datasets/v1.2
[7] such as https://data.hplt-project.org/one/monotext/de_map.txt
[8] This file ("shard", numbered as 3) is present for all of the four languages investigated. Given our experience with initial and final files from any collection, we have chosen the 3rd shard under the assumption that it might be more "random" than the shards 1 and 2. No explicit experiments to confirm this have been made, though.
[9] https://framenet.icsi.berkeley.edu, version `framenet_v17-1`
[10] https://framenet-constructicon.hhu.de/framenet/frameindex, downloaded list of entries 10/18/2023
[11] https://wordnetcode.princeton.edu/wn3.1.dict.tar.gz
[12] http://hdl.handle.net/11858/00-097C-0000-0001-4880-3
[13] https://github.com/fucikova/SynSemClass_multi/commits/main/Lexicons, downloaded version 10/19/2023
[14] https://github.com/cu-clear/verbnet/tree/master/verbnet3.4, last commit 11/10/2022
[15] http://hdl.handle.net/11858/00-097C-0000-0023-4337-2
[16] https://github.com/propbank/propbank-frames/releases/tag/v3.1, last stable version 9/1/2016
[17] http://alanakbik.github.io/UniversalPropositions_German/index.html, version downloaded 10/18/2023
[18] https://grammis.ids-mannheim.de/verbs: downloaded 3/24/2021 as a list of verbs only
[19] http://grial.edu.es/web/en/downloads-access, request version `lexico_verbal_sensem_espanol-1.1`
[20] https://clic.ub.edu/corpus/system/files/2022-01/ancoralex-es-2.0.3.zip
[21] http://hdl.handle.net/11858/00-097C-0000-0023-4338-F
[22] http://hdl.handle.net/11234/1-3524
[23] Counting different language versions of FrameNet, WordNet and SynSemClass separately.

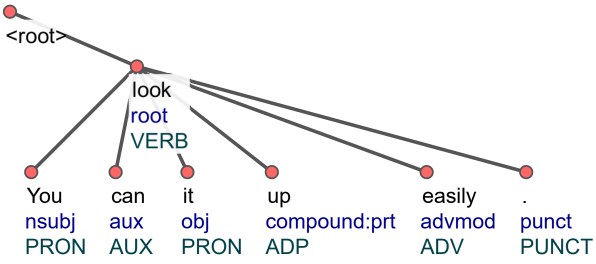

Figure 1: Using a dependency parser output for identifying phrasal verbs - example: (look_up) in the sentence *You can look it up easily.*

also includes most of PropBank verbs), GUP, E-VALBU, AnCora, PDT-Vallex and VALLEX, which display mostly syntactic features like valency, even though they contain some semantic features as well:

Language	Semantic Lexicons	Syntactic(-semantic) Lexicons
English	4	2
German	2	2
Spanish	2	1
Czech	2	2

3. Data Preprocessing

Given the nature of lexical resources, especially those referring to eventive word senses (or meanings, as the semantic ones inevitably do), the plain texts cannot be used directly, since the various forms (especially for highly inflective languages like Spanish or Czech) do not match the lexical entries, which are typically verb lemmas or other base forms, often even in the form of a multiword expression (MWE), such as for reflexive verbs (de: *sich verstellen*, cs: *šířit se*, etc.) or phrasal verbs (en: *look up* - see Fig. 1 for an example of using the output of the **UDPipe** parser for identifying a phrasal particle (up) attached to a verb (look), using the compound:prt dependency relation). Text analysis has to be used to get the lemmas or base verb forms to match against the lexical units in the lexical resource entries. In addition, some words types have to be excluded due to their non-content nature, such as modal verbs–these are normally not included as an entry in lexical semantic resources. This requires even deeper analysis that just getting the lemmas.

