Label Bias in Symbolic Representation of Meaning

Marie Mikulová and **Jan Štěpánek** and **Jan Hajič**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{mikulova,stepanek,hajic}@ufal.mff.cuni.cz

Abstract

This paper contributes to the trend of building semantic representations and exploring the relations between a language and the world it represents. We analyse alternative approaches to semantic representation, focusing on methodology of determining meaning categories, their arrangement and granularity, and annotation consistency and reliability. Using the task of semantic classification of circumstantial meanings within the Prague Dependency Treebank framework, we present our principles for analyzing meaning categories. Compared with the discussed projects, the unique aspect of our approach is its focus on how a language, in its structure, reflects reality. We employ a two-level classification: a higher, coarse-grained set of general semantic concepts (defined by questions: where, how, why, etc.) and a fine-grained set of circumstantial meanings based on data-driven analysis, reflecting meanings fixed in the language. We highlight that the inherent vagueness of linguistic meaning is crucial for capturing the limitless variety of the world but it can lead to label biases in datasets. Therefore, besides semantically clear categories, we also use fuzzy meaning categories. We support this position with a brief annotation experiment.

1 Motivation

Natural language is a very powerful way of describing the world. Communication using natural language is remarkably efficient because it allows the use of a finite grammar and lexicon to describe a potentially infinite set of situations, knowledge, emotions (i.e. *content*, as we will simplistically refer to the communicated reality in this paper). The means of language have many meanings. The meanings expressed may be relatively vague in relation to the content being described. The properties of natural language, such as ambiguity or vagueness, therefore pose challenging problems for symbolic representations of meaning.

The research question we tackle in this contribution can be illustrated by the examples (1)–(7).

(1) John worked **quickly**.
(2) John worked **with a chisel**.
(3) John worked **with a wood**.
(4) John worked **with a colleague**.
(5) John worked **with / without a smile**.
(6) **With his skills** John worked **with success**.
(7) John worked **behind the house**.

How can we describe the meanings of the highlighted expressions in examples (1)–(7)? One may simply state that, in all examples, some circumstance of John's working is expressed and to use one very coarse-grained category "circumstance" for all expressions (cf. a single label *Adverbial* in the Universal Conceptual Cognitive Annotation project (Abend and Rappoport, 2013)). However, it is clear that the circumstance in (7) is semantically considerably distinct from the circumstances expressed in (1)–(6). It seems that a finer distinction into spatial and, let's say, "broad manner-related" circumstances would be more appropriate. But it is also evident that the circumstances in (1)–(6) differ. Some more significantly, some less so. Are *to work with a chisel* and *with wood* the same semantic category? Should a semantic classification distinguish between *with a smile* and *without a smile*? The question posed in this paper is: what granularity should semantic classification have, and, more importantly, what should determine this granularity? This also raises a question for linguistic annotation: On how fine-grained categories can human annotators agree?

2 Introduction

Meaning representation has long been an important task in computational linguistics, yet it remains challenging for both machines and human annotators. New or extended symbolic representations of meaning are continuously being proposed (e.g., Uniform Meaning Representation (UMR; Van Gy-

sel et al., 2021), Abstract Meaning Representation (AMR; Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA; Abend and Rappoport, 2013), Deep Universal Dependencies (Droganova and Zeman, 2019), Parallel Meaning Bank (Abzianidze et al., 2017)).

Meaning representation (semantic role labelling, word sense disambiguation) is typically modelled by means of a dictionary or pre-defined set of meaning categories, and a meaning is then captured through the best-fitting label from this set. Most of these approaches have a primary focus on verbs with varying degree of elaborate classification of the verb participant semantic roles (e.g., VerbNet (Kipper et al., 2008), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), PDT-Vallex (Urešová et al., 2024b), SynSemClass (Urešová et al., 2024a)), and there are also broader databases for word senses in general, such as WordNet (Miller, 1995), OntoNotes (Hovy et al., 2006).

Relatively few frameworks have focused on comprehensive accounts of non-participant (adjuncts, adverbials, circumstants) roles, though they are very frequent and contribute crucial semantics to sentences. In this respect, we have to mention the Xposition project (or SNACS - Semantic Network of Adposition and Case Supersenses; Schneider et al., 2018; Gessler et al., 2022), which focuses on the semantics of prepositions and it is relatively close to our project. In this project, 52 so-called supersenses are distinguished and organized into a multi-level hierarchy. At the highest level, circumstances, participants, and configurations (noun attributes) are differentiated. The set of labels is partially up to three levels deep, but in terms of expressed meaning, it is relatively coarse-grained.

This contribution aims to critically consider the trend of building semantic representations, highlighting its challenges, and limitations in addressing the following issues in the task of semantic classification of circumstances (outlined in Sect. 1):

(i) The arrangement and granularity of meaning categories, principles upon which a semantic classification can be built to ensure its credibility, explainability, broadness in coverage, and suitability for consistent manual annotation of real texts;

(ii) The relation between language and the world it describes, the boundaries of linguistic meaning and the role of context and knowledge in determining semantic categories for linguistic annotation – arguably one of the most challenging questions in current computational linguistics.

Our semantic classification is developed within the Prague Dependency Treebank (PDT) framework (Hajič et al., 2020). The description of circumstantial meanings is based on a large volume of real examples that PDT corpora provide and the proposal is subsequently used to enrich the semantic annotation in these corpora (for the upcoming release in 2026). We support our approach with a pilot annotation and evaluate the results.

The paper is organized as follows: In Sect. 3, based on the analysis of recent projects dealing with semantic annotation, we discuss key points of meaning representation: description models (Sect. 3.1), granularity of semantic roles (Sect. 3.2), and consistency and reliability of annotation (Sect. 3.3). In Sect. 4, we describe our project on the semantic classification of circumstants within the PDT framework, applying these key points. The annotation experiment is presented in Sect. 5. Our position and findings are summarized in Sect. 6. Supportive material is provided in Appendix A.

3 Meaning Representation Key Points

In the semantic representation projects, labels are determined more or less intuitively (often without any apparent underlying theory), which results in varying granularity both within a single classification and across different semantic representation systems. Different degrees of granularity and (dis)arrangement of categories, as well as their (un)clear definition, influence the reliability and consistency of annotated data. We are aware of the complexity (and unresolvability) of these issues, but we believe that it is important to raise and explore them, seeking guidance toward their solution.

3.1 Linguistic Meaning and what is Beyond

Regarding semantics, questions about the relation between (extra-linguistic) content and linguistic meaning, which have been repeatedly raised in philosophy, logic, and linguistics (Frege, 1892; Saussure, 1916; Wittgenstein, 1953), are now relevant again. In the proposals of semantic representations, the distinction between these two domains is not always clearly made, which leads to unclear principles in the design of the representations. Resolving this issue should be an integral part of defining any semantic representation, especially given its direct implications for portability to other languages.

