SynSemClass for German: Extending a Multilingual Verb Lexicon

Peter Bourgonje[1], Karolina Zaczynska[1], Julián Moreno-Schneider[1],
Georg Rehm[1], Zdenka Uresova[2], and Jan Hajič[2]

[1] DFKI GmbH, Speech and Language Technology, Berlin, Germany
[2] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Prague, Czech Republic

Abstract. We present the concept of extending a multilingual verb lexicon also to include German. In this lexicon, verbs are grouped by meaning and by semantic properties (following frame semantics) to form multilingual classes, linking Czech and English verbs. Entries are further linked to external lexical resources like VerbNet and PropBank. In this paper, we present our plan also to include German verbs, by experimenting with word alignments to obtain candidates linked to existing English entries, and identify possible approaches to obtain semantic role information. In addition, we identify German-specific lexical resources to link to. This small-scale pilot study aims to provide a blueprint for extending a lexical resource with a new language.

Keywords: Linked Lexicon · Semantics · Synonymy · Parallel Corpus

1 Introduction

In Natural Language Processing (NLP), lexical resources play an important role for supporting a computer's understanding of human language. Such machine-readable resources not only list lexical surface forms, but often also provide additional syntactic and semantic properties, focusing on particular word groups [10, 33, 38], use cases [9], or languages other than English[3] [13, 27, 42]. While recent neural technologies have proven to be very successful at a number of NLP tasks learning from unannotated data only (i. e., without using any form of explicitly encoded knowledge external to the language data itself) [8, 26, 28], these approaches rely on the availability of large amounts of data, which may not always be available for the desired language or domain. Additionally, certain semantic properties may not be sufficiently picked up on by a system trained on large amounts of unannotated data only [4, 45]. Moreover, such systems are, by design, sensitive to bias in the training data [36, 3].

This paper focuses on a verb lexicon in which synonym classes are defined in terms of both semantic and syntactic properties. Computational verb lexicons,

[3] Which is, for many paradigms and tasks, the most popular language in NLP research and Language Technology applications [29, 30].

such as VerbNet, have proven to be useful in supporting a wide range of NLP tasks and applications, including information extraction [21], sentence similarity [41] and event extraction [44]. With regard to the use of VerbNet for event extraction, a new release of the lexicon [5] includes a modified version of the semantic representation of verbs to provide an improved representation of event and subevent structures in language.[4] Unlike the majority of monolingual verb lexicons, a multilingual aligned verb lexicon supports deeper understanding and comparability of the usage of verbs in different languages, and simultaneously provides a close interaction between syntactic and semantic features of verbs. Additionally, a multilingual resource of this kind is able to provide a broader range of applications, among others, cross-lingual search.

In this paper, we outline our plans to extend an existing, bi-lingual (Czech and English) verb lexicon in which verbs are grouped into synonym classes both meaning-wise (verb senses, semantic roles) and structurally (valency arguments). To ease adaptation and increase compatibility with existing resources, verb classes and individual entries are linked to existing lexical resources where possible. The existing lexicon is described in [39] and we outline our plans to both validate this lexicon and its classes and extend it to include a subset of German verbs as well. To demonstrate our plans, we perform a small-scale pilot study enabling us to test the outcome and reliably estimate the time and resources needed for the extension plan.