We have used the `UDPipe` tool,[24] capable of performing tagging, lemmatization and syntactic (dependency) analysis in order to find just those verbs for which we need to compute the coverage, and in the right base form, including MWEs.[25]

The UDPipe in version 2 is trained on the Universal Dependencies v2 (Nivre et al., 2020) datasets for more than 100 languages. We have used the 2.12 models (named `<prefix>-ud-2.12-230717`) as follows:[26]

- for Czech: prefix `czech-pdt`,
- for English: prefix `english-ewt`,
- for German: prefix `german-gsd`,
- for Spanish: prefix `spanish-ancora`.

The following attributes (columns in the CoNLL-U format) of the `UDPipe` output have been used:

- the LEMMA column to get the lemma or base form of a reflexive or particle,
- the UPOS column to search for the values of VERB, PRON, and ADP that signal the relevant words,
- the DEPREL column to find components of verbal MWEs (phrasal and reflexive),
- the FORM column to distinguish Czech reflexives *se, si*.

Based on them, we have constructed a “matching-ready” form for each VERB token in the data. While the use of the LEMMA column is obvious, the additional information (especially the syntactic relation for compounds using a particle (`compound:prt`), which the analyzer recognized as being dependent of the VERB) allowed us to match also phrasal verbs (such as en: *break away, look up*), verbs with separated prefix in German (such as de: *mitgehen*) and reflexives (cs: *smát se*, de: *sich vorstellen*). The manual inspection of the lexical resources used enabled us then to correctly form the final matching lexical string (pronoun/particle before/after the verb lemma, joined by space, comma or underscore).

Given that (a) the texts are relatively noisy in terms of various formatting problems, missing spaces etc., and (b) the UDPipe tool still makes some (albeit rare) mistakes even on correct verbs (typically in short or nonstandard contexts), we have computed “maximum noise” figures that show how much the coverage might be influenced (to the worse) by these (possibly) non-verbs. The figures are based on reliable, manually curated sources of lexicons or verbal lemma lists extracted from them. Examples of non-verbs are strings such as in en: *25build*, de: *Ursachen* or

[24] `https://ufal.mff.cuni.cz/udpipe/2`

[25] The syntactic dependency analysis has been used only for the various types of MWE identification, however.

[26] `https://ufal.mff.cuni.cz/udpipe/2/models#universal_dependencies_212_models`

Language	Verb lemmas	Occurrences in corpus	Possible noise (lemmas)	Possible noise (tokens)	Attested lemmas	Attested tokens
English	34,526	9,197,397	80.57%	2.78%	6,710	8,941,473
German	56,443	6,244,389	95.40%	18.00%	2596	5,120,287
Spanish	57,481	8,843,767	95.35%	13.54%	2675	7,646,483
Czech	37,156	5,462,633	79.81%	7.96%	7501	5,027,974

Table 1: Filtering out corpus noise

cs: *implantát*, wrongly analyzed or "guessed" by UDPipe to be VERBs. The statistics on this "maximal" noise are summarized in Table 1: the noise in terms of lemmas is very high, but the token counts are influenced much less.

Language	Verbs excluded	Percent
English	*be can could have may make must will would*	5.44%
German	*dürfen haben können mögen müssen sein sollen wollen*	3.90%
Spanish	*deber poder querer saber ser soler*	1.56%
Czech	*být dělat lze muset moci mít smět*	10.91%

Table 2: Modal (and other) verbs excluded, in %

In addition, modals, and copulas have been excluded (Table 2). These either do not possibly represent verbs that would be expected to have a separate entry in lexical semantic resources, or are ambiguous enough not to be included in these resources. Given that the texts cannot be (as of yet) analyzed fully semantically for a better matching, they have been excluded, too.

4. Results

The resulting coverage on the final set of lemmas tested for coverage is presented in Table 3, with the "winners" in each language in bold. The basis for the coverage percentages (last column) is still the original number of verb occurrences in the texts used, i.e., the third column as seen in Table 1,[27] minus the excluded modals and other such verbs, as seen in Table 2.