Languages differ significantly in the meaning categories they express and the formal means they use to do so (cf. Comrie, 1989; Croft, 2003; Haspelmath, 2010 in general; Levinson and Wilkins, 2006 for spatial circumstants). A cross-language semantic representation cannot simply be proposed in the domain of linguistic meaning. However, the representation in the content domain is a task of a completely different nature, mainly in two aspects (cf. Hajičová and Sgall, 1980):

(i) while there is a clear support in the form of analysed language for the representation of linguistic meaning, it is difficult, if not impossible, to find the principles and criteria by which semantic categories in the content domain are determined;

(ii) while a representation of linguistic meaning is one of the levels of the language system, a representation of the content is beyond language itself and is the object of interdisciplinary study.

The language-independent semantic representation has to be approached by trial and error (cf. the development of semantic categories from a complicated multi-layer hierarchy (Schneider et al., 2015) to a simpler hierarchy (Schneider et al., 2018) in the SNACS project) or refined with the incorporation of any new language (cf. interesting comparison of English, Chinese, and Czech in the AMR framework; Xue et al., 2014). The language-independent representation may lead to a small number of very general categories (in UCCA, only one category (*Adverbial*), later 7 (Wang et al., 2021), were established for circumstants), or, on the contrary, to the postulation of more and more subtle structuring (cf. several hundred semantic categories for prepositional phrases in the Preposition Project, Litkowski and Hargraves, 2021). Intuitively designed, language-independent categories vary in granularity even within a single framework. E.g., according to the UMR guidelines (Bonn and et al., 2022), both the circumstants in the sentences *He decorated the room in a creative way* and *Lindbergh crossed the Atlantic in the Spirit of St. Louis* are labelled with the same *Manner* category. In contrast, the circumstant in *I read it in the newspaper* is labelled with the subtle category *Medium*.

We argue that the level of linguistic meaning (the meaning of a sentence is determined by its structure and the meanings of its constituents; cf. also the notion of compositionality (Partee, 2004; Szabó, 2022) or literal meaning (Searle, 1978)) should be considered as starting point for further semantic-pragmatic interpretation of the sentence semantics in which knowledge of the context and general knowledge of the world are applied; cf. ideas postulated in Function Generative Description (Sgall et al., 1986; Sgall, 1995); these questions were reopened by Bender et al., 2015 (cf. also Dinu et al., 2018; Li et al., 2021).[1]

3.2 Arrangement and Granularity

The concept of semantic categories is a widely accepted practice for labelling the meanings of both core and non-core participants. However, as we already mentioned, there is no consensus among linguists on how to define and delimit these categories, which results in considerably diverse set of labels – varying both in quantity and in level of semantic granularity (the verb-oriented projects PropBank, FrameNet, and VerbNet are compared in Petukhova and Bunt, 2008, for an interesting comparative research for prepositional phrases, see O'Hara and Wiebe, 2009).[2]

The repertoire of semantic categories is closely related to their interrelations. Traditionally, semantic categories are organized (if they are organized at all) in a hierarchy (WordNet, FrameNet and partially in OntoNotes and SNACS). In the UMR project, it is proposed to organize semantic categories not through a strict hierarchy, but rather in a lattice-like architecture, in which categories can also divide the semantic space into overlapping domains (Van Gysel et al., 2019).

However, is a hierarchy or lattice a good solution for organizing meanings for the linguistic annotation tasks? The assumption of semantic categories that are mutually disjoint and have clear boundaries

[1] The idea of distinguishing between formally expressed meaning and "real" meaning is also applied in the SNACS project (Hwang et al., 2017): each prepositional phrase is assigned two labels, both selected from the same set of 52 supersenses. One label represents the meaning conveyed by the preposition itself (approximately the domain of linguistic meaning), while the other represents the semantic role that would be expected based on the predicate or the situation (approximately the domain of content).

[2] In SNACS, the set of 52 supersenses is roughly the same granularity as the functors in PDT (cf. Sect. 4; Scivetti and Schneider, 2023). For example, three labels are distinguished within spatial meanings: `GOAL`, `SOURCE`, and the hierarchically superior category `LOCUS`. These categories correspond approximately to the PDT functors `DIR3`, `DIR1`, and `LOC` respectively (see Table 1). The aim of the current project is to achieve a more fine-grained classification within these broad categories. For example, we intend to describe the various locations of the cat in relation to the dog in instances such as this one (taken from the SNACS documentation to illustrate the `LOCUS` category): *The cat is on top of / off / beside / near the dog* via the fine-grained subfunctors `surface`, `outside`, `beside` (cf. Table 4). SNACS's supersenses make no such fine distinctions.

has already been questioned many times (see Kilgarriff, 1997; Hanks, 2000; Tuggy, 1993). While some form of arrangement can serve as a helpful tool, at the same time, it leads to inconsistencies in cases where very different meanings are combined. A lattice structure seems to be more appropriate, but it does not resolve semantically complex cases (e.g., *at his party* is an answer to the questions *When?* and *Where did he laugh?* and merges location and time; the example is from Clematide and Klenner, 2013 study on (coarse-grained) meanings of German prepositions).

We argue that the distinction between the centre of language and its periphery (well known in linguistics throughout its modern development; Daneš, 1966) should also be applied on the semantic level. The meaning disambiguation is either straightforward – making category selection (even fine-grained) clear – or the meaning is more or less complex and vague (where none of the categories fits completely, or more than one fits partially; Mani, 1998; Hanks, 2000; Sgall, 2002; Erk et al., 2013). In such cases, determining the appropriate category is always debatable, regardless of the arrangement approach (none, hierarchy, lattice). Inter-annotator agreement in such instances tends to be low. This notion also matches results in cognitive linguistics: mental categories show "fuzzy boundaries" and different levels of granularity in the course of reasoning (see Rosch, 1975; Hobbs, 1985; Hampton, 2007).

As Sgall (2002) points out, without a certain degree of indistinctness of meaning it would not be possible to capture with limited means the unlimited range of the world we perceive and speak of. The fuzzy meanings are not only a precondition of the natural language universality but also one of its consequences (cf. also Mani, 1998). These properties of natural language communication – vagueness and underspecification – pose challenges for semantic representation. As computational linguists, how can we address this issue? We need a flexible annotation scheme that enables annotators to capture and articulate their interpretations of ambiguous or fuzzy cases, facilitating subsequent analysis and generalization.

3.3 Reliability and Consistency

Reliable and accurate labels are crucial for classification models. While it is a common practice to collect multiple annotations to ensure high-quality labels, these are often condensed into a single "gold" label through majority voting. However, this approach leads to significant information loss and uncertain ground truth labels in applications with high label variance (cf. Uma et al., 2021). Many NLU tasks provide evidence of annotator disagreement (e.g., Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Zhang and de Marneffe, 2021; Jiang et al., 2023 investigate disagreement in NLI tasks; Erk et al., 2013 provide a summary and discussion of inter-annotator agreement in WSD tasks;[3] Wein, 2025 examine disagreement in AMR framework),[4] and a growing body of research aims to develop learning methods that do not rely on the single gold-label assumption (cf. Erk et al., 2013; Dumitrache et al., 2019; Plank, 2022; Gruber et al., 2024).

