# Meaning Annotation Experience. A Tribute to Petr Sgall

**Marie Mikulová, Jan Štěpánek,**
**Barbora Štěpánková, Jarmila Panevová, Eva Hajičová**
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
mikulova@ufal.mff.cuni.cz

**Abstract**

We present ongoing work on annotating fine-grained semantic distinctions for circumstantial meanings, focusing on spatial expressions. We describe our theoretical background, and annotation process, as well as how we evaluate the results obtained. Using multiple independent annotations across the 3-million-token, genre-diverse Prague Dependency Treebank - Consolidated corpus of Czech data, we analyse inter-annotator agreement, recurrent disagreement patterns, and the limits of semantic categorization. Our results highlight the inherent vagueness of linguistic meaning. We also propose strategies for handling disagreement, such as weighted annotations, intermediate labels, and fuzzy labels that preserve annotation nuance. This work builds on the legacy of Petr Sgall and the Functional Generative Description theory that underpins the multi-layer form–meaning framework.

**Keywords:** meaning, semantics, vagueness, annotation, human label disagreement

## 1. Motivation

The world we live in—and the world we talk about—is immensely complex. Natural language is inherently ambiguous and vague (Sgall, 2002; Piantadosi et al., 2012); for example, no language can provide the means to distinguish between all possible locations of objects in such a world. Linguistic means (e.g., prepositions) typically serve to describe a wide range of extra-linguistic facts. We illustrate both the diversity of the world and the poly-functionality of linguistic expressions with examples from the Internet (1)–(3); see also Fig. 1.[1]

(1) *How the pumpkin got **on the tower**?*

(2) *The babies were first placed **on the tower** in 2000. Ten sculptures of toddlers climbing up and down...*

(3) *Some students standing behind a pillar of the Academic Center started shouting something about a guy **on the Tower** shooting people.*

In the examples (1)–(3), there is always the same expression from a formal perspective: the preposition *on* with the noun *tower*. However, the actual locations referred to this expression differ. Let's assume localizations (a)-(d):

(a) **at the top** of the specified place
(b) **on the outer surface** of the specified place
(c) **in the upper part** of the specified place
(d) **inside** the specified place

Localization (a) is referred to (1), as evidenced by the accompanying picture of the Cornell University tower in the respective text (cf. first picture in Fig. 1). Localization (b) is intended in (2), as illustrated by the picture, this time of the Prague tower (cf. the right picture in Fig. 1). In (3), the prepositional phrase *on the tower* expresses a rather complex localization (which includes (b), (c), (d)): the shooter was located in the upper part of the specified place, but not entirely at its top like the pumpkin in (1); he was on the tower's observation deck, but unlike the babies in (2), he was not only on the outer surface but also inside the tower (compare also the middle illustration in Fig. 1). The complexity of this localization is evidenced by the fact that while Texas Monthly magazine calls the article about the University of Texas tower shooting *The Madman **on** the Tower*, Time magazine reports on the same event under the title *The Madman **in** the Tower*.[2] This variability highlights how challenging it is for any semantic representation to capture such fine-grained distinctions.

## 2. Introduction

This paper contributes to the ongoing effort to build semantic representations,[3] emphasizing that the inherent vagueness of linguistic meaning is crucial for capturing the boundless diversity of the world.

[1] Examples (1)–(3) and the pictures in Fig. 1 are taken from the following websites:
https://cornelldailysun.github.io/pumpkin-feature/;
https://www.ourbeautifulprague.com/babies-on-the-tower-and-on-kampa/;
https://www.texasmonthly.com/true-crime/the-madman-on-the-tower.

[2] Prepositions highlighted by the authors of the paper.

[3] Enhanced Rhetorical Structure Theory (Zeldes et al., 2025), Uniform Meaning Representation (Van Gysel et al., 2021), Deep Universal Dependencies (Droganova and Zeman, 2019), Xposition project (Gessler et al., 2022); cf. also the survey papers: Ma et al. (2025); Sadeddine et al. (2024); Dobnik et al. (2022).



*Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026*, pages 160–171
11 May 2026. 