The resources are ordered from the "most semantic" ones (FrameNet, WordNet, SynSemClass) to the "least semantic," such as the valency lexicons used for the Spanish and Czech treebank annotation. With the exception of German E-VALBU valency lexicon, the more syntactically-oriented lexicons show higher coverage (with PropBank showing its maturity with the highest coverage of all the lexicons), while among the semantic ones, WordNet wins for English (and overall), but has poor coverage for Czech. However, WordNet –except for the hierarchical relations among its synsets– does not offer additional semantic (or even syntactic) information for annotation or other applications, as opposed to FrameNet(s), VerbNet or SynSemClass. From these richly annotated semantic/syntactic resources, SynSemClass for English (and to a certain extent, for Czech) offers the best coverage, followed very closely by VerbNet.

While keeping the original verb occurrences counts despite the noise in the data, as presented in Table 1, lowers the coverage due to possibly dubious verbs being counted, the filtering of modals and copulas (Table 2), on the other, hand inevitably increases the coverage. However, we deem it fair to do so, as it is not expected that these verbs would have an entry in semantic lexical resources (WordNet is an exception, but for comparison purposes, it has been simply treated the same way).

The controversial point might be the exclusion of verbs like *to be*, *to have* or *to do*, since they do have, depending on context or use, its own semantic "content" meaning (e.g., existential *to be*) and are (or should) be covered in resources like FrameNet, VerbNet or SynSemClass. However, even with the UDPipe analysis, it would be difficult to distinguish, e.g., the many senses of *to have* and its counterparts in the other languages. We thus hope that by excluding them, the coverage will be closer to the actual one than by *not* excluding them. This has been done uniformly across all the resources, with the aim of minimizing its influence on the differences among the resources when comparing them.

5. Conclusions and Future Work

As described and presented in our paper, we have tried to quantify the coverage of widely used lexical resources, mainly those reflecting semantics, on recent internet texts. The results vary widely, with some of the most popular resources (Word-

[27] Table 1 serves only as an indication of noise in the data, but the possibly problematic verbs have *not* been excluded from the coverage computation.

Lexical resource	Language	Coverage (tokens)	Coverage (percent)
FrameNet	English	7,464,343	85,82%
FrameNet	German	2,460,251	41.00%
WordNet	**English**	8,465,366	**97.33**%
WordNet	Czech	2,586,432	53.15%
SynSemClass	English	7,959,432	91.51%
SynSemClass	German	3,339,962	55.66%
SynSemClass	Spanish	6,627,769	76.13%
SynSemClass	Czech	4,125,642	84.77%
VerbNet	English	7,657,626	88.04%
SenSem	Spanish	4,732,521	54.36%
PropBank	English	8,433,779	96.97%
EngVallex	English	8,275,981	95.15%
E-VALBU	German	3,235,501	53.92%
GUP	**German**	4,729,042	**78.81**%
AnCora	**Spanish**	7,508,545	**86.25**%
PDT-Vallex	**Czech**	4,374,973	**89.90**%
VALLEX	Czech	4,239,811	87.12%

Table 3: Coverage of all the lexical resources used

Net and PropBank/EngVallex, all for English) and some others showing relatively high (albeit not perfect, as expected) coverage. For the non-English languages, the situation is substantially worse – with exceptions, such as AnCora for Spanish and PDT-Vallex and SynSemClass for Czech.

The matching algorithm can still be substantially improved. The (syntactic) `UDPipe` parser can still provide more information than we have been able to use, such as proper distinction between auxiliary and modal verbs and the content-bearing ones, etc. Of course a good semantic parser would be the ultimate solution to use, alleviating the need for approximations and exclusions –provided that the parser would be trained on an annotation matching the lexical resources (which by itself is a non-trivial task to do for 17 resources), which differ in the treatment of reflexive particles, phrasal verbs, MWEs in general, treatment of light verbs, and in the semantic labeling schemas.