Spatial functors		**Temporal functors**	
LOC	where	TWHEN	when
DIR1	where from	TSIN	since when
DIR2	which way	TTILL	till when
DIR3	where to	THL	how long
Causal functors		TFHL	for how long
CAUS	why	THO	how often
AIM	for what purpose	TFHRW	from when
CNCS	despite what	TOWH	to when
COND	under what conditions		
INTT	with what intention		
Manner and other functors			
MANN	how	EXT	how much
ACMP	accompanied by	MEANS	by means of
BEN	benefit of	REG	with regard to
CPR	comparison with	RESL	what result
CRIT	according to	RESTR	except for
DIFF	with what difference	SUBS	on behalf of
CONTRD	against what	HER	inheritance

Table 1: PDT functors for circumstants

4 Prague Dependency Treebank

We develop our semantic classification of circumstants within the Prague Dependency Treebank (PDT) project. The PDT framework is unique in its attempt to systematically include and link different layers of language including a semantic representation at deep syntactic annotation layer called tectogrammatical. Regarding the current trend in the development of semantic representations in the field of computational linguistics, it

[3] IAA is generally relatively low (66.5% to 86%) in corpora that use fine-grained sense distinctions (WordNet, FrameNet) and higher (more than 90 %) in those with more coarse-grained categories (OntoNotes).

[4] The SNACS 52-label set was used to annotate *The Little Prince* novel in English (Schneider et al., 2018), Hindi (Arora et al., 2021), Korean (Hwang et al., 2020), and Mandarin Chinese (Peng et al., 2020). IAA ranges from 75% to 93%. The results from the annotation of the SNACS project show higher agreement on linguistic meaning than on content domain.

should be highlighted that in the latest version PDT-C 2.0 (Hajič et al., 2024), there is a large amount of genre-diversified data (more than 3 million tokens) manually annotated with an interlinked semantic, syntactic, and morphological annotation. The annotation scenario of PDT is based on the original, well-developed theory of language description, so-called Functional Generative Description (FGD; Sgall et al., 1986) and was reflected in several detailed annotation manuals available from the project web site.[5]

4.1 Linguistic Meaning Layer

In Sect. 3.1, we stated that semantic representation requires distinguishing between the domain of linguistic meaning and the domain of (extra-linguistic) content. The highest tectogrammatical layer in the multi-layer PDT scheme is conceived as a layer of linguistic meaning. It captures complex semantic annotations of a sentence: predicate-argument structure, fine-grained classification of semantic roles, semantic counterparts of morphological categories, topic-focus articulation, information structure, grammatical coreference, ellipsis. Later, annotations extending beyond the level of linguistic meaning – such as coreference, bridging, or discourse relations were added.

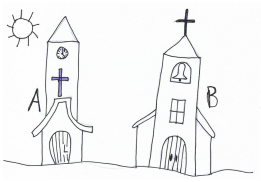

Figure 1: Same linguistic meaning and different content
A *There is a cross **on** the church tower.*
B *There is a cross **on** the church tower.*

In the PDT framework, we now focus on fine-grained classification of circumstances. We illustrate the semantic level at which our semantic classification operates using Fig. 1 and 2 and the examples below them. Our goal is to describe how a given language (in our case, Czech) reflects reality through its form and structure – that is, we describe linguistic meaning rather than content or reality itself. Therefore, our categories for spatial meanings do not distinguish the difference in the placement of the cross in images A and B (in Fig. 1) because the language itself does not make this distinction

[5] https://ufal.mff.cuni.cz/pdt-c

Figure 2: Different linguistic meaning and same content
*A tree grows **beside** the house.*
*A tree grows **near** the house.*

(the same preposition is used for both placements). On the other hand (cf. Fig. 2), we differentiate between placement "beside" something and placement "near" something, as these meanings are formally differentiated: the prepositions *beside* and *near* are not interchangeable in all contexts (cf. the proposal of spatial meaning labels in Table 4 in Appendix A). The tectogrammatical representations of sentences capture language specific patterning of the extra-linguistic content.

4.2 Two-level Semantic Classification

Regarding the arrangement and granularity of semantic categories (Sect. 3.2), we employ a two-level semantic classification of circumstants: a coarse-grained classification into *functors* (see Table 1) and a fine-grained into *subfunctors* (based on the FGD theory and first described in Panevová, 1980). While functor labelling has already been completed in the PDT corpora, the set of subfunctors is currently the focus of our research.

Functors are language-independent concepts defined by questions we ask about specific circumstances. This means that the way someone may ask (*how*, *when*, *where*, *why*, etc.), determines the granularity of the functor classification (see Table 1). Functors (although several dozen are distinguished) describe circumstantial meanings only as generalized categories and, from the perspective of linguistic meaning, they reflect only a rough classification.

A fine-grained subcategorization of circumstants into *subfunctors* involves delimiting subtle semantic distinctions within a single functor while sharing the basic semantics of that functor (answer the same question on the circumstance). The circumstants assigned different functors are not substitutable when answering a question about particular circumstance, i.e. the question "How did he work?" cannot be answered by a spatial circumstant as in (7); this question is answered by a manner circumstant (as in (1)–(6)), which may have different

sub-meanings (subfunctors). The fine-grained classification of circumstants is language-specific and based on the notion of linguistic meaning. We aim to create a set of meaning categories that have formal support in the language (see description of our methodology in Mikulová, 2024).

4.3 Fuzzy Meanings

In Sect. 3.2, we indicated that we need a set of labels that account for the high degree of vagueness in language. It becomes evident (see also Sect. 5.3) that in addition to clear, well-differentiated meanings, there are fuzzy cases, both at the level of functors and subfunctors, and that the situation is not uniform across all circumstants. While in spatial and temporal domains, the system of questions (*where*, *where from*, *where to*, etc.) is instructive and divides the conceptual time-space straightforwardly into discrete subdomains (see Table 1; ambiguous cases include the aforementioned example *at his party*, in which temporal and spatial localizations are expressed at the same time), in the manner-related domain, the basic question *how* yields diverse responses as we outlined by (1)–(6). Moreover, not all manner-related circumstants can be questioned by *how* (in (6), the only response to the question *How did John work?* is the circumstant *with success*, while the response *with his skills* is less suitable, even impossible). Therefore, we do not treat all variable manner-related circumstants as representatives of a single functor. To divide this heterogeneous group of meanings, we formulate specific questions: *with regard to what* (REG; for *with his skills* in (6)), *by means of what* (MEANS; (2)), *accompanied by what* (ACMP; (4)); see Table 1.

A similar situation arises at the level of subfunctors. While spatial and temporal meanings are typically expressed through formal means in Czech (and other languages; e.g., *before* vs. *after*, *above* vs. *below*; see the proposal of subfunctors for the LOC functor in Table 4), languages generally lack special formal means for distinguishing fine-grained subtypes within manner-related and other meanings. An exception is, e.g., the expression of +/- opposition (as in (5)). In the manner-related domain, a limited number of forms are used for various meanings (see the same form *with* used for various meanings in (2)–(6)). To distinguish subtle meaning categories, we look for other linguistic criteria. We mainly apply the principle of form substitutability (see more in Mikulová, 2024). E.g., the Czech preposition *s* 'with' in the MEANS-tool meaning (2) can be replaced by the preposition *pomocí* 'with the help of', whereas for the MEANS-material meaning (3) this substitution is not possible; in the ACMP-community meaning (4), the preposition *s* 'with' can be replaced by *společně s* 'together with', etc.