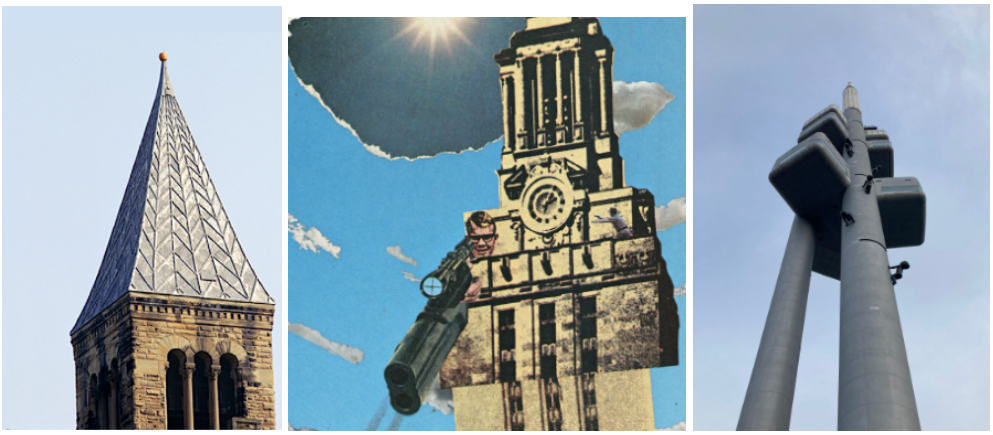

Figure 1: 'On the tower' examples illustrating different placements of pumpkin, shooter, babies on a tower, all described using the same prepositional phrase.

While underspecification, vagueness, and ambiguity rarely hinder everyday communication, they pose challenges when designing a semantic classification that is fine-grained, cognitively plausible, distinguishable, and human-understandable.[4] In this paper, we present an update on our work on designing and annotating fine-grained semantic distinctions in the expression of spatial, temporal, manner, and other circumstances in Czech (cf. previous contributions to this topic: Mikulová, 2024; Mikulová et al., 2025a) within the framework of the Prague Dependency Treebank (Mikulová et al., 2026). We address the following issues:

(i) granularity of meaning categories to ensure its credibility, broadness in coverage, and suitability for consistent manual annotation of real texts;

(ii) the relation between language and the world it describes.

We are now completing the annotation of fine-grained semantic distinctions in the spatial domain. We have produced multiple independent annotations of over 126,500 circumstants in the Prague Dependency Treebank - Consolidated 2.0, a 3-million-token corpus of genre-diverse Czech texts (Hajič et al., 2024). In the paper, we summarize the key results of this large-scale annotation, assess annotator disagreement, and discuss how the annotations will be represented in the final dataset.

The concept of extensively annotated corpora within the Prague Dependency Treebank framework is rooted in the Functional Generative Description (FGD), one of the most influential modern Czech linguistic theories (Sgall et al., 1986). The theory provides a multi-layer framework based on form-meaning relation, capturing semantic and syntactic phenomena such as valency and information structure.

[4] A considerable amount of work has emerged in this area; cf. specialized workshops: Pyatkin et al. (2024); Roth and Schlechtweg (2025); Lai and Wein (2025).

2026 marks the centenary of **Petr Sgall**, the founder of FGD. Petr Sgall greatly influenced natural language processing and helped develop computational linguistics in the Czech Republic. This paper pays tribute to his enduring impact on linguistics and language technologies at the occasion of what would have been his hundredth birthday.

The paper consists of two parts. The first chapters outline the theoretical background established in FGD: multi-layer annotation scheme (Sect. 3); the notion of linguistic meaning (Sect. 4); the distinction between meaning and content (Sect. 5); and vagueness in language (Sect. 6). The second part describes the annotation of real texts: task definition (7.1), guideline development (7.2), the annotation process (7.3), and result evaluation (7.4), followed by a discussion of disagreement treatment (Sect. 8). The paper concludes in Sect. 9.

