Automatic Suggestions Help Extending Eventive Ontology: A Case Study on SynSemClass

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

Abstract

Despite substantial recent progress in many areas of NLP, semantic tasks remain particularly challenging. One such task is the creation (extension, or *annotation*) of semantic ontologies. In this work, we present a case study on the eventive SynSemClass ontology, focusing on the challenges of *semantic annotation* – that is extending the ontology with new lexical units and/or new concepts – both with and without automatic support. We consider two strategies for generating annotation suggestions: (i) a knowledge-driven approach based on a small, carefully curated corpus of verbal valency frames, and (ii) a corpus-driven approach using lemma-based suggestions from a large raw text collection, disregarding semantic homonymy. Our findings show that ontology annotation is inherently difficult, and that automatic annotations statistically significantly reduce this difficulty both in terms of inter-annotator agreement and when compared with gold expert annotations. We discuss the implications for semantic resource creation and extension, as well as the limits of automation in ontology annotation.

Keywords: semantic annotation, evaluation, synonyms, event types, lexical resource, ontology

1. Introduction

Semantic annotation often faces significant challenges, including annotation complexity, language ambiguity, and difficulty achieving a rigorous evaluation. In this work, we present a case study of the SynSemClass (Urešová et al., 2025) eventive concept ontology which represents a hierarchy of concepts (and corresponding words for expressing them in multiple languages) denoting events or states.

In order to achieve a consistent semantic annotation, comparison of methods, and rigorous evaluation, we consider several levels of increasingly more extensive machine-assisted annotation and either a manual or automatic evaluation procedure.

The annotation setting was either purely manual or using a hybrid approach: annotating both manually and automatically at once, that is, employing human annotators as well as automatic systems. Moreover, automatic suggestions were generated from two resources: a knowledge-driven approach based on a relatively small, curated corpus of verb senses, and a corpus-driven approach using lemma-based suggestions.[1]

The evaluation procedures were manual or automatic, where the manual evaluation (Hinze et al., 2019) should compare human annotations with gold annotations provided by an expert who ensures correctness and provides insightful feedback to annotators, while automatic metrics are based on measures such as inter-annotator agreement between annotators, precision, recall, and F1 against a gold standard.

The variety of experimental settings in our study allowed us to ask the following questions:

- **How difficult is semantic annotation of events and states?** We examine this both with and without automatic assistance from deep learning models trained on large text corpora and investigate the predicting factors for semantic annotation difficulty.
- **How does sense ambiguity affect automatic suggestions quality?** We expect the quality of automatic suggestions based solely on the surface lemma to decline as the number of senses per lemma increases.
- **What type of automatic annotation suggestions is more beneficial?** When should we prefer suggestions derived from a smaller, expertly annotated corpus, and when those based on surface-lemma patterns from a much larger raw corpus?
- **Does the number of suggestions improve automatic suggestions quality?** We investigate whether increasing the number of candidate suggestions leads to measurable gains in quality.

Our experiments reveal that semantic ontology annotation is challenging; however, our analysis

[1] The lemmas in the large corpus have been derived from the raw text by an automatic POS tagger and lemmatizer UDPipe (Straka and Straková, 2017).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7804–7815
11–16 May 2026.

demonstrates that automatic suggestions significantly improve both inter-annotator agreement and annotators' alignment with the expert control sample manually annotated by the annotation lead.

Furthermore, we find that knowledge-based suggestions extracted from a carefully curated, albeit smaller, corpus of verbal senses are superior to automatic suggestions based on surface lemmas harvested from a vastly larger raw corpus. This advantage arises because support derived solely from text naturally declines as semantic ambiguity increases.

2. Background

The tasks described in this work involve annotation aimed at extending a semantic ontology in which the classes are semantic notions of (types or concepts of) events or states. Before describing the case studies in detail, we provide concise background information on the original resource.

2.1. SynSemClass Resource

The semantic annotation evaluation described here is used for extending the SynSemClass (SSC) hierarchical, eventive concept ontology. The ontology links its entries called "classes" to several existing lexical resources with similar goals. Each class corresponds to a "concept" of an event or state, and is populated with words possibly expressing the concept (in several languages), effectively forming a "multilingual synonym class" (Urešová et al., 2025). A SSC class can be exemplified by the class *behave* (Fig. 1).

The SynSemClass ontology further links the semantic behavior of classes (concepts) with the syntactic (valency) structure of the words associated with the class. The relation is captured by mapping the *semantic roles* that are part of every class description to the various valency frames[2] as defined for the individual words (verbs) in the class. In this regard, the SynSemClass classes with semantic roles are a more "semantic" type of "frames", generalizing over the more (morpho-)syntactically-oriented valency frames.

The SynSemClass classes are annotated in Czech, English, German, and Spanish. This case study describes the work on extending the Czech part.