However, there are still a relatively large number of cases whose meaning is difficult to describe, where none of the well-defined labels fit well, or some overlap, even though the content described may be quite simple and clear. How can we describe the meaning of the circumstant in (8)?

(8) *Šel do kampaně **s** novou **iniciativou**.*
'He went into the campaign **with** a new **initiative**.'

To account for this situation, we introduce:

– special labels to capture generalizable fuzzy cases; e.g., we introduce the event label (see Table 4 in Appendix A) for the cases where the meanings of place and time overlap.

– special labels for distinction between central, clear meanings and complex ones (such as in (8)); cf. CIRC and side-effect labels in Table 5.

We also allow annotators to select more than one category from a list. When using a fuzzy category, annotators are required to provide a description of the meaning, thereby collecting material for further research.

5 Label Bias Experiment

The position described in Sect. 4 is supported here by a brief annotation experiment.[6]

5.1 Design

In line with the research questions that we want to address, and the annotators that we have available, we choose the following experiment design.

We examine two annotation tasks:

Task 1: Annotation of fine-grained meanings (subfunctors) within the spatial functor LOC (*where*). The spatial meanings are well-definable and formally distinguished. The proposed set of 24 labels used for the experiment is in Table 4 in Appendix A. A high inter-annotator agreement is expected.

Task 2: Annotation of meanings (both functors and subfunctors) for circumstants expressed by the polyfunctional preposition *s* 'with'. In addition to several clear meanings, the preposition

[6]Input data and experimental annotations are freely available at https://github.com/ufal/Subfunctor-annotation-experiment-2025.

Annotator	2 options (%)		Not shared (%)	
	Task 1	Task 2	Task 1	Task 2
A	11.25	13.3	6.50	17.6
B	6.25	17.6	3.75	15.2
C	9.25	13.0	2.50	13.8
D	1.25	4.0	2.50	11.3

Table 2: Percentage of sentences where each annotator selected two options or did not share the selected labels with any other annotator.

also expresses a range of less clear-cut, difficult-to-describe meanings. The proposed set of 26 labels is in Table 5.[7] In this experiment, we aim to evaluate the reliability of the taxonomy and the complexity of the task compared to Task 1.

For Task 1, 400 sentences were randomly selected from the PDT-C dataset, ensuring proportional representation of all forms in the sample. For Task 2, 500 sentences were randomly selected, ensuring proportional representation of all original functors. Each task was annotated by the same 4 annotators (A, B, C, D). In both tasks, if annotators were uncertain about the label choice, they could provide one alternative label and add an explanatory comment.

5.2 Results

To assess the complexity of the tasks and the reliability of the proposed sets of labels for consistent annotation, we evaluated both tasks from different perspectives. To compare the annotators, we measured how often they selected two options and how often the labels they proposed were not shared by any other annotator (see Table 2). In Task 1, the annotators were more confident and the choice of an option not shared by others was much rarer.

Giving the annotators the possibility to select an alternative label in the annotation made measuring inter-annotator agreement more complex than usually. For an initial estimation, we calculated Cohen's κ (Cohen, 1960) for each pair of annotators ignoring the alternative labels (see Table 3). With the exception of the pair A–B, all other pairs surpassed 0.8 in Task 1 and 0.6 in Task 2 (see Table 6 in Appendix A for more details). Also note that with the exception of annotators B and C (who agreed less in the second task, rank 2 versus 4) the pairs would be ranked the same by κ.

We also calculated Krippendorff's coefficient α (Krippendorff, 1980) to get a single number incorporating all the annotators. We removed the label `other` from the data prior to the calculation, as there could be different reasons why two annotators selected it for a given sentence; the second option was considered if `other` was the first option. The coefficient for Task 1 was calculated as $\alpha_1 = \mathbf{0.865}$, which shows a high degree of agreement, while $\alpha_2 = \mathbf{0.648}$ for Task 2 indicates poor agreement. However, we have not taken the second choice into account.[8]

A_i	A_j	κ	
		Task 1	Task 2
A	B	0.787	0.548
A	C	0.803	0.603
A	D	0.813	0.636
B	C	0.877	0.629
B	D	0.872	0.641
C	D	0.893	0.668

Table 3: Cohen's κ for each pair of annotators (considering the 1st label only) in both the tasks.

To show which subfunctors competed against each other most of the time we plotted a confusion matrix. We did not have golden data for comparison, so we created them: we used the data as "votes" for the correct subfunctor for each sentence.[9] There were still 6 sentences in Task 1 and 29 sentences in Task 2 that did not have a clear winner, so we let a fifth annotator break the ties. When populating the matrix, we considered each option separately, so we can understand the experiment as having 8 annotators, from whom only one half annotates all the data. Normalizing the matrix per rows clearly shows which subfunctors were confused most of the time or behaved similarly (see Fig. 3).[10] The numbers on the diagonal of the confusion matrix normalized per rows show the precision of the annotators, in the matrix normalized per columns, they show the recall. These two numbers are also shown together with the frequency of each subfunctor in Fig. 7 in Appendix A. We can observe how precision and recall differ in the two tasks: in Task 1, both values are relatively high and only drop around the middle of the graph, i.e., for less frequent subfunctors. In Task 2, the values are scattered almost from the beginning.

[7] The annotators assigned both functors and subfunctors in Task 2, but we used only subfunctors in the following calculations (functor is always implied by the subfunctor).

[8] Finding a satisfactory measure of agreement in this situation exceeds the scope of this paper.

[9] The first option had 1 vote, the second option had 0.95 votes, and the special value `other` had a penalty of 0.03.

[10] The other matrices are in Appendix A.

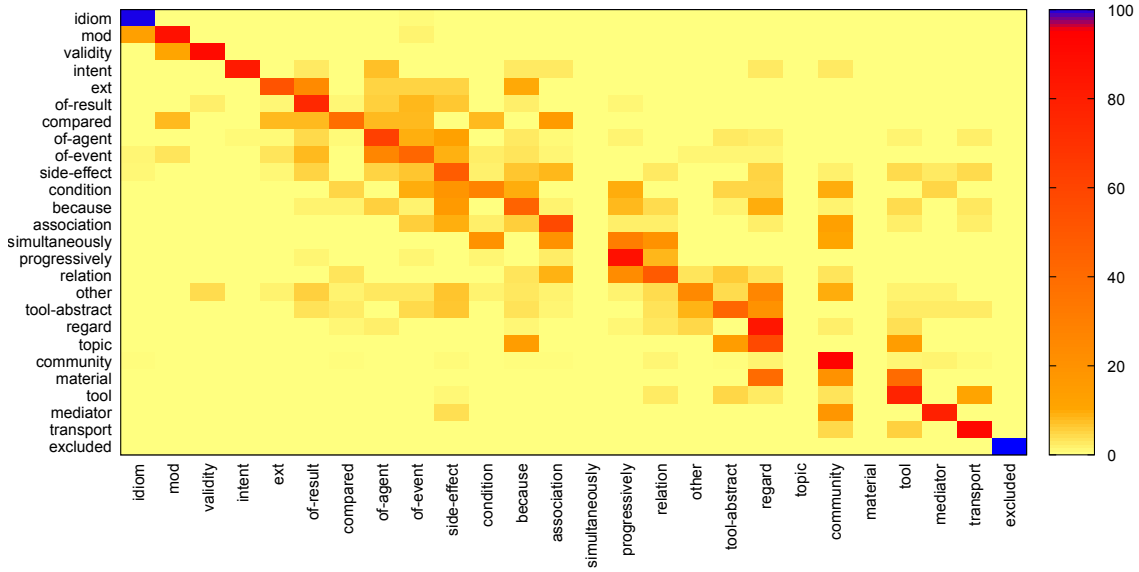

Figure 3: Confusion matrix for Task 2. It is calculated for each annotator against the created "golden data" and the values are summed for each pair of subfunctors. The matrix is normalized per rows, values are sorted to move the large values towards the diagonal as described in (Thoma, 2017) to group similarly behaving labels together.