## 3. Multi-layer Language Description

*It is indisputable that language has a complex, multi-level structure, and therefore the process of describing a language must also be structured in a certain way.*[5] (Sgall, 2006)

Our classification of circumstants is developed within the **Prague Dependency Treebank** (PDT) framework.[6] The long-term process of building the PDT corpora, as well as the current research on the circumstantial meanings, has repeatedly convinced us that a complex multi-layer annotation scheme is well founded both theoretically and computationally. Multi-layer language description views the form–meaning relation as composed of several layers, each with distinct functions contributing to overall meaning. The PDT multi-layer

[5] The original quote is in Czech: *Je nesporné, že jazyk má složitou, mnohavrstevnou strukturu, takže i postup popisu jazyka musí být určitým způsobem strukturován.*

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

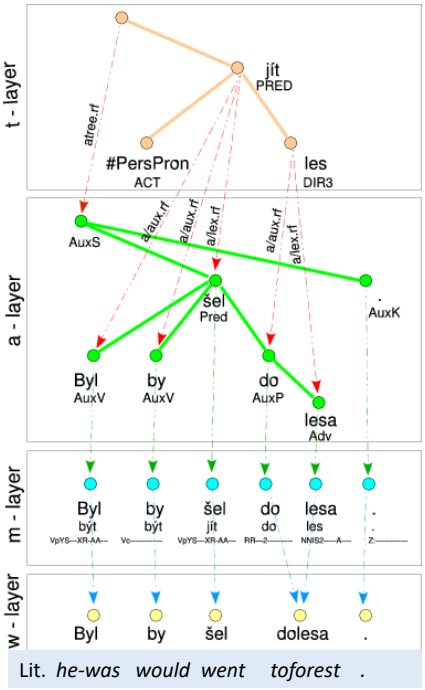


Figure 2: PDT multi-layer annotation scheme illustrated on the example of the Czech sentence: *Byl by šel do lesa.*'lit.: He-was would went to forest.'

architecture (originally proposed by Petr Sgall, e.g., in Sgall, 1965), schematically illustrated in Fig. 2, and described in numerous studies (most recently in Mikulová et al., 2026) is based on form–meaning relation and enables a comprehensive description of the relations between morphological properties, syntactic functions, and expressed meanings. It contributes to higher accuracy in language description and to the overall data consistency (Hajičová et al., 2022; Mikulová et al., 2025b).

The highest layer in PDT scheme is the layer of meaning (*t-layer* in Fig. 2),[7] while the lower layers capture surface structure (*a-layer*) and morphological properties (*m-layer*).[8] For example, the spatial circumstant *do lesa* 'to forest' (cf. Fig. 2) is represented at the t-layer by a single node with the so-called *functor* `DIR3` (meaning "where to"), which corresponds at the a-layer to a two-node prepositional structure, and at the m-layer, the morphological properties of these words are captured by a 15-character tag, indicating, among other things, their POS, case, number, and gender. In the original input text (*w-layer*), there is a typo: the prepositional phrase is incorrectly written without a space.

[7] It captures complex semantic annotations of a sentence: predicate-argument structure, semantic roles, semantic counterparts of morphological categories, topic-focus articulation, coreference, ellipsis.

[8] The other lower layers contain the input text, or, where applicable, transcription and audio.

The PDT multi-layer annotation makes it easy to examine the relations between form and meaning. It is obvious that there is no one-to-one correspondence between meaning categories and forms. For example, the form *do*+2[9] expresses not only the meaning "where to" (`DIR3`, as in Fig. 2), but also the meaning "until when" (e.g., *do pondělí* 'until Monday', `TTILL`) or quantificational modification (*do dvaceti let* 'under the age of twenty', `EXT`). And vice versa, the meaning "where to" is expressed not only by the form *do*+2, but also by a wide range of other forms, e.g., *nad*+4 (*nad les* 'above forest'), *poblíž*+2 (*poblíž lesa* 'near forest'). From these examples, we can see that functors capture circumstantial meanings only as generalized categories and from the perspective of semantic description, reflect only a coarse classification. These instances of the general meaning "where to" differ in a more specific locations ("into the specified place", "above the specified place", "behind the specified place", "near the specified place", etc.). The introduction of a set of "narrower" meanings, the so-called *subfunctors*, makes it possible to capture these semantic distinctions.[10]

The annotation of functors has already been completed in the PDT-C corpus. For the upcoming release in 2028, we further enrich the semantic annotation in the corpus by subfunctor annotation. In accordance with the principles on which the corpus is built, we establish a repertoire of subtle meaning categories and specify the formal means by which partial meanings are expressed. Methodologically, however, we proceed in the opposite direction: for particular forms, we determine their semantic functions. Morphosyntactic description and linguistic meaning representation are related but pursue opposite goals: the former uses semantic equivalence to establish formal patterning, while the latter uses formal distinctions to uncover the structure of meaning categories and compare semantic concepts (cf. Haspelmath, 2010).