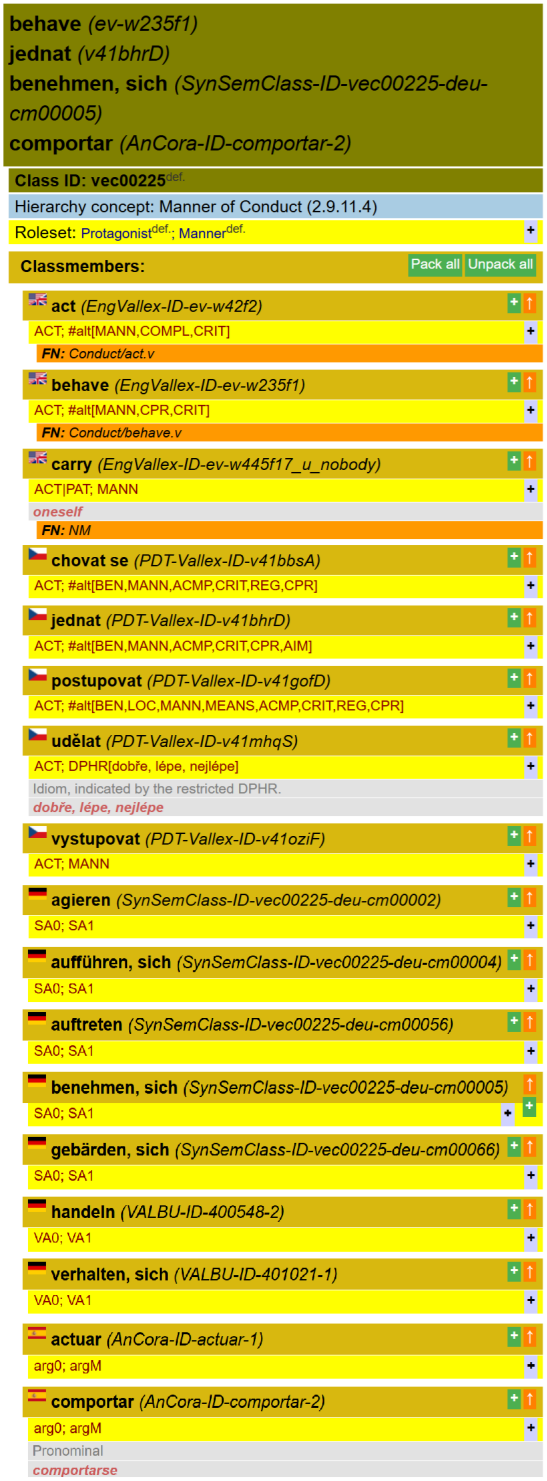

Figure 1: Simplified example of the class *behave*

2.2. The Annotation Process

The annotation process for the enrichment of SynSemClass ontology consists of three main stages:

- **A. Preparation** Unprocessed valency frames from PDT-Vallex 4.5 (Urešová et al., 2024) are selected and automatic suggestions are generated (or no suggestions are applied for Task 1). These suggestions are produced using corpus-based methods or language models,

[2]Each valency frame usually presents one verbal sense. Accordingly, the terms frame and sense are treated as synonymous in this context.

depending on the task setup (see Sect. 4).

- **B. The First Annotation Phase** (covering the tasks described in this paper). The prepared data are presented to the annotators whose task is, using their linguistically and cognitively based judgment:

  1. *To select (assign) one of the suggested classes, propose some other class of the SSC 5.5 classes, or indicate that the suitable class is not present in the ontology yet*. This is considered the key annotation step.

  2. *To specify possible restrictions or notes* for the selected classes, where applicable.

  3. *To assign the appropriate hierarchical concept* under which the new class (for the valency frames without any suitable synonymous class in the current SynSemClass version) should be added (typically a new sub-concept for a more specific subset of an existing concept must be created).

- **C. The Second Annotation Phase (frame annotations)** is not part of the present study; therefore, this stage is not described here.

3. Production of Suggestions

We follow the pre-annotation approach by Straková et al. (2023) who fine-tuned a multilingual language model for extending one of the previous versions of SynSemClass (Uresova et al., 2022).

Our pre-annotation model is a fine-tuned multilingual RemBERT model (Chung et al., 2021) with 559M trainable parameters. The objective of the model is to estimate the probability for an input surface lemma in a sentence context to belong to each of the 1511 SynSemClass classes. For training the model, we used the examples of verbs assigned to classes from the already annotated SynSemClass sentences (Urešová et al., 2025). We adopted the training hyperparameters from Straková et al. (2023).

During inference, the model harvests new example sentences containing the new verbs from a corpus and estimates the probability of the SynSemClass classes. The top K ($K = 5$) predicted classes are then presented to the annotators.

For inference, we used two sources of input sentences: For the lemma-based suggestions, we collected new example sentences representing the surface lemmas corresponding to the annotated valency frames from the SYN v4 (Hnátková et al., 2014), (Křen et al., 2016). As a lemma can represent a manifestation of several valency frames, we expect this approach to produce weaker suggestions especially for lemmas with many valency frames. Therefore, we also employed the PDT-C 2.0 (Mikulová et al., 2026) (Hajič et al., 2024). The verbs in this corpus are manually annotated with their valency frames, which allows for exact targeting of the example sentences for each input frame.