5.3 Data Analysis

As expected, the experiment confirmed (in all measured aspects) that the annotation of fine-grained meanings in the (more manageable and formally fixed) spatial domain (Task 1) leads to more consistent annotation than the annotation of formally less distinct manner-related meanings (Task 2). In both tasks, some labels show high IAA, while others are frequently confused. Data analysis reveals competing labels.

In Task 1, there are significantly more cases with high IAA (e.g., in (9), there was 100% agreement on the meaning of `front`, in (10) on `near`), and groups of labels that were confused with each other are less common. A detailed analysis shows that cases where the form cannot be relied upon unambiguously exhibit the most hesitation and disagreement. E.g., in (11), the annotators disagreed on whether the polyfunctional preposition *u* 'beside/at' expresses the localization "beside a given place" (`adjacency`, *u divadla je škola* 'there is a school beside the theater') or a more general localization "within a given place" (`within`, *pracuje u divadla* 'he works at theater'). Disagreements typically occur with meanings of localization within a given place (`within`, `surface`, `area`), where several basic forms (*v*, *na*, *u* 'in/at/on') compete and the nature of the given place is also important (whether it has an interior and a surface); cf. (12) with competition of `area` and `inside` meanings.

(9) *Stará paní stála **před statkem**.*
'The old lady stood **in front of a farm**.'

(10) *Bydlí **blízko závodu**.*
'She lives **near a factory**.'

(11) *Dělala **u plničky** kostkového cukru.*
'She worked **at** [lit. by, beside] a sugar cube **filler**.'

(12) *Cvičila **na louce**.*
'She was exercising **in** [lit. on] **a meadow**.'

(13) ***S** psacím **strojem** se nedalo psát.*
'It was impossible to write **with the typewriter**.'

(14) ***S přibývajícím věkem** zjišťuje, že už nemá kamarády.*
'**With increasing age**, he finds out he has no friend.'

(15) ***S** velkými **oběťmi** zde udržují bezpečnost.*
'They maintain safety here **with** great **sacrifices**.'

(16) *Společnost nemá **s** těmito **akciemi** žádné plány.*
'The company has no plans **with** these **shares**.'

In Task 2, we observe high agreement only for a few clearly and narrowly semantically defined meanings, such as `community` (4), `transport`, or `tool` (13). Regarding less concrete and more abstract meanings, the label for the mutual conditionality of two events (`progressively`, (14)) shows high agreement. For other cases, the confusion matrices show which labels are closely related, and the IAA of these cases decreases. Although in the literature (Fillmore, 1994; Bonami et al., 2004) manner circumstants are usually distinguished according to their relation to an agent (5), event (1), or result (6), in real examples these distinctions are

often difficult to make. E.g., in (15) all three subfunctors (`of-agent`, `of-event`, and `of-result`) were assigned, and no single label prevailed.

The high variability of labels in many examples leads to low values of both precision and recall. E.g., the `tool-abstract` label shows very low precision. Often, when this label was used, the final agreement was on a different label. On the other hand, `regard` label has a low recall (below 60%), meaning that annotators mostly disagreed on it, however when this label was used, it was mostly in cases where there was majority agreement (e.g., in (16), `regard` label won over `tool-abstract`). The `tool-abstract` label was also assigned as an alternative label in (8). This example showed zero agreement among the 4 annotators, other assigned labels were: `mediator`, `association`, `community`, `side-effect` and the fifth annotator chose `mediator` and `side-effect`.

For further annotation, it is necessary to evaluate in which cases the disagreements occurred due to insufficient guidelines, and their improvement will lead to greater consistency. Annotators used the special fuzzy labels less than expected and tended to assign a specific meaning. This seems to be a good practice, as the merging of various labels into a fuzzy one can always be done afterwards; on the contrary, different perspectives are valuable for further investigation.

6 Conclusion

This paper puts under scrutiny the annotation of circumstantial meanings in the Prague Dependency Treebank, addressing challenges in meaning representation. Our approach centres attention on the intricate relation between language and the world it describes, emphasizing the need for a classification system that accommodates both clear-cut and vague meanings. Our two-level classification balances broad semantic concepts with fine-grained distinctions, reflecting linguistic meaning. We introduce fuzzy meaning labels for cases where rigid classification fails. An annotation experiment confirms this perspective, showing varying levels of annotator agreement, from unanimous to none. By incorporating fuzzy labels and multiple annotations, we enhance the precision and explanatory power of semantic descriptions. Ongoing development within the Prague Dependency Treebank will further refine and extend this framework.
Description of language is far from complete.

Limitation

Our experiment has several limitations. We are aware that the two tasks are not fully comparable – in the Task 1, the selected circumstants varied in form but belonged to the same semantic domain, while in the Task 2, the circumstants had the same form but differed in semantic domain. More importantly, the possibility to select a second alternative label prevented the use of standard evaluation methods, making it difficult to apply conventional metrics for assessing annotation reliability. In addition, the lack of gold standard data poses a challenge. Due to the nature of the task, such data cannot exist. Our study serves as a basis for future efforts to establish a gold standard rather than relying on one from the outset.

Acknowledgments

The research reported in the paper was supported by the Czech Science Foundation under the projects GA23-05238S and GX20-16819X. The work described herein has also been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (`https://lindat.cz`), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

We would also like to thank all our outstanding annotators for not working like machines, but for thinking critically during annotation and pointing out the shortcomings of the annotation guidelines. Without their efforts, this contribution would not have been possible.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a Multilingual Corpus of Translations Annotated with Compositional Meaning Representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2021. SNACS Annotation of Case Markers and Adpositions in Hindi. In *Proceedings of the Society for Computation in Linguistics 2021*, pages

454–458, Online. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Emily M. Bender, Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of Interpretation: On Grammar and Compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 239–249, London, UK. Association for Computational Linguistics.

Olivier Bonami, Danièle Godard, and Brigitte Kampers-Manhe. 2004. Adverb classification. In *Handbook of French Semantics*, pages 143–184, Stanford, California. Center for the Study of Language and Information.

Julia Bonn and et al. 2022. Uniform Meaning Representation (UMR) 0.9 Specification.

Simon Clematide and Manfred Klenner. 2013. A pilot study on the semantic classification of two German prepositions: Combining monolingual and multilingual evidence. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 148–155, Hissar, Bulgaria.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

William Croft. 2003. *Typology and universals*. Cambridge University Press.