## 4. Linguistic Meaning Layer

*Without distinguishing the level of meaning it is difficult to imagine an integrated description of language, since the linguistic structuring of semantic and pragmatic issues has to be described independently on what we assume to be the "real" or "actual" structure of the world.* (Hajičová and Sgall, 1980)

The top layer in the PDT multi-layer scheme (at which we annotate subfunctors) is conceived as

[9] In the paper, the nouns are indicated by number of morphological case, i.e. 2 for noun in *Genitive*, 3 for *Dative*, 4 for *Accusative*, 6 for *Locative*, 7 for *Instrumental*.

[10] The need for a finer classification of functors was first described in Panevová (1980).

| Subfunctor | Forms | Example |
|---|---|---|
| `above` | *nad* 'above/over' | *nad lesem* 'above the forest' |
| `adjacency` | *u*, *při* 'by' | *u lesa* 'by the forest' |
| `alongside` | *podle*, *podél* 'along' | *podél lesa* 'along the forest' |
| `among` | *mezi* 'among' | *chodit mezi stromy* 'to walk among trees' |
| `area` | *po* 'on/around' | *chodit po domě* 'walk around the house' |
| `around` | *okolo*, *kolem* 'around' | *kolem lesa* 'around the forest' |
| `behind` | *za* 'behind/beyond' | *za lesem* 'behind the forest' |
| `below` | *pod* 'below/under' | *pod lesem* 'under the forest' |
| `beside` | *vedle* 'beside/next to' | *vedle lesa* 'next to the forest' |
| `between` | *mezi* 'between' | *cesta mezi dvěma lesy* 'path between two forests' |
| `direction` | *na+4*, *směrem na+4* 'towards' | *jet směrem na Prahu* 'to go towards Prague ' |
| `facing` | *čelem k* 'facing' | *čelem k lesu* 'facing the forest' |
| `foreground` | *v čele* 'at the head of' | *v čele kolony* 'at the head of the column' |
| `front` | *před* 'in front of' | *před lesem* 'in front of the forest' |
| `inside` | *v* 'in', *uvnitř* 'inside' | *v lese* 'in the forest' |
| `middle` | *uprostřed* 'in middle of' | *uprostřed lesa* 'in the middle of the forest' |
| `near` | *blízko*, *poblíž* 'near' | *blízko lesa* 'near the forest' |
| `opposite` | *naproti* 'opposite' | *naproti lesu* 'opposite the forest' |
| `otherside` | *přes*, 'across' | *hodit kámen přes řeku* 'to throw a stone across the river' |
| `outside` | *stranou*, *mimo* 'outside' | *stranou lesa* 'outside the forest' |
| `side` | *po boku* 'alongside' | *po boku manželky* 'alongside the wife' |
| `surface` | *na* 'on' | *nová barva na domě* 'new paint on the house' |
| `through` | *přes*, *skrz* 'through' | *strčit ruku skrz mříž* 'to put a hand through the bars' |
| `within` | *na*, *u* 'at/on/in' | *svatba na věži* 'wedding on the tower' |

Table 1: Core subfunctors and selected forms for spatial circumstants

a layer of linguistic meaning. It captures the way semantic distinctions are structured within a given language. It differs from other domains that reflect non-linguistic structuring of (cognitive or ontological) content, primarily in two aspects (taken from the timeless article by Hajičová and Sgall (1980)):

(i) *while there is a clear support in the form of analysed language for the representation of linguistic meaning, no clear criteria have been found for the classification of units in the content/knowledge domain*,

(ii) *while a representation of meaning is one of the levels of the language system, a representation of the content is beyond language*.