4. Annotation Process Details

In the annotation, we worked with three groups of words (more precisely, with word senses according to the PDT-Vallex version 4.5 valency lexicon) not found in the SynSemClass ontology. In general, the groups differed among each other along two main axes: semantic ambiguity, determined by the number of valency frames per word (its lemma), and the presence or absence of the automatic class suggestions. Highly ambiguous words (lemmas) have been defined as those with three or more valency frames (corresponding roughly to the word's senses) in the PDT-Vallex 4.5 valency lexicon. For each group, we have devised a task:

- **Task 1: High-Amb Without Suggestions** – In this task, we processed valency frames that fall under a highly ambiguous lemma (with an average of 17.85 frames per lemma). They were processed entirely manually without the use of any automatic suggestions. The reason for such an approach was that these frames were not covered by the manually curated corpus PDT-C 2.0, having thus no example sentences for determining the context needed for generating the suggestions. For the comparative evaluation statistics, we excluded erroneous PDT-Vallex 4.5 frames and frames with multiple meanings.[3]

- **Task 2: Low+High-Amb With Suggestions** – In this task, we processed the valency frames covered in sentence examples from PDT-C 2.0, which allowed us to provide annotators with annotation suggestions. The following setup was applied: a) for frames with a highly ambiguous lemma (avg. 17.04 frames per lemma), the annotators received suggestions provided by a fine-tuned LLM based solely on PDT-C 2.0 sentence examples; b) for the frames belonging to low-ambiguity lemmas (avg. 1.31 frames per lemma), the annotators received not only suggestions provided by a fine-tuned

[3] For Light Verb Constructions (LVCs), i.e., frames containing a CPHR slot with multiple words, whose meanings determined the assignment to SynSemClass classes. For example, the LVC "mít pocit, představu, vzpomínku" (to have a feeling, an idea, a memory) is formally represented as one frame - "ACT (1) CPHR ({pocit, představa, vzpomínka, . . . }.4)," representing three meanings of "cítit" (to feel), "představovat si" (to imagine), and "vzpomínat" (to remember). This would require more complicated evaluation measure(s).

Ambiguity	Task 1 High	Task 2 High	Task 2 Low	Task 3 Low
Annotated	463	614	380	354
Shared	441	68	99	349
Annotators	6	4	4	4

Table 1: Number of frames annotated by at least one annotator, number of shared frames (valid for the study), and number of annotators in the shared annotations, across all tasks.

LLM based on PDT-C 2.0, as well as suggestions based on the surface lemmas taken from SYN v4 (Hnátková et al., 2014).[4] For the evaluation, only a small sample of annotations processed by multiple annotators was taken into consideration.

- **Task 3: Low-Amb With Suggestions** – In this task, we processed low-ambiguity valency frames (avg. 1.51 frames per lemma) without examples in PDT-C 2.0. The annotators thus only received LLM suggestions based on the lemmas taken from SYN v4, rather than frame-based suggestions, due to the absence of contextual examples.

Our annotators are undergraduate students and junior researchers in the field of theoretical linguistics. Our chief annotator is an internal senior linguistic expert. All annotators are native Czech speakers and have received prior training on the same task (using data obtained through a different pre-processing method); therefore, no additional training was required. The annotators were compensated on an hourly basis.

The final number of annotated frames is presented in Table 1.

4.1. Task 1: High-Amb Without Suggestions

The concrete objective of Task 1 was to manually assign each valency frame (word sense) to the appropriate, existing SSC class, or, if no such suitable class was deemed to exist, to propose a new class and indicate its corresponding hierarchical concept (HIC).[5]

As already stated, no automatic suggestions have been presented to the annotators, given the absence of PDT-C 2.0 examples for the selected frames.

The annotation materials were distributed to the annotators as an Excel table whose header included a description of the individual columns as follows:

- **Column A** contained a frame from PDT-Vallex 4.5.
- **Column B** contained the ID of the verb (lemma).
- **Column C** displayed the lemma.
- **Column D** indicated how many frames the lemma had in PDT-Vallex 4.5.
- **Column E** showed how many of the frames in Column D showed had not yet been processed in SSC.
- **Columns F, H, and J** were reserved for the class/classes to which the frame was assigned.
- **Column L** recorded an alternative expression of the frame's meaning (a near-synonym), ideally as a single word but possibly as multiple words for clarity.
- **Column M** was designated for the hierarchical concept.
- **Column N** was reserved for annotators' notes.

Columns B, C, D, and E were included for internal processing and did not require input from the annotators.

The annotation process consisted of the following steps:

1. **Displaying the frame.**
2. **Understanding the meaning.** The annotators were instructed to consider the meaning of the verb or idiomatic expression represented by the frame and find the closest similar verbs already annotated in SynSemClass. They were using the SSC search tool (Petliak et al., 2024) and the tool for the SynSemClass annotation (SynEd) (Fučíková et al., 2023). In case of an idiomatic expression, this was recorded in column L. For non-idiomatic expressions, the annotators proceeded to the next step.
3. **Assignment of existing classes.** If one or more classes matched the annotated frame, they were written in descending order of applicability to columns F, H, and J.

[4] SYN v4 (Křen et al., 2016) is part of the SYN series of synchronic corpora of written Czech compiled within the framework of the Czech National Corpus project (CNC). The CNC is an academic project founded in 1994 at the Faculty of Arts, Charles University, Prague, and maintained by the Institute of the Czech National Corpus (https://korpus.cz). It systematically documents Czech and related languages, providing free access to registered users interested in authentic language usage.