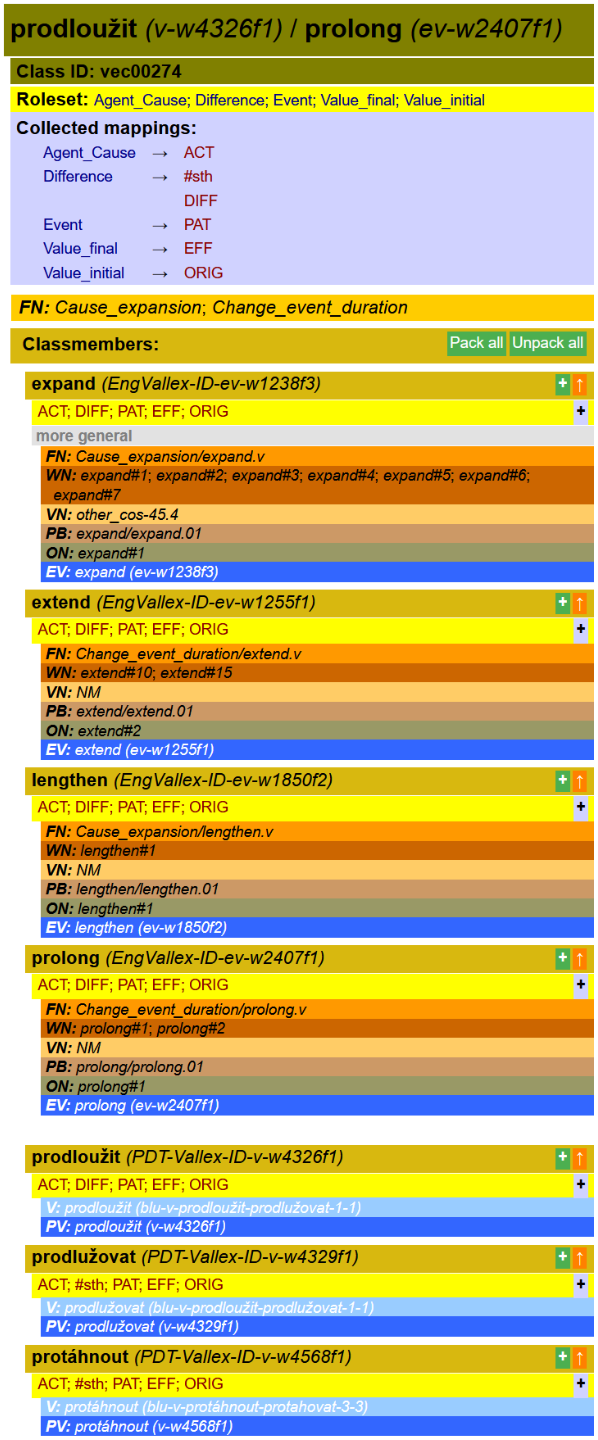
We first provide a brief description of the existing, bi-lingual lexicon and its key properties (Section 2). Then, we explain the corpus we use for the pilot study (Section 3), followed by the procedure to extract word alignments (Section 4.1) and semantic role properties (Section 4.2) including a description of the resources we link to (Section 5). Finally, Section 6 sums up our key findings and provides an outlook on the full-scale study we intend to perform as future work.

2 SynSemClass

The SynSemClass lexicon currently[5] groups Czech and English verbs by meaning and structural properties. It contains 145 synonym classes with 3,515 Czech and English verb senses; 2,027 in English and 1,488 in Czech. Each class is assigned a set of semantic roles and the prototypical meaning of the synonym class representing the English and Czech verb sense. The lexicon was developed in a bottom-up fashion. Class member candidates originate from actual corpus examples (the Prague Czech-English Dependency Treebank [12]), starting off with 200 semi-randomly chosen Czech verbs (and their valency information, coming from the monolingual PDT-Vallex [40]), and going from Czech to English and vice versa, with manual adjudication steps in between. There is no specific

[4] Event detection is our main use case, see, e. g., [32], [22], [31].

[5] At the time of writing; a new version, SynSemClass3.0, with 455 classes containing approx. 3,800 verbs on each side was published at the end of December 2020, see http://hdl.handle.net/11234/1-3439.



prodloužit *(v-w4326f1)* / **prolong** *(ev-w2407f1)*

Class ID: vec00274

Roleset: Agent_Cause; Difference; Event; Value_final; Value_initial

Collected mappings:

Agent_Cause	→	ACT
Difference	→	#sth DIFF
Event	→	PAT
Value_final	→	EFF
Value_initial	→	ORIG