We began addressing circumstantial meaning categories within the domain of linguistic meaning. This domain focuses on how a language reflects reality through its form and structure; consequently, our spatial subfunctors do not describe the exact placement of the pumpkin, the shooter, or the babies in (1)–(3) in Sect. 1, because the language itself (in our case, Czech) does not distinguish them—the same formal means are used for all three placements.[11] This strategy appears to work well for core spatial, temporal, and other circumstantial meanings; cf. Tab. 1, which summarizes our core subfunctors for the spatial circumstants.

[11] To address this, we introduced `surface` subfunctor for "on the surface" meaning and `within` for underspecified location "somewhere in there"; cf. Tab. 1.

## 5. And What is Beyond?

*The level of meaning may be considered to constitute a suitable starting point for semantic-pragmatic interpretation of the sentence, i.e. of an analysis of its cognitive (ontological) content.* (Sgall, 1995)

*The interplay of semantics and pragmatics in the structure of natural language is far too complex a matter to be dealt with by simply including some pragmatic features in the structuring of meaning.*(Sgall et al., 1986)

However, within spatial circumstants not only basic, literal spatial distinctions are expressed, but also a wide range of abstract, metaphorical, or otherwise extended meanings, which may differ from the basic one to varying degrees. Compare examples (4)–(6) with the prepositional phrase *v novinách* 'in newspaper', where the core spatial meaning `inside` is present only in (4). The other examples express transferred meanings: in (5), the newspaper is understood as an institution, and in (6) it refers to the content of the newspaper as a (literary) work. These readings appear to be desirable to distinguish in addition to the core meanings.

There are two main ways to handle these transferred meanings: assign a general "non-core" meaning label or try to divide them into finer categories. It is clear that language understanding is not based solely on linguistic meaning, but also on further semantic-pragmatic interpretation, during which the interpreter draws on contextual informa-

| Subfunctor | Examples |
|---|---|
| event | *potkat se na návštěvě* 'to meet on a visit', *odejít do války* 'to go to war' |
| state | *být v domácnosti* 'to be a stay-at-home mother', *dostat se do bezpečí* 'to get to safety' |
| aim | *hnát se do útoku* 'to rush into attack', *přijmout někoho do služby* 'to take him into service' |
| institute | *pracovat ve škole* 'to work at school', *odejít od Siemensu* 'to leave (from) Siemens' |
| institute-p | *jít k holiči* 'to go to the barber', *přijít od doktora* 'to come from the doctor' |
| ingroup | *nejmenší ve třídě* 'the smallest in the class', *vmísit se do davu* 'to blend into the crowd' |
| function | *odejít z funkce vedoucího* 'to resign from a position as manager' |
| work | *hledat humor v knize* 'to search for humour in a book' |
| media | *písnička v rozhlase* 'a song on the radio', *dostat se na obrazovku* 'to get on the screen' |
| actinfo | *uvést něco v prohlášení, ve zprávě* 'to state something in a statement, in a message' |
| domain | *působit v zemědělství* 'to work in agriculture' |
| placings | *doběhnout na třetím místě* 'to finish in third place' |
| level | *pozdvihnout zábavu na vyšší rovinu* 'to took entertainment to a higher level' |
| value | *uzavřít obchodování na 48 centech* 'to close trading at 48 cents' |

Table 2: Other subfunctors for spatial domain

tion and general world knowledge. The proposed "non-core" subfunctors are semantic-pragmatic, grounded in context and human knowledge.

(4) ***V novinách*** *najdete i přílohu.*
'You will find an addendum **in the newspaper**.'

(5) ***V*** *našich* ***novinách*** *pracovat nemůžete.*
'You cannot work **at** (lit. in) **our newspaper**.'

(6) *Ten inzerát jsem našel* ***v novinách****.*
'I found the advertisement **in the newspaper**.'

In delimiting these subfunctors, we rely on only a few linguistic "crutches". The most important is the principle of substitutability of forms (cf. Mikulová, 2024). For instance, in the "institution" meaning (in contrast to the core meaning "inside"; cf. (5)), Czech allows in some cases the use of the preposition *u*+2 'by/at' (e.g., *pracuje u novin* 'to work at a newspaper'). Another useful clue is a cross-linguistic perspective: cases in which a given meaning is formally distinguished in another language. For example, in English, the meaning 'in the newspaper' as an institution is formally distinguished from the simple spatial meaning (*in* vs. *at*). Tab. 2 summarizes the proposed "non-core" subfunctors.