[5] For further details on the hierarchy of classes (HICs) in SynSemClass, see Urešová et al. (2025). In short, the classes are organized in a hierarchy tree using "more general"/"more specialized" relation between the HICs.

4. **Proposal of a new class.** If no suitable class was found in SSC, “x” was entered in column F, and instead the annotators attempted to specify the HIC (in column M) under which the new class should belong.
5. **Adding notes.** Column N was reserved for annotation notes.

4.2. Task 2: Low+High-Amb With Suggestions

In this task, valency frames with representative sentences in PDT-C 2.0 were processed. All valency frames received the frame-based suggestions, and only the valency frames with a low-ambiguity lemma (at most two senses) received also the lemma-based suggestions.

For the second task, the annotation materials were distributed to the annotators again as an Excel table, where two rows were included: one with five frame-based class suggestions and one with lemma-based class suggestions (empty for high-ambiguity lemmas).

The first row of the file contained the following column headers:

- **Column A**: lemma,
- **Column B**: frame,
- **Column C**: indication whether the suggestions are frame-based (F) or lemma-based (L),
- **Column D**: indication of the prioritized processing - For lemmas with a single valency frame in PDT-Vallex 4.5, the lemma-based suggestions are prioritized (marked as "1"), as they originate from a larger corpus and offer broader coverage. In contrast, for lemmas with multiple valency frames, the precedence is given to frame-based suggestions, since they were derived from a more precise and semantically disambiguated corpus.
- **Column E**: number of frame/lemma occurrences found in PDT-C 2.0/SYN v4, and
- **Column F**: the highest-ranked suggestion from L or F.

The remaining columns were completed by the annotators for each lemma/frame suggestions as follows:

- **Rating of suggestions:** Each of the five automatically generated suggestions in the row were manually evaluated and annotated as either accepted (*y* or *r_y*) or rejected (*n* or *r_n*)[6] in the column adjacent to the suggestion.
- **Overall verdict**: In column G, the annotator indicated whether any of the suggestions were applicable (*x*); whether all suggestions were rejected and another existing class was proposed (*class ID*); or whether a new class needed to be created for the concept (*0*).

[6]The r_ prefix means “rather” (yes or no), or a “weak” accept or reject.

4.3. Task 3: Low-Amb With Suggestions

For Task 3, an Excel table with valency frames, having at most two senses per lemma in PDT-Vallex 4.5, has been prepared. This file followed the same format as that used for Task 2 and contained all relevant information for each frame, including the lemma, the frame identifier, and other details. Since these valency frames were not used in PDT-C 2.0 example sentences, only lemma-based suggestions were provided in Task 3. The annotators reviewed the five automatic suggestions and confirmed or adjusted them, similar to Task 2.

5. Results

5.1. Automatic Suggestions Quality

Before carrying out the main analysis, we assessed the quality of the automatic suggestions using information retrieval evaluation measures implemented in `pytrec_eval`, a Python interface to TREC’s evaluation tool (Van Gysel and de Rijke, 2018). Table 2 shows recall and mean average precision (MAP) of the five automatic suggestions offered to the four annotators (upper part) and to the expert annotator (lower part). Each suggestion was explicitly annotated as accepted or rejected by each of the four annotators, allowing them to accept up to five suggestions. As the recall at 5 exceeded 60% in all tasks with suggestions, we proceeded with further analysis. In the gold data annotated by the chief annotator, only one gold class was selected, leading to overall lower recall and MAP scores of the automatic suggestions. Task 1 was annotated in the most complicated setting (high ambiguity, without presented suggestions) and as these verbs were not covered in PDT-C 2.0, only lemma-based suggestions were generated and evaluated ex-post (marked with †). We discuss the challenges of lemma-based suggestions in the high-ambiguity setting in Section 5.3.

5.2. The Effect of Frame-Based Suggestions on Annotation Quality

Table 3 shows inter-annotator agreement (IAA, Fleiss’ Kappa) between annotators in both low- and high-ambiguity annotation settings, with and without suggestions. The IAA for Task 1 annotated without automatic support is modest, at 39.73. This

Ambiguity Suggestions Type	Task 1 High Without L	Task 2 High With F	Task 2 Low With L+F	Task 3 Low With L
Suggestions vs. annotations (up to 5 selected)				
Recall	26.44†	62.10	67.11	61.66
MAP	18.04†	50.47	45.16	44.77
Suggestions vs. gold (1 selected)				
Recall	23.64†	41.94	42.86	40.24
MAP	16.59†	32.85	28.79	27.57

Table 2: Recall and Mean Average Precision (MAP) at rank 5 of automatic suggestions quality for low- and high- semantic ambiguity conditions, measured against four annotators (upper) and gold data by expert annotator (lower). "F" stands for frame-based suggestions from PDT-C 2.0, "L" for lemma-based suggestions from SYN v4. † marks annotation without suggestions evaluated against suggestions generated ex-post.

Amb. Suggest.	Task 1 High None	Task 2 High F	Task 2 Low L+F	Task 3 Low L
IAA	39.73	47.88	57.70	49.15
Acc.	54.80	64.34	71.72	66.33

Table 3: Inter-annotator agreement (Fleiss' Kappa) and Accuracy w.r.t. gold data annotated by annotation lead on the first class choice.

is unsurprising, since annotators must correctly assign one of the 1511 semantic ontology classes or decide that a new concept should be created.