FN: *Cause_expansion; Change_event_duration*

Classmembers:

expand *(EngVallex-ID-ev-w1238f3)*

ACT; DIFF; PAT; EFF; ORIG

more general

FN: *Cause_expansion/expand.v*
WN: *expand#1; expand#2; expand#3; expand#4; expand#5; expand#6; expand#7*
VN: *other_cos-45.4*
PB: *expand/expand.01*
ON: *expand#1*
EV: *expand (ev-w1238f3)*

extend *(EngVallex-ID-ev-w1255f1)*

ACT; DIFF; PAT; EFF; ORIG

FN: *Change_event_duration/extend.v*
WN: *extend#10; extend#15*
VN: *NM*
PB: *extend/extend.01*
ON: *extend#2*
EV: *extend (ev-w1255f1)*

lengthen *(EngVallex-ID-ev-w1850f2)*

ACT; DIFF; PAT; EFF; ORIG

FN: *Cause_expansion/lengthen.v*
WN: *lengthen#1*
VN: *NM*
PB: *lengthen/lengthen.01*
ON: *lengthen#1*
EV: *lengthen (ev-w1850f2)*

prolong *(EngVallex-ID-ev-w2407f1)*

ACT; DIFF; PAT; EFF; ORIG

FN: *Change_event_duration/prolong.v*
WN: *prolong#1; prolong#2*
VN: *NM*
PB: *prolong/prolong.01*
ON: *prolong#1*
EV: *prolong (ev-w2407f1)*

prodloužit *(PDT-Vallex-ID-v-w4326f1)*

ACT; DIFF; PAT; EFF; ORIG

V: *prodloužit (blu-v-prodloužit-prodlužovat-1-1)*
PV: *prodloužit (v-w4326f1)*

prodlužovat *(PDT-Vallex-ID-v-w4329f1)*

ACT; #sth; PAT; EFF; ORIG

V: *prodlužovat (blu-v-prodloužit-prodlužovat-1-1)*
PV: *prodlužovat (v-w4329f1)*

protáhnout *(PDT-Vallex-ID-v-w4568f1)*

ACT; #sth; PAT; EFF; ORIG

V: *protáhnout (blu-v-protáhnout-protahovat-3-3)*
PV: *protáhnout (v-w4568f1)*

Fig. 1. SynSemClass entry for *p*rolong/prodloužit synonym class in online browser

model or lexicographic theory behind it, even though the underlying syntactic-semantic lexicons for Czech and English (PDT-Vallex and EngVallex), which provide the (current) sense distinctions, are based on the Functional Generative Description theory [35]. The notion of synonymy used is based on the “loose” definition of synonymy by Lyons and Jackson [18, 15], or alternatively and very closely, on both “near-synonyms” and “partial synonyms” as defined by Lyons [19, 7] or “plesionyms” as defined by Cruse [6]. The SynSemClass lexicon is also available online.[6] For more details on its creation process as well as inter-annotator agreement numbers see [39]. An example of a current SynSemClass class entry for (*prolong/prodloužit*) is depicted in Fig. 1.

In the following sections, we describe our plans to expand this bi-lingual lexicon to German, based on the preliminary results of a pilot study.

3 Corpus

The original lexicon was created from a bi-lingual and (automatically) word-aligned corpus, based on a part of the Penn Treebank [20] manually translated into Czech by a professional translator. The Czech sentences were automatically processed to annotate them with morphological information and dependency parses. To maintain this data-driven basis of the original lexicon linking Czech and English verbs on the basis of their usage in a corpus, we thus need either a parallel Czech-German corpus or a parallel English-German corpus.

After settling upon a sentence-aligned parallel corpus, word alignments need to be extracted to establish links between either English or Czech verbs on the source side and German verbs on the target side. Because German is typologically closer to English than to Czech (German and English are West-Germanic languages, Czech is a Slavic language), we expect word alignment tools, exploiting syntactic information, to perform better on an English-German parallel corpus. A large number of candidate corpora are listed in the OPUS corpus browser[7]. Because several of these originate from a particular domain (European Parliament meeting transcripts [16], movie subtitles [17] or Wikipedia [43]), but SynSemClass verbs are not tuned to any particular domain or genre, we simply select the largest resource (which also does not seem to be targeted at one particular domain or genre), i. e., ParaCrawl[8]. The English-German part of ParaCrawl contains over 82 million parallel sentences, with 1.5 billion tokens on the German and 1.6 billion tokens on the English side.