Where no such "linguistic crutch" exists, the above-mentioned method of trial and error becomes necessary, and we are aware that in this area, in particular, careful evaluation is required: to what extent annotators will agree on these categories, and how well the proposed classification covers the data (see Sect. 7.4).

## 6. How Do We Understand?

*Without a certain degree of indistinctness of language meaning (i.e., of the units of the layer of functions of expressions in the language system) it would not be possible to capture with limited means the unlimited range of the world we perceive and speak of.* (Sgall, 2002)

And what lies at the end of this process? Is it even possible to sort something as inherently vague as language into meaningful categories? It turns out that after we classify the basic and clearly identifiable meanings (Tab. 1) and single out the predominantly abstract ones (with respect to the domains of the annotated texts; Tab. 2), we are still left with a relatively large number of cases whose meaning is difficult to describe, or where it is unclear how fine-grained the analysis should be: should we distinguish in the semantic representation between (4) and (7), or between (6) and (8)? It becomes evident that in cases where the annotation is not supported by the language itself—namely when non-core meanings are involved—and is based instead on knowledge and understanding, factors that rely heavily on human judgment and are therefore subjective, the annotations often diverge (cf. Mikulová et al., 2025a).

(7) ***V novinách*** *byla díra.*
'There was a hole **in this newspaper**.'

(8) *Ty lži se objevily jen* ***v*** *těchto* ***novinách****.*
'Those lies appeared only **in this newspaper**.'

This raises a broader epistemological issue: to what extent can semantic annotation strive for objectivity when the boundaries between meaning categories are fluid and context- and knowledge-dependent? Even with carefully designed guidelines, annotators may interpret borderline cases differently, not because of a lack of training, but because language itself does not provide stable cues. As a result, annotation schemes inevitably reflect theoretical assumptions about meaning segmentation, highlighting the limits of any attempt to fully systematize meaning. Semantic categorization is inherently approximate. However, examining the sources of disagreement and the limits of meaning annotation (see Sect. 8) provides a valuable insight into how speakers conceptualise space (and content more broadly).

| - | a12 | a09 | a07 | a01 | a10 | a06 | a03 | a15 | a14 | a11 | a04 | a05 | a02 | a00 | a13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a08 | 86 | 93 | 87 | 86 | 94 | 88 | 95 | 89 | 87 | 88 | 85 | 88 | 87 | 81 | 84 |
| a12 | - | 89 | 85 | 84 | 90 | 87 | 90 | 87 | 85 | 87 | 86 | 86 | 78 | 78 | 81 |
| a09 | | - | 90 | 88 | 92 | 92 | 92 | 89 | 86 | 90 | 81 | 94 | 96 | 86 | 86 |
| a07 | | | - | 88 | 90 | 89 | 90 | 88 | 87 | 89 | 90 | 90 | 89 | 79 | 85 |
| a01 | | | | - | 91 | 90 | 91 | 86 | 93 | 89 | 90 | 90 | 90 | 79 | 83 |
| a10 | | | | | - | 94 | 95 | 94 | 92 | 93 | 94 | 93 | 91 | 91 | 89 |
| a06 | | | | | | - | 94 | 92 | 89 | 93 | 92 | 92 | 91 | 86 | 86 |
| a03 | | | | | | | - | 95 | 89 | 94 | 91 | 96 | 87 | 88 | 87 |
| a15 | | | | | | | | - | 90 | 92 | 95 | 95 | 92 | 85 | 87 |
| a14 | | | | | | | | | - | 91 | 93 | 87 | 91 | 86 | 84 |
| a11 | | | | | | | | | | - | 92 | 93 | 92 | 86 | 87 |
| a04 | | | | | | | | | | | - | 93 | 96 | 83 | 91 |
| a05 | | | | | | | | | | | | - | **98** | 87 | 86 |
| a02 | | | | | | | | | | | | | - | 80 | 88 |
| a00 | | | | | | | | | | | | | | - | 79 |

Table 3: Inter-annotator agreement over all the tasks. The percentage of non-empty intersections of annotated values in all the data. Special labels are ignored.