The crucial question is whether and how the semantic annotation benefits from the addition of automatic suggestions. Our first experiment therefore starts with an addition of highly curated, putatively precise suggestions from manually annotated frame corpus. A comparison between a high-ambiguity Task 1 annotated without suggestions and similarly highly ambiguous Task 2 annotated with the frame-based suggestions shows an increase of the IAA on the first class choice from 39.73 to 47.88.[7]

However, inter-annotator agreement can be problematic, since multiple annotators may produce divergent interpretations of the same annotation. Therefore, we believe that manual evaluation constitutes a central element for the evaluation of semantics, and we compared the same two sets against the gold annotations by a chief annotator. In comparison against the expert annotations, the annotators improved in accuracy on first class choice from 54.80 to 64.34.[8]

[7] statistically significant at $\alpha = 0.05$ with $p = 0.001$ using Welch's two-sample t-test

[8] statistically significant at $\alpha = 0.05$ with $p = 0.029$ using Welch's two-sample t-test

5.3. Challenges of Lemma-Based Suggestions

A highly curated, manually annotated resource with valency frames is rare and may not cover all cases. We now turn our attention to suggestions based on surface lemma extracted from a large unannotated corpus.

An immediate caveat is that suggestions based on surface lemmas inevitably conflate distinct semantic senses, an effect likely to be more pronounced under highly ambiguous experimental conditions. We hypothesize that frame-based suggestions will disambiguate better than lemma-based ones, but, on the other hand, the frame-based suggestions might be less robust as they are harvested from a limited resource[9] compared to a larger unannotated corpus.[10]

To evaluate the severity of the expected trend, we had annotators manually accept or reject five suggestions in Task 2 and Task 3. For Task 1 annotated without suggestions at the time of annotation, we generated the suggestions on the side and evaluated them against the independently chosen classes ex-post. Indeed, Figure 2 clearly demonstrates almost linear decline of quality of annotations based on surface lemmas with increasing ambiguity (orange and blue line), while, on the contrary, the quality of frame-based suggestions holds (green line). This happens despite the vast difference in sizes between the resources.

Another piece of supporting evidence of the frame-based suggestions superiority is the increase of IAA between the Task 3 annotated with supplied lemma-based suggestions and the Task 2 annotated with both frame-based and lemma-

[9] 175k sentences with 367k annotated frames of the PDT-C 2.0 tectogrammatical layer

[10] first 3.35M sentences of SYN v4

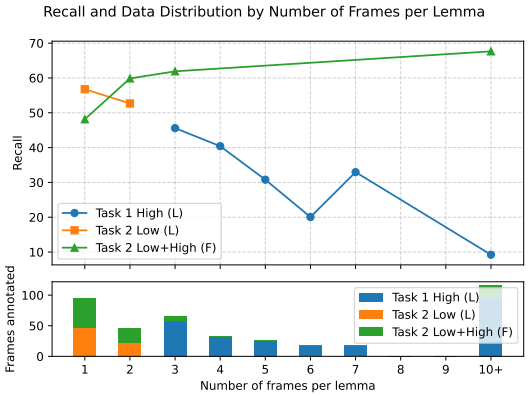

Figure 2: Recall at rank 5 for automatic suggestions, shown as a function of semantic ambiguity. Categories with fewer than 7 instances were excluded due to high variance.

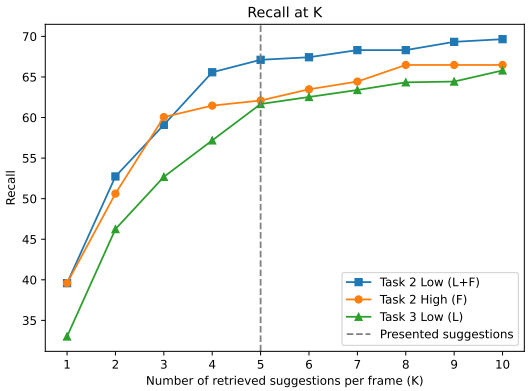

Figure 3: Recall with respect to number of retrieved automatic annotation suggestions (K). Annotations for $K > 5$ are only partial.

based suggestions from 49.15 to 57.70 (Fleiss' Kappa), and the corresponding increase in accuracy against the expert annotator's gold data from 66.33 to 71.72.

Despite the expected decrease in the quality of lemma-based suggestions compared to frame-based ones, we still consider the lemma-based suggestions to be useful. In future work, we plan to examine the annotation process with lemma-based suggestions in a low-ambiguity domain and compare it to annotation without any suggestions in the same setting. Furthermore, it will be valuable to directly compare Task 1 (no suggestions), Task 2 (frame-based suggestions), and a new task (lemma-based suggestions) in a high-ambiguity environment. Such an experiment is becoming increasingly relevant as annotation extends to languages that lack resources containing example sentences with manually annotated frames or with disambiguated word senses.