[6] https://lindat.mff.cuni.cz/services/SynSemClass

[7] http://opus.nlpl.eu

[8] https://paracrawl.eu

4 Method

4.1 Word Alignments

For the extraction of word alignments we use MGIZA [23]. Our small scale pilot study is based on the first 5 million sentences of the EN-DE ParaCrawl corpus, containing approx. 94 million German and approx. 98 million English tokens.

Further narrowing the scope of our pilot study, we select the canonical forms of the English classes starting with *a* in SynSemClass, i. e., the following 13 verbs: *agree*, *allow*, *announce*, *applaud*, *approach*, *approve*, *arise*, *arrest*, *assert*, *assume*, *attend*, *avoid*, *await*. For each we automatically extract the most frequent alignments with a cut-off of 0.2%, meaning that if the particular English verb was aligned to a particular German word or phrase in more than 0.2% of cases (in English) it was selected and discarded otherwise. This list was then manually checked by one of the authors of this paper in order to eliminate the many irrelevant entries among the automatically extracted list. Examples are verb/noun ambiguity (at this point, we only had word alignments, no part-of-speech-tag information yet), such as for *approach*, which was aligned to the German *ansatz* (*approach*, but only in the noun sense) in 28% of cases, and to *nähern* (*approach* in the verb sense) in 24% of cases. Another frequent reason for filtering out automatically extracted alignments was the co-extraction of pronouns (*erlauben* ***es***, *allow* ***it***) or particles (***zu*** *genehmigen*, ***to*** *approve*), where the actual verb was among the list already. For some relatively infrequent verbs (such as *applaud*, occurring only 65 times (in infinitival form) in our 5 million sentence subset of the corpus, compared to 3,448 for *agree* or 10,989 for *allow* (again, counting infinitival forms only), some obviously non-sensical alignments still made it past the 0.2% threshold, such as *spielen ihre rolle bis grenzen möglichen mithin spendest beifall* ("*play their role to the limits possible therefore give applause*") for *applaud*.

After manually processing the automatically extracted alignment list, we were left with 100 German root forms of verbs as candidate entries (7.7 verbs per seed verb on average, with the most alignments (16) for *approve* and *arise*, and the least alignments (3) for *await*). The next step is to obtain more structural and semantic information for these candidate entries.

4.2 Semantic Role Labeling

In addition to meaning, the semantic roles that a class can assign are important for the clustering of verbs in SynSemClass. In the creation of the Czech-English lexicon, the semantic roles (SRs) are "mostly taken from FrameNet" [39, p. 13]. The German equivalent, the collaborative *FrameNet des Deutschen*[9], does not specify SRs, but does link to the original FrameNet [2], and SR information could be retrieved from there, in the same way this was done for the Czech-English SynSemClass lexicon. This will be consulted with the SynSemClass entries in the future, in order to keep a common set of roles for each class.

[9] https://gsw.phil.hhu.de/framenet/frameindex

Alternatively, the 2009 CoNLL shared task included SR labeling for seven different languages (including German), inspiring many automated approaches [11]. More recently, inspired by transformer architectures and their multilingual capabilities, there have been attempts at contributing to the SR labeling task using neural approaches [14, 37]. Such automated procedures support a more data-driven specification of the SRs of particular verbs and are able to specify this information for verbs that occur in the corpus (which, in our case, is rather large). Their downside is the expected quality of the output; [14] report F_1-scores ranging from 81.41 for German and 91.00 for English, demonstrating that at least for German, such automatic SR labeling systems still have a considerable margin of error. Manually curated resources such as FrameNet (and its German equivalent) are likely to provide better quality for the verbs they cover, but obviously will not help us for verbs not included in the lexical resource.

For our study, we searched our 100 root forms of German verbs in *FrameNet des Deutschen*. Because these are sometimes described using the verb (e. g., for "besuchen"[10]) and sometimes using the corresponding noun (e. g., for "Verhaftung"[11]), we made sure that the search string would match both the verb and corresponding noun. The German FrameNet contains 834 entries, and only 23 of our 100 verbs were found in this way. For these entries, we can thus obtain SR information through the link to the English FrameNet entry. For the remaining 77 entries, however, we must resort to other means of getting SR information. We consider the approach of [14], who made their code publicly available, a promising start for processing sentences containing verbs that are not included yet in the German FrameNet. Given their F_1-score, which is impressive but still leaves considerable room for improvement, the output can be manually checked for individual entries before including them in SynSemClass as German verbs.