František Daneš. 1966. The relation of centre and periphery as a language universal. *Travaux linguistiques de Prague*, 2:9–21.

Georgiana Dinu, Miguel Ballesteros, Avirup Sil, Sam Bowman, Wael Hamza, Anders Sogaard, Tahira Naseem, and Yoav Goldberg, editors. 2018. *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*. Association for Computational Linguistics, Melbourne, Australia.

Kira Droganova and Daniel Zeman. 2019. Towards deep Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics*, pages 144–152, Paris, France. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A Crowdsourced Frame Disambiguation Corpus with Ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring Word Meaning in Context. *Computational Linguistics*, 39(3):511–554.

Charles J Fillmore. 1994. Under the circumstances (place, time, manner, etc.). In *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society: General Session Dedicated to the Contributions of Charles J. Fillmore (1994)*, pages 158–172.

Gottlob Frege. 1892. On sense and reference.

Luke Gessler, Austin Blodgett, Joseph C. Ledford, and Nathan Schneider. 2022. Xposition: An Online Multilingual Database of Adpositional Semantics. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1824–1830, Marseille, France. European Language Resources Association.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More Labels or Cases? Assessing Label Variation in Natural Language Inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024. Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-5813.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank -

Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Eva Hajičová and Petr Sgall. 1980. Linguistic Meaning and Knowledge Representation in Automatic Understanding of Natural Language. In *Proceedings of the 8th International Conference on Computational Linguistics*, pages 67–75, Tokyo, Japan. International Committee on Computational Linguistics.

James A Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive science*, 31(3):355–384.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1/2):205–215.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Jerry R Hobbs. 1985. Granularity. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*. Elsevier.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double Trouble: The Problem of Construal in Semantic Annotation of Adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 178–188, Vancouver, Canada. Association for Computational Linguistics.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean Adposition Semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically Valid Explanations for Label Variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42:21–40.

Klaus Krippendorff. 1980. *Content Analysis*. Sage Publications, Newbury Park, CA.

Stephen C Levinson and David P Wilkins. 2006. *Grammars of space: Explorations in cognitive diversity*, volume 6. Cambridge University Press.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021. Syntax Role for Neural Semantic Role Labeling. *Computational Linguistics*, 47(3):529–574.

Ken Litkowski and Orin Hargraves. 2021. The Preposition Project. *arXiv:2104.08922*.

Inderjeet Mani. 1998. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning*, pages 245–257, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Marie Mikulová. 2024. Fine-grained Classification of Circumstantial Meanings within the Prague Dependency Treebank Annotation Scheme. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 7314–7323, Torino, Italia. European Language Resources Association and International Committee on Computational Linguistics.

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

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? *Preprint*, arXiv:2010.03532.

Tom O'Hara and Janyce Wiebe. 2009. Exploiting Semantic Role Resources for Preposition Disambiguation. *Computational Linguistics*, 35(2):151–184.

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

Jarmila Panevová. 1980. *Formy a funkce ve stavbě české věty*. [Forms and Functions in Czech Sentence Construction]. Academia, Prague, Czech Republic.

Barbara Hall Partee. 2004. *Compositionality in Formal Semantics: Selected Papers of Barbara H. Partee*. Blackwell, Hoboken, USA, Malden, MA.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A Corpus of Adpositional Supersenses for Mandarin Chinese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5986–5994, Marseille, France. European Language Resources Association.

Volha Petukhova and Harry Bunt. 2008. LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco. European Language Resources Association.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Ferdinand de Saussure. 1916. Cours de linguistique générale, ed. *C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive Supersense Disambiguation of English Prepositions and Possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A Hierarchy with, of, and for Preposition Supersenses. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA. Association for Computational Linguistics.

Wesley Scivetti and Nathan Schneider. 2023. Meaning Representation of English Prepositional Phrase Roles: SNACS Supersenses vs. Tectogrammatical Functors. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 68–73, Nancy, France. Association for Computational Linguistics.

John R. Searle. 1978. Literal Meaning. *Erkenntnis*, 13(1):207–224.

Petr Sgall. 1995. From Meaning via Reference to Content. In *Karlovy Vary studies in reference and meaning*, pages 172–183. Filosofia Publications, Prague, Czech Republic.

Petr Sgall. 2002. Freedom of language: its nature, its sources, and its consequences. In *Prague Linguistic Circle Papers: Travaux du cercle linguistique de Prague nouvelle série. Volume 4*, pages 309–329. John Benjamins Publishing Company.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague/Dordrecht.

Zoltán Gendler Szabó. 2022. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Standford, USA.

Martin Thoma. 2017. Analysis and Optimization of Convolutional Neural Network Architectures. Master's thesis, KIT – University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(2):273–290.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Zdeňka Urešová, Cristina Fernández Alcaina, Peter Bourgonje, Eva Fučíková, Jan Hajič, Eva Hajičová, Georg Rehm, Kateřina Rysová, and Karolina Zaczynska. 2024a. SynSemClass 5.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics), Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-5808.

Zdeňka Urešová, Alevtina Bémová, Eva Fučíková, Jan Hajič, Veronika Kolářová, Marie Mikulová, Petr Pajas, Jarmila Panevová, and Jan Štěpánek. 2024b. PDT-Vallex: Czech valency lexicon linked to treebanks 4.5 (PDT-Vallex 4.5). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, http://hdl.handle.net/11234/1-5814.

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm, Sook-kyung Lee, Michael Regan, and William Croft. 2019. Cross-Linguistic Semantic Annotation: Reconciling the Language-Specific and the Universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 1–14, Florence, Italy. Association for Computational Linguistics.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI-Künstliche Intelligenz*, 35(3-4):343–360.

Zhuxin Wang, Jakob Prange, and Nathan Schneider. 2021. Subcategorizing Adverbials in Universal Conceptual Cognitive Annotation. In *Proceedings of the Joint 15th Linguistic Annotation Workshop and 3rd Designing Meaning Representations Workshop*, pages 96–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shira Wein. 2025. Ambiguity and Disagreement in Abstract Meaning Representation. In *Proceedings of*

Context and Meaning: Navigating Disagreements in NLP Annotation, pages 145–154, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische untersuchungen*. Macmillan, New York, USA.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an Interlingua, But Close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 1765–1772, Reykjavik, Iceland. European Language Resources Association.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying Inherent Disagreement in Natural Language Inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