5.4. Suggestion Quantity and Recall

We are also interested in the optimal number of suggestions presented to the annotators, studying how this number affects the recall of the correct ontology class. Figure 3 shows an increasing trend in recall as the number of retrieved suggestions grows. Note that annotators fully annotated exactly $K = 5$ presented suggestions. For $K > 5$, we adopt a strict approach for partial information retrieval metrics: a class is considered positive only if it was explicitly selected by an annotator (i.e., the annotator rejected all five presented classes and freely suggested a new class, consistent with a retrieved class that was not initially presented). All other classes are considered negative, even though an annotator might have selected them if they had been presented.

Figure 3 further shows that recall begins to plateau before $K = 5$, indicating that the chosen cutoff adequately captures most of the achievable recall. Additional suggestions beyond this point are likely to yield only marginal improvements while increasing the annotation effort.

6. Related Work

The work described here relates closely to various lexical resources that include similar information as SynSemClass, namely, information on syntactic and semantic patterns of English verbs and information on hierarchical relations among their concepts. Examples of such sources are WordNet (WordNet, 2010), (Fellbaum, 1998), EuroWordNet (EuroWordNet Consortium, 1998), (Pianta et al., 2002a; Ellman, 2003), VerbNet (Schuler et al., 2023), (Schuler, 2006), or FrameNet (Baker et al., 1998). Also, various ontologies of events, such as BabelNet (Navigli and Ponzetto, 2010), or The Rich Event Ontology (Brown et al., 2017). As far as we are aware, none of these have used LLM-based preprocessing either for their initial creation (understandingly, given the timeframe) or for their extension.

Recent studies increasingly explore the interaction between ontologies and LLMs to improve knowledge representation, alignment, and ontology engineering. LLMs have been used successfully to support ontology matching and alignment tasks, combining structural and lexical information with retrieval-augmented or self-training approaches. LLMs have been successfully applied to ontology matching by combining lexical and structural information and using retrieval-augmented or self-training approaches.

Early works such as OLaLa (Hertling and Paulheim, 2023) demonstrated the feasibility of zero-

shot and few-shot prompting for ontology alignment. More recent systems, including Giglou et al. (2024), Song et al. (2025), and Nguyen et al. (2025), integrate context generation and retrieval-based techniques to improve semantic matching. Hybrid approaches such as Giglou et al. (2025) further combine LLMs with knowledge graph embeddings, illustrating how symbolic and neural representations can complement each other in ontology engineering. Except for the last one, none of the works have been used in a similar context to ours, and none for Czech.

Mikulová et al. (2022) evaluates the effect of automatic parser and/or linguistically-based (rule-formulated) checks on the same data available to the annotators, and their influence on annotation quality and efficiency. This experiment (on Czech language) confirmed that the pre-annotation is an efficient tool for faster manual syntactic annotation which increases the consistency of the resulting annotation without reducing its quality.

Straková et al. (2023) investigated the use of fine-tuned language models for pre-annotating data by adding descriptive verbs to SynSemClass 4.0 (Uresova et al., 2022). Building on their approach to generating automatic pre-annotations from surface lemmas, we extend the method by incorporating knowledge-driven pre-annotations based on a small, carefully curated corpus of verbal valency frames. This setup enables a direct comparison of the usefulness of data-driven and knowledge-driven sources of pre-annotations. While the main outcome of Straková et al. (2023) was a confirmed correlation between automatic scores and human annotations, our work goes further by directly comparing inter-annotator agreement (IAA) and accuracy against gold-standard data. We also reveal the limitations of lemma-based suggestions in settings with high ambiguity.

7. Conclusions

We extended the SynSemClass ontology with 1,811 words (more precisely, valency frames corresponding to those words' senses) in three annotation tasks differing in the degree of ambiguity and the type of automatic suggestions. At least four annotators worked on a shared subset of 957 frames, which enabled an analysis of the impact of automatic annotation support. We conclude that:

- Both inter-annotator agreement and accuracy (measured against the chief annotator's gold data) increased statistically significantly when using knowledge-driven, frame-based suggestions derived from corpus sentences with manually annotated valency frames.
- Suggestions based on surface lemmas in sentence contexts pose challenges, despite being sourced from a much larger corpus. We showed that their quality declines as ambiguity increases, in contrast to the stable performance of frame-based suggestions.
- The five annotation suggestions presented represent a satisfactory balance between recall and annotation cost.

Building on previous findings and the preliminary lemma-based experiments, we assume that lemma-based suggestions still provide benefits for low-ambiguity verbs with only one or two possible meanings. Future work should therefore include a direct comparison between annotation with and without lemma-based suggestions for such verbs.

The SynSemClass project description including its history and previous releases is available at https://ufal.mff.cuni.cz/synsemclass, with a browser for the latest SynSemClass 5.5 version available at https://lindat.mff.cuni.cz/services/SynSemClass55.

For full replicability, the code and data used for this analysis are available on GitHub at https://github.com/ufal/SynSemClassLREC2026, and as a snapshot of the repository hosted by the LINDAT/CLARIAH-CZ LRI at http://hdl.handle.net/11234/1-6112.