5 Linking to Existing Resources

The original SynSemClass lexicon is linked to a range of resources (see Section 3.2 in [39]), including popular resources like FrameNet [2], VerbNet [34] and PropBank [24]. By linking the new German entries to the existing classes in SynSemClass, we thus establish a link between our German verbs and, among others, VerbNet entries. As for German-specific resources, we consider linking to *FrameNet des Deutschen* an important way of connecting SynSemClass to existing lexical resources for German. Additional resources we consider linking to are 1) GermaNet [13], a lexical resource for German that contains nouns, verbs and adjectives and groups them by synsets and defines relations between these synsets in the WordNet tradition, and 2) Universal Proposition Banks [1][12], which provides a list of German verbs annotated with frame and role labels linked to the English Proposition Bank (PropBank) [25]. Furthermore, we

[10] https://gsw.phil.hhu.de/framenet/frame?id=441
[11] https://gsw.phil.hhu.de/framenet/frame?id=499
[12] https://github.com/System-T/UniversalPropositions

plan to explore if meaningful links to resources available in the Linguistic Linked Open Data cloud (LLOD) can be established.

Such links between SynSemClasses with German verbs in them and existing German resources will probably have to be established manually; the German FrameNet has an intuitive search interface that we also used in Section 4.2.

6 Conclusion

This paper presents our plan to expand a bi-lingual (Czech and English) lexicon of verbs to a multilingual verb lexicon by including German verbs. The existing SynSemClass lexicon groups verbs by their meaning and by semantic role properties and links them internally (Czech and English) and externally (to existing resources such as FrameNet, VerbNet and PropBank). To expand the lexicon to include German verbs, we thus need 1) correspondences between German verbs and their English and Czech counterparts, and 2) SR information for the German verbs. We plan to obtain this using 1) word alignments extracted from a parallel English-German corpus, and 2) exploiting existing German resources (*FrameNet des Deutschen*) in combination with recent advances in automatic semantic role labeling approaches. We executed a small-scale pilot study using 5 million of the 82 million aligned sentences in a candidate corpus, and using only 13 entries from SynSemClass. This allows us to estimate the time and resources required to perform the full-scale exercise.

Extracting word alignments from a parallel corpus (for which we use MGIZA) is a time-consuming process, but mostly takes compute time (over 50 hours on a single laptop (i7 2.20Ghz, 24GB RAM), but we estimate that at least half of this can be optimised through parallelisation. Manual filtering of irrelevant alignments for our 13 pilot verbs took approx. one hour. We do note that in the creation process of the SynSemClass lexicon, the authors went back and forth between Czech and English in three steps [39, p. 13] (Figure 2), to find more alignment candidates. In our pilot study we only perform the first step, which already expands the seed size of 13 English verbs to a list of 100 German candidates (after filtering). Including more alignment steps will thus further increase the size of the candidate set, but by a smaller factor (i. e., we expect additional alignment steps to increase by a factor considerably smaller than 7.7).

Collecting the semantic role information was done completely manually in our pilot study and took ca. 0.5 hours, but resulted in this information for only 23% of the entries. We did not yet experiment with automatic approaches to semantic role labeling. This procedure will take relatively cheap compute processing time; the amount of time and effort needed to manually amend results before the entries can be included in SynSemClass remains to be seen.

Acknowledgment The work presented in this paper has received funding from the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (Wachstumskern no. 03WKDA1A), and from the projects LINDAT/CLARIAH-CZ of the MEYS Czech Republic (no. LM2018101) and LUSyD of the GAČR (no. GX20-16819X).