A Appendix

Subfunc tor	Forms	Example
above	*nad* 'above/over'	*nad domem* 'above the house'
adjacency	*u*, *při* 'by'	*u domu* 'by the house'
alongside	*podle*, *podél* 'along'	*podél domu* 'along the house'
among	*mezi* 'among'	*chodit mezi domy* 'to walk among houses'
area	*po* 'on/around'	*chodit po domě* 'walk around the house'
around	*okolo*, *kolem* 'around'	*kolem domu* 'around the house'
behind	*za* 'behind/beyond'	*za domem* 'behind the house'
below	*pod* 'below/under'	*pod domem* 'under the house'
beside	*vedle* 'beside/next to'	*vedle domu* 'next to the house'
between	*mezi* 'between'	*cesta mezi domy* 'path between houses'
distr	*po* 'on'	*vysedávají po hospodách* 'hang out in pubs'
event	*na*, *při* 'on/at'	*na návštěvě* 'on a visit'
facing	*čelem k* 'facing'	*čelem k domu* 'facing the house'
foreground	*v čele* 'at the head of'	*v čele kolony* 'at the head of the column'
front	*před* 'in front of'	*před domem* 'in front of the house'
ingroup	*mezi* 'among'	*mezi auty vede Škoda* 'Skoda leads among cars'
inside	*v* 'in', *uvnitř* 'inside'	*v domě* 'in the house'
middle	*uprostřed* 'in middle of'	*uprostřed domu* 'in the middle of the house'
near	*blízko*, *poblíž* 'near'	*blízko domu* 'near the house'
opposite	*naproti* 'opposite'	*naproti domu* 'opposite the house'
outside	*stranou*, *mimo* 'outside'	*stranou domu* 'outside the house'
side	*po boku* 'alongside'	*po boku manželky* 'alongside the wife'
surface	*na* 'on'	*na domě* 'on the house'
within	*na*, *u* 'at/on/in'	*pracuje u divadla* 'work at the theater'
OTHER		

Table 4: Subfunctors (and selected forms) for LOC functor (meaning "where")

Func	Subfunctor	Example
ACMP	community	*pracovat s kolegou* 'to work with a colleague'
	association	*prodávat s byty i pozemky* 'to sell with apartments also land'
	excluded	*s výjimkou Jana pracují všichni* lit. 'with exception of Jan'
MANN	of-event	*pracovat s obtížemi* 'to work with difficulties'
	of-agent	*pracovat s nadšením* 'to work with enthusiasm'
	of-result	*pracovat s úspěchem* 'to work with success'
MEANS	tool	*pracovat s lopatou* 'to work with a shovel'
	tool-abstr	*obtěžovat se zprávami* 'to bother with news'
	transport	*jet s autem* 'to go with a car'
	material	*pracovat se dřevem* 'work with wood'
	mediator	*jet s cestovkou* 'to go with a tour guide'
EXT	ext	*pracovat s velkou intenzitou* 'to work with great intensity'
COND	because	*pracovat s přinucením* lit. 'to work with coercion'
	progress	*s jarem roste nálada* 'with spring comes a rise in mood'
	relation	*změnila se vznikem klubu* 'it changed with establishment of club'
	condition	*pracovat se sluncem nad hlavou* 'to work with sun overhead'
AIM	intent	*pracovat s cílem uspět* 'work with the aim of succeeding'
REG	regard	*s přírodou není všechno v pořádku* 'all is not well with nature.'
	topic	*s tou kytarou si vzpomínám, že...* 'with that guitar I remember...'
TWHEN	simult	*souběžně s konferencí* 'simultaneously with conference'
TSIN	validity	*s účinností od ledna* lit. 'with efficiency from January'
CPR	compared	*je se mnou stejně stará* 'she is the same age as (lit. with) me.'
MOD	mod	*s největší pravděpodobností odjel* lit. 'he left with highest probability'
CIRC	side-effect	*přijet s bábovkou* 'to arrive with a cake'
	idiom	*dělej se sebou něco* 'do something with yourself'
OTHER	other	

Table 5: Functors and subfunctors for circumstants expressed by Czech preposition *s* 'with'.

A_1	A_2	κ_1	p_{o1}	p_{e1}	κ_2	p_{o2}	p_{e2}
A	B	**0.787**	0.800	0.063	**0.548**	0.584	0.080
A	C	**0.803**	0.815	0.063	**0.603**	0.634	0.078
A	D	**0.813**	0.825	0.063	**0.636**	0.666	0.083
B	C	**0.877**	0.885	0.064	**0.629**	0.658	0.078
B	D	**0.872**	0.880	0.065	**0.641**	0.670	0.081
C	D	**0.893**	0.900	0.065	**0.668**	0.694	0.077

Table 6: Details of Cohen's κ calculation: the relative observed agreement p_o and hypothetical probability of agreement by chance p_e for each pair of annotators and both the tasks.

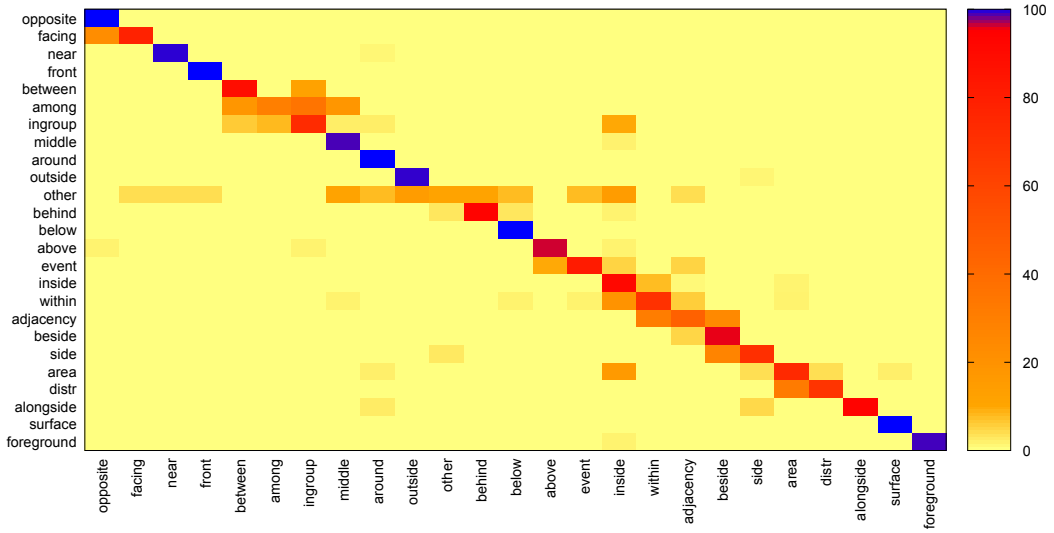

Figure 4: Confusion matrix for Task 1. A confusion matrix was calculated for each annotator against the created "golden data" and the values were summed for each pair of subfunctors. The matrix was normalized per rows, values were sorted to move the large values towards the diagonal as described in (Thoma, 2017) to group similarly behaving subfunctors together.