8. Acknowledgements

Jana Straková was supported by the Johannes Amos Comenius Programme (P JAC) project No. CZ.02.01.01/00/22_008/0004605, Natural and anthropogenic georisks, funded by the Ministry of Education, Youth and Sports of the Czech Republic (MEYS CR), and the remaining authors by the project Uniform Meaning Representation (UMR), Project No LUAUS23283, also funded by MEYS CR. The work described herein uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (projects LM2018101 and LM2023062, funded by MEYS CR). We also acknowledge the use of the corpora created, maintained, and made available by the Czech National Corpus (project No. LM2023044, funded also by MEYS CR).

We thank our annotators for their linguistic expertise and diligent work in developing this ontology.

9. Ethics Statement

All resources and data used in this study originate from publicly available linguistic corpora and lexical databases (such as PDT-Vallex 4.5, SynSemClass 5.5 or PDT-C 2.0). Data annotation was performed by paid annotators who gave their informed consent prior to participation. All annotators were compensated fairly for their time and effort. No personally identifiable or sensitive information was

collected or used in this study. The procedures followed institutional and professional ethical standards. The resulting materials have been released for research purposes under an open license (via the LINDAT/CLARIAH-CZ repository – see the last two paragraphs in Sect. 7).

10. Limitations

The greatest limitation of this study lies in the design of the annotation tasks, which were optimized primarily to streamline the annotation process. This resulted in an incremental experimental setup, progressing from Task 1 through Task 2 to Task 3. While this design enabled clear and direct comparisons between annotations without automatic support and those assisted by frame-based suggestions in a high-ambiguity environment, it also introduced certain constraints. In particular, it limited our ability to directly evaluate lemma-based suggestions in low-ambiguity contexts. Nevertheless, the experiments provided valuable insights into the challenges of applying lemma-based suggestions under high ambiguity and facilitated a comparison with frame-based suggestions under the same conditions.

Furthermore, this study was limited to Czech data only. Future research should examine whether the findings generalize to other languages.

11. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M. Sanfilippo, and Laure Vieu. 2022. DOLCE: A descriptive ontology for linguistic and cognitive engineering1. *Applied Ontology*, 17(1):45–69.

Susan Brown, Claire Bonial, Leo Obrst, and Martha Palmer. 2017. The Rich Event Ontology. In *Proceedings of the Events and Stories in the News Workshop*, pages 87–97, Vancouver, Canada. Association for Computational Linguistics.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Silvie Cinková. 2006. From PropBank to EngVallex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2170–2175, Genova, Italy. ELRA.

Jeremy Ellman. 2003. Eurowordnet: A multilingual database with lexical semantic networks: Edited by Piek Vossen. Kluwer Academic Publishers. 1998. isbn 0792352955, 179 pages. *Natural Language Engineering*, 9:427 – 430.

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

Christiane Fellbaum. 2005. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 2–665. Elsevier.

Christiane Fellbaum and Piek Vossen. 2007. Connecting the universal to the specific: Towards the global grid. In *Intercultural Collaboration*, pages 1–16, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Corpus-based multilingual event-type ontology: annotation tools and principles. In *Proceedings of the 21st International Workshop on*

Treebanks and Linguistic Theories, pages 1–10, Washington, D.C., USA. Association for Computational Linguistics, Association for Computational Linguistics.

Eva Fučíková, Cristina Fernández-Alcaina, Jan Hajič, and Zdeňka Urešová. 2024. Textual coverage of eventive entries in lexical semantic resources. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15835–15841, Torino, Italy. European Language Resources Association.

Hamed Babaei Giglou, Jennifer D'Souza, Sören Auer, and Mahsa Sanaei. 2025. Ontoaligner meets knowledge graph embedding aligners. *arXiv preprint*. ArXiv:2509.26417.

Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. 2024. Llms4om: Matching ontologies with large language models. In *European Semantic Web Conference (ESWC) 2024*. ArXiv:2404.10317.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Yuan He, Jiaoyan Chen, Hang Dong, and Ian Horrocks. 2023. Exploring large language models for ontology alignment. In *ISWC 2023, Posters & Demos*. ArXiv:2309.07172.

Sven Hertling and Heiko Paulheim. 2023. Olala: Ontology matching with large language models. In *K-CAP 2023*.

Annika Hinze, Ralf Heese, Alexa Schlegel, and Adrian Paschke. 2019. Manual semantic annotations: User evaluation of interface and interaction designs. *Journal of Web Semantics*, 58:100516.

Milena Hnátková, Michal Křen, Pavel Procházka, and Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164, Reykjavik, Iceland. European Language Resources Association (ELRA).

Svetlozara Leseva and Ivelina Stoyanova. 2022. *Stative verbs: Conceptual structure, hierarchy, systematic relations*, pages 68–114. Prof. Marin Drinov Publishing House of BAS.

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

Jiří Materna. 2014 [cit. 2024-11-14]. *Probabilistic Semantic Frames [online]*. Doctoral theses, dissertations, Masaryk University, Faculty of Informatics, Brno. SUPERVISOR : Karel Pala.

Marie Mikulová, Jiří Mírovský, Milan Straka, Pavlína Synková, Barbora Štěpánková, Jan Štěpánek, and Jan Hajič. 2026. Prague dependency treebank - consolidated 2.0: Enriching a complex annotation scheme. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. 2022. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2909–2918, Marseille, France. European Language Resources Association.

George Miller and Christiane Fellbaum. 2007. Wordnet then and now. *Language Resources and Evaluation*, 41:209–214.