References

1. Akbik, A., Guan, X., Li, Y.: Multilingual aliasing for auto-generating proposition Banks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 3466–3474. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/C16-1327
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: 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. p. 86–90. ACL '98/COLING '98, Association for Computational Linguistics, USA (1998). https://doi.org/10.3115/980845.980860, https://doi.org/10.3115/980845.980860
3. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5454–5476. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.485, https://www.aclweb.org/anthology/2020.acl-main.485
4. Boleda, G.: Distributional semantics and linguistic theory. CoRR **abs/1905.01896** (2019), http://arxiv.org/abs/1905.01896
5. Brown, S.W., Bonn, J., Gung, J., Zaenen, A., Pustejovsky, J., Palmer, M.: VerbNet representations: Subevent semantics for transfer verbs. In: Proceedings of the First International Workshop on Designing Meaning Representations. pp. 154–163. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3318, https://www.aclweb.org/anthology/W19-3318
6. Cruse, A.: Lexical Semantics. Cambridge University Press, UK (1986)
7. Cruse, A.: Meaning in Language. An Introduction to Semantics and Pragmatics. Oxford University Press. Oxford, UK (2000)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
9. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06. pp. 417–422 (2006), SENTIWORDNET: A publicly available lexical resource for opinion mining
10. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, MA (1998)
11. Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., Zhang, Y.: The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task. pp. 1–18. Association for Computational Linguistics, Boulder, Colorado (Jun 2009), https://www.aclweb.org/anthology/W09-1201

12. 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.: Announcing prague czech-english dependency treebank 2.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). pp. 3153–3160. ELRA, European Language Resources Association, İstanbul, Turkey (2012)
13. Hamp, B., Feldweg, H.: GermaNet – A Lexical-Semantic Net for German. In: In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. pp. 9–15 (1997), https://www.aclweb.org/anthology/W97-0802
14. He, S., Li, Z., Zhao, H.: Syntax-aware multilingual semantic role labeling. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5350–5359. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1538, https://www.aclweb.org/anthology/D19-1538
15. Jackson, H.: Words and Their Meaning. Routledge (1988)
16. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit. pp. 79–86. AAMT, AAMT, Phuket, Thailand (2005), http://mt-archive.info/MTS-2005-Koehn.pdf
17. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1147
18. Lyons, J.: Introduction to Theoretical Linguistics. Cambridge University Press (1968)
19. Lyons, J.: Linguistic Semantics. Cambridge University Press (1995)
20. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993), https://www.aclweb.org/anthology/J93-2004
21. Mausam, Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 523–534. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), https://www.aclweb.org/anthology/D12-1048
22. Moreno-Schneider, J., Srivastava, A., Bourgonje, P., Wabnitz, D., Rehm, G.: Semantic Storytelling, Cross-lingual Event Detection and other Semantic Services for a Newsroom Content Curation Dashboard. In: Popescu, O., Strapparava, C. (eds.) Proceedings of the Second Workshop on Natural Language Processing meets Journalism – EMNLP 2017 Workshop (NLPMJ 2017). pp. 68–73. Copenhagen, Denmark (9 2017), 7 September
23. Och, F.J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics **29**(1), 19–51 (2003)
24. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (Mar 2005). https://doi.org/10.1162/0891201053630264, https://doi.org/10.1162/0891201053630264

25. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics **31**(1), 71–106 (2005). https://doi.org/10.1162/0891201053630264, https://www.aclweb.org/anthology/J05-1004
26. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1202, https://www.aclweb.org/anthology/N18-1202
27. Postma, M., van Miltenburg, E., Segers, R., Schoen, A., Vossen, P.: Open Dutch WordNet. In: Proceedings of the Eight Global Wordnet Conference. Bucharest, Romania (2016)
28. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
29. Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Piperidis, S., Deligiannis, M., Galanis, D., Gkirtzou, K., Labropoulou, P., Bontcheva, K., Jones, D., Roberts, I., Hajic, J., Hamrlová, J., Kačena, L., Choukri, K., Arranz, V., Vasiļjevs, A., Anvari, O., Lagzdiņš, A., Meļņika, J., Backfried, G., Dikici, E., Janosik, M., Prinz, K., Prinz, C., Stampler, S., Thomas-Aniola, D., Pérez, J.M.G., Silva, A.G., Berrío, C., Germann, U., Renals, S., Klejch, O.: European Language Grid: An Overview. In: Calzolari, N., Béchet, F., Blache, P., Cieri, C., Choukri, K., Declerck, T., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 3359–3373. European Language Resources Association (ELRA), Marseille, France (5 2020)
30. Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajic, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Pérez, J.M.G., Meertens, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Bars, L.L., Anastasiou, D., Auksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., Smedt, K.D., Garabík, R., Gavriilidou, M., Gromann, D., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Odijk, J., Ogrodniczuk, M., Rögnvaldsson, E., Rosner, M., Pedersen, B., Skadina, I., Tadić, M., Tufiș, D., Váradi, T., Vider, K., Way, A., Yvon, F.: The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In: Calzolari, N., Béchet, F., Blache, P., Cieri, C., Choukri, K., Declerck, T., Isahara, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020). pp. 3315–3325. European Language Resources Association (ELRA), Marseille, France (5 2020)
31. Rehm, G., Schneider, J.M., Bourgonje, P., Srivastava, A., Nehring, J., Berger, A., König, L., Räuchle, S., Gerth, J.: Event Detection and Semantic Storytelling: Generating a Travelogue from a large Collection of Personal Letters. In: Caselli, T., Miller, B., van Erp, M., Vossen, P., Palmer, M., Hovy, E., Mitamura, T. (eds.) Proceedings of the Events and Stories in the News Workshop. pp. 42–51. Association for Computational Linguistics, Vancouver, Canada (8 2017), co-located with ACL 2017
32. Schneider, J.M., Bourgonje, P., Nehring, J., Rehm, G., Sasaki, F., Srivastava, A.: Towards Semantic Story Telling with Digital Curation Technologies. In: Birnbaum,

L., Popescu, O., Strapparava, C. (eds.) Proceedings of Natural Language Processing meets Journalism – IJCAI-16 Workshop (NLPMJ 2016). New York (7 2016)

33. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania (2006), http://verbs.colorado.edu/ kipper/Papers/dissertation.pdf
34. Schuler, K.K., Palmer, M.S.: Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, USA (2005), aAI3179808
35. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in its Semantic and Pragmatic Aspects. D. Reidel, Dordrecht (1986)
36. Shah, D.S., Schwartz, H.A., Hovy, D.: Predictive biases in natural language processing models: A conceptual framework and overview. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5248–5264. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.468, https://www.aclweb.org/anthology/2020.acl-main.468
37. Shi, P., Lin, J.: Simple BERT models for relation extraction and semantic role labeling. CoRR **abs/1904.05255** (2019), http://arxiv.org/abs/1904.05255
38. Stede, M.: DiMLex: A Lexical Approach to Discourse Markers. In: Exploring the Lexicon - Theory and Computation. Edizioni dell'Orso, Alessandria (2002)
39. Uresova, Z., Fucikova, E., Hajicova, E., Hajic, J.: SynSemClass linked lexicon: Mapping synonymy between languages. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. pp. 10–19. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.globalex-1.2
40. Urešová, Z., Štěpánek, J., Hajič, J., Panevova, J., Mikulová, M.: PDT-vallex: Czech valency lexicon linked to treebanks (2014), http://hdl.handle.net/11858/00-097C-0000-0023-4338-F, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague
41. Wali, W., Gargouri, B., Hamadou, A.B.: Sentence similarity computation based on wordnet and verbnet. Computación y Sistemas **21** (2017)
42. Wang, S., Bond, F.: Building the Chinese open Wordnet (COW): Starting from core synsets. In: Proceedings of the 11th Workshop on Asian Language Resources. pp. 10–18. Asian Federation of Natural Language Processing, Nagoya, Japan (Oct 2013), https://www.aclweb.org/anthology/W13-4302
43. Wolk, K., Marasek, K.: Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. CoRR **abs/1509.08881** (2015), http://arxiv.org/abs/1509.08881
44. Xiang, W., Wang, B.: A survey of event extraction from text. IEEE Access **7**, 173111–173137 (2019)
45. Yanaka, H., Mineshima, K., Bekki, D., Inui, K., Sekine, S., Abzianidze, L., Bos, J.: HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019). pp. 250–255. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/S19-1027, https://www.aclweb.org/anthology/S19-1027