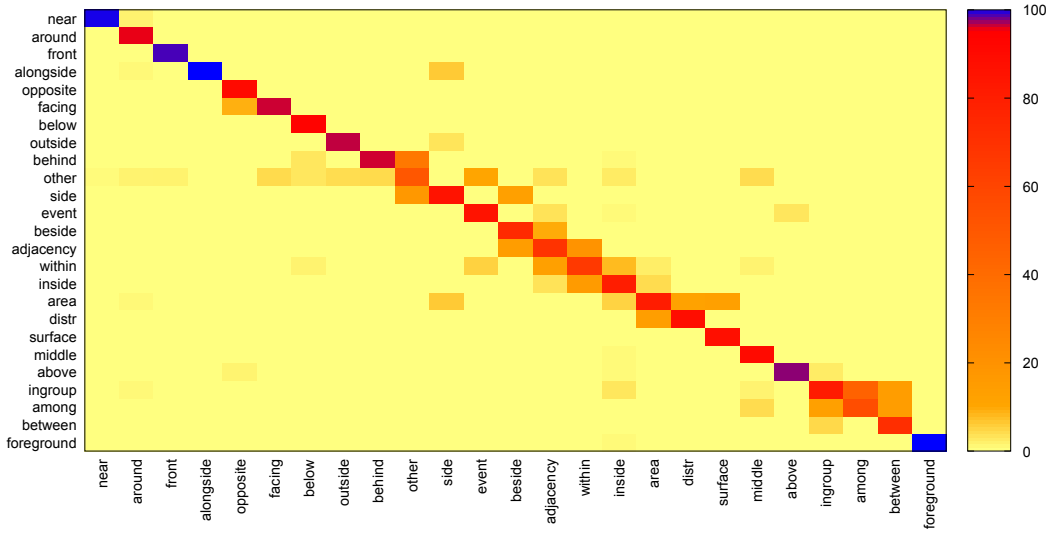

Figure 5: Confusion matrix for Task 1, normalized per columns. See Figure 4 for more details.

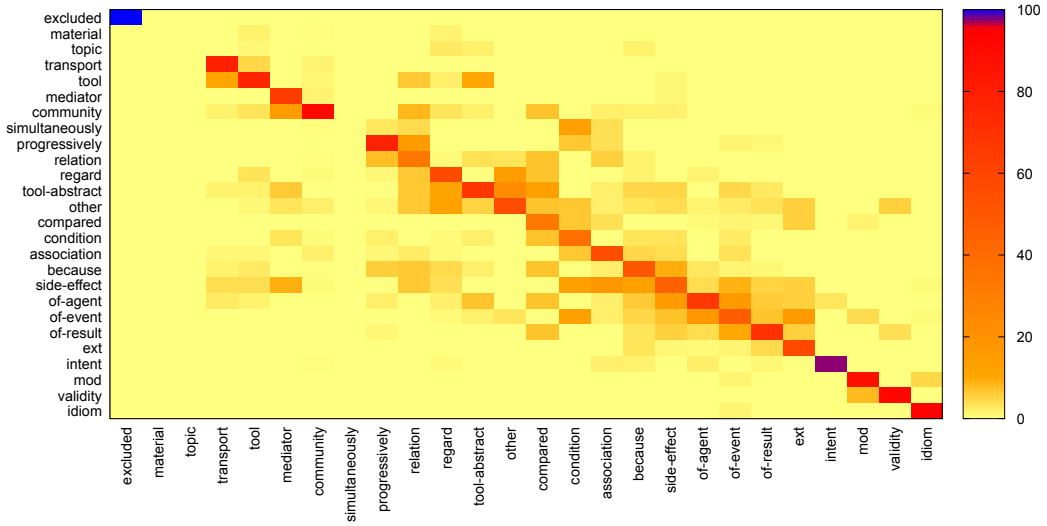

Figure 6: Confusion matrix for Task 2, normalized per columns. See Figure 4 for more details. See Figure 3 for the matrix normalized per rows.

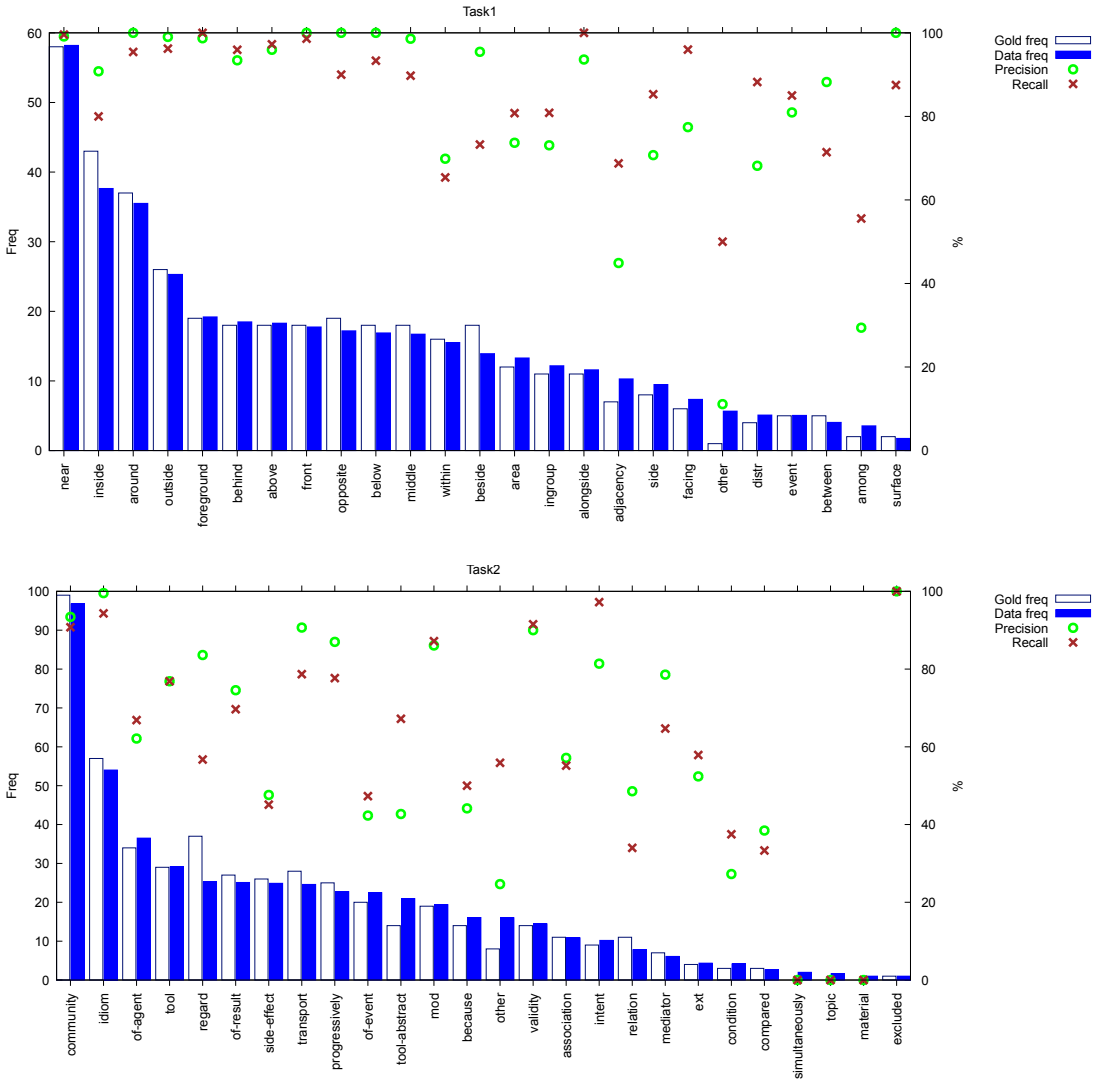

Figure 7: Comparison of subfunctor frequencies in the annotated data and in the golden data. To make frequencies comparable, the number of occurrences of each subfunctor in a sentence was divided by the number of all the values assigned by all the annotators to the sentence. Also shown are precision and recall for each subfunctor.