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

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Lam Nguyen, Erika Barcelos, Roger French, and Yinghui Wu. 2025. Kroma: Ontology matching with knowledge retrieval and large language models. *arXiv preprint*. ArXiv:2507.14032.

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/CLARIN digital library. http://hdl.handle.net/11858/00-097C-0000-0001-4880-3.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Nataliia Petliak, Cristina Fernandéz Alcaina, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2024. Search tool for an event-type ontology. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 66–70, Torino, Italia. ELRA and ICCL.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002a. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002b. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended theory and practice. *Unpublished Manuscript*.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Yiping Song, Jiaoyan Chen, and Renate A. Schmidt. 2025. Genom: Ontology matching with description generation and large language model. *arXiv preprint*. ArXiv:2508.10703.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Jana Straková, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Extending an event-type ontology: Adding verbs and classes using fine-tuned LLMs suggestions. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 85–95, Toronto, Canada. Association for Computational Linguistics.

Dan Tufis, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, methods, results and perspectives – A general overview. *Romanian Journal of Information Science and Technology Special Issue*, 7:9–43.

Zdeňka Urešová, Cristina Fernández-Alcaina, Eva Fučíková, and Jan Hajič. 2023a. SynSemClass Czech and English Annotation Guidelines. Technical Report 73, UFAL MFF UK.

Zdeňka Urešová, Eva Fučíková, Cristina Fernández Alcaina, and Jan Hajič. 2023b. Synsemclass 5.0. available from the lindat/clariah-cz digital repository. http://hdl.handle.net/11234/1-5230.

Zdeňka Urešová, Eva Fučíková, and Jan Hajič. 2025. Creating hierarchical relations in a multilingual event-type ontology. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 240–249, Vienna, Austria. Association for Computational Linguistics.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2019. Parallel Dependency Treebank Annotated with Interlinked Verbal Synonym Classes and Roles. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 38–50, Paris, France. Université Paris Sorbonne Nouvelle, Association for Computational Linguistics.

Zdenka Uresova, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, and Jan Hajic. 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.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *SIGIR*. ACM.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer.

Piek Vossen. 2002. Wordnet, EuroWordNet and Global WordNet. *Revue Française de Linguistique Appliquée*, 7.

Piek Vossen, Laura Bloksma, and Horacio Rodriquez. 1998. The EuroWordNet Base Concepts and Top Ontology. Workingpaper, Vrije Universiteit.

12. Language Resource References

EuroWordNet Consortium. 1998. *EuroWordNet*. Global WordNet Association. Available at: https://globalwordnet.org/resources/euro-wordnet/.

Jan Hajič and Eduard Bejček and Alevtina Bémová and Eva Buráňová and Eva Fučíková and Eva Hajičová and Jiří Havelka and Jaroslava Hlaváčová and Petr Homola and Pavel Ircing and Jiří Kárník and Václava Kettnerová and Natalia Klyueva and Veronika Kolářová and Lucie Kučová and Markéta Lopatková and David Mareček and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Michal Novák and Petr Pajas

and Jarmila Panevová and Nino Peterek and Lucie Poláková and Martin Popel and Jan Popelka and Jan Romportl and Magdaléna Rysová and Jiří Semecký and Petr Sgall and Johanka Spoustová and Milan Straka and Pavel Straňák and Pavlína Synková and Magda Ševčíková and Jana Šindlerová and Jan Štěpánek and Barbora Štěpánková and Josef Toman and Zdeňka Urešová and Barbora Vidová Hladká and Daniel Zeman and Šárka Zikánová and Zdeněk Žabokrtský. 2024. *Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0)*. Institute of Formal and Applied Linguistics, Charles University. LINDAT/CLARIAH-CZ digital library.

Křen, Michal and Cvrček, Václav and Čapka, Tomáš and Čermáková, Anna and Hnátková, Milena and Chlumská, Lucie and Jelínek, Tomáš and Kováříková, Dominika and Petkevič, Vladimír and Procházka, Pavel and Skoumalová, Hana and Škrabal, Michal and Truneček, Petr and Vondřička, Pavel and Zasina, Adrian. 2016. *SYN v4: large corpus of written Czech*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Schuler, Karin Kipper and Brown, Susan and Stabler, Julia and Palmer, Martha. 2023. *VerbNet*. University of Colorado Boulder. A broad-coverage verb lexicon providing syntactic and semantic information for English verbs.

Urešová, Zdeňka and Alcaina, Cristina Fernández and Bourgonje, Peter and Fučíková, Eva and Hajič, Jan and Hajičová, Eva and Kolářová, Veronika and Rehm, Georg and Rysová, Kateřina and Zaczynska, Karolina. 2025. *SynSemClass 5.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Urešová, Zdeňka and Bémová, Alevtina and Fučíková, Eva and Hajič, Jan and Kolářová, Veronika and Mikulová, Marie and Pajas, Petr and Panevová, Jarmila and Štěpánek, Jan. 2024. *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.5 (PDT-Vallex 4.5)*. LINDAT/CLARIAH-CZ digital library at Institute of Formal and Applied Linguistics, Charles University.

WordNet. 2010. *WordNet: A Lexical Database for English*. Princeton University. Available at: https://wordnet.princeton.edu/.