In the future, we also plan to extend the set of lexical resources for which coverage is computed, and redo those for which (if and when) new versions become available. If interest persists, we will publish a "dashboard" where further figures on coverage on these and possibly additional resources will be posted.

As is becoming common practice, we have packaged and published the data on which this paper builds as well as its full outputs, to allow for verification.[28]

[28] http://hdl.handle.net/11372/LRT-5444

6. Acknowledgements

The work described herein has been supported by the Grant Agency of the Czech Republic under the EXPRO program as project "LUSyD" (project No. GX20-16819X), by the Ministry of Education, Youth and Sports under the Inter-Excellence programme, project "Universal Meaning Representation" (project No. LUAUS23283). In addition, we acknowledge the use of data resulting from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee, grant number 10052546. It has also used resources hosted, i.a., by the LINDAT/CLARIAH-CZ Research Infrastructure (projects LM2018101 and LM2023062, supported by the Ministry of Education, Youth and Sports of the Czech Republic). We would also like to thank our team of annotators on the SynSemClass ontology work under the LUSyD project, and also to the authors of all the other 16 lexical resources used, without whom this project and paper could not exist.

7. Bibliographical References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 397–407, Beijing, China. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 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.

Wes Feely, Claire Bonial, and Martha Palmer. 2012. Evaluating the coverage of verbnet. In *Joint 8th ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation, Pisa, Italy, October 3-5, 2012*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

A. Fernández, G. Vázquez, and I. Castellón. 2004. Sensem: base de datos verbal del español. In *IX Ibero-American Workshop on Artificial Intelligence*, pages 155–163. IBERAMIA. Puebla de los Ángeles, Mexico, ISBN: 968-863-786-6.

Kata Gábor, Marianna Apidianaki, Benoît Sagot, and Éric Villemonte de La Clergerie. 2012. Boosting the coverage of a semantic lexicon by automatically extracted event nominalizations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1466–1473, Istanbul, Turkey. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of LREC*, volume 2006.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

R. Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *ACL*.

Sanni Nimb, Bolette Pedersen, and Sussi Olsen. 2021. DanNet2: Extending the coverage of adjectives in DanNet based on thesaurus data (project presentation). In *Proceedings of the 11th Global Wordnet Conference*, pages 267–272, University of South Africa (UNISA). Global Wordnet Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. 2011. Czech WordNet 1.9 PDT. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Pollet Samvelian, Pegah Faghiri, and Sarra El Ayari. 2014. Extending the coverage of a MWE database for Persian CPs exploiting valency alternations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4023–4026, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ut Seong Sio and Luís Morgado da Costa. 2022. Enriching linguistic representation in the Cantonese Wordnet and building the new Cantonese Wordnet corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 70–78, Marseille, France. European Language Resources Association.

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

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020. SynSemClass linked lexicon:

Mapping synonymy between languages. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 10–19, Marseille, France. European Language Resources Association.

Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, and Jan Hajič. 2022. Making a semantic event-type ontology multilingual. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1332–1343, Marseille, France. European Language Resources Association.

Alexander Ziem. 2020. *Word meanings as frames: a framework model for the analysis of lexical meanings*, page 27–56. Stauffenburg, Tübingen.

8. Language Resource References

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

Kubczak, Jacqueline. 2014. *Valenzwörterbuch E-VALBU*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University,http://hdl.handle.net/11372/LRT-1162.

Lopatková, Markéta and Kettnerová, Václava and Vernerová, Anna and Bejček, Eduard and Žabokrtský, Zdeněk. 2020. *VALLEX 4.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University,http://hdl.handle.net/11234/1-3524.

Pala, Karel and Čapek, Tomáš and Zajíčková, Barbora and Bartůšková, Dita and Kulková, Kateřina and Hoffmannová, Petra and Bejček, Eduard and Straňák, Pavel and Hajič, Jan. 2011. *Czech WordNet 1.9 PDT*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.

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

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