## 7. Annotation Process

In this section, we describe practical experience with what we have theoretically discussed above, including the specific course and outcomes of our annotation. More details are in the Appendix A.

### 7.1. Defining the Task

The semantic annotation of circumstantial meanings throughout the entire corpus involves assigning subfunctors to nodes expressing spatial, temporal, manner, causal, and other types of circumstances. We began with the spatial domain. Within the PDT framework, we distinguish four functors for spatial meanings: `LOC` ("where"), `DIR1` ("from where"), `DIR2` ("which way"), and `DIR3` ("where to"). In annotating subfunctors, we take advantage of the fact that the functors are already annotated in the corpus. To ensure greater consistency, we do not annotate all spatial circumstants at once (there are approximately 126,500 occurrences, expressed by roughly 120 different forms).[12] Instead, we proceed in smaller tasks defined by functor and form—e.g., the functor `DIR3` expressed by *do*+2.

### 7.2. Developing Guidelines

The set of subfunctors for a functor–form combination is proposed on the basis of an analysis of at least 200 randomly selected corpus examples for each such combination. We use the ForFun database[13] (Mikulová and Bejček, 2018), which is extracted from the corpus and organizes its formal and semantic annotations in a user-friendly tool. When designing subfunctors, we make use of available Czech dictionaries[14] in which the preposition meanings are described to a certain extent. Most of the proposed subfunctors are shared across all spatial functors, while several more specific subfunctors are developed individually for each one. An overview of the proposed subfunctors is provided in Tab. 1 and 2.

### 7.3. Performing Annotation

Based on the proposed sets of subfunctors, we carry out multiple annotations of all occurrences of each functor–form combination throughout the entire corpus. In addition to selecting a mandatory subfunctor (or marking the case as problematic, e.g., an error in the functor assignment; such cases are not included in the calculations presented here in Sect. 7.4), annotators can also use a special label indicating an abstract or idiomatic meaning. If annotators are uncertain about the appropriate subfunctor, they are allowed to provide one more option and add an explanatory comment. Each occurrence is annotated by at least three different annotators.

### 7.4. Evaluating Results

One of the key stages in meaning annotation is a careful evaluation of the results. We assess both the annotators and the annotations themselves. We measure the extent to which annotators agree (Sec. 7.4.2) and identify which annotators disagree the most (Sec. 7.4.1). Furthermore, we examine the extent of annotator agreement for each proposed subfunctor (Sect. 7.4.3) and how often individual subfunctors co-occur (Sect. 7.4.4).

[12] A form means a preposition (including a wide range of complex ones) plus a case combination here, i.e. we do not count adverbs, subordinate clauses.

[13] http://hdl.handle.net/11234/1-2542

[14] https://prirucka.ujc.cas.cz/

| Considered | L+O+ | L+O- | L-O+ | L-O- |
|---|---|---|---|---|
| $\alpha$ | 0.706 | 0.710 | 0.727 | 0.731 |

Table 4: Krippendorff's $\alpha$. We use four different ways to calculate it: with or without the special label (L+/L-) and with or without considering option order (O+/O-).

#### 7.4.1. Inter-Annotator Agreement

As the annotators were given the possibility to use two labels, the simple classic methods of measuring their agreement are not directly applicable. To get a general overview of the consistency of the annotation, we calculate a simplified agreement as shown in Tab. 3. The highest agreement is 98%, the lowest 78%. By summing the table (without omitting the mirrored values) by rows (or columns), we can get the "most different" annotators and further inspect how their annotation differs to others.

#### 7.4.2. Krippendorff's Alpha

We also calculate Krippendorff's coefficient $\alpha$ to measure the overall inter-annotator agreement (Tab. 4). The coefficient stays above 0.667 recommended by Krippendorff for at least tentative conclusions, and when broken by task (i.e. functor and form), two tasks achieve over 0.8, the satisfactory threshold for firm conclusions (while some others drop below 0.667). Ignoring the order of options does not change the value much (mostly because a single option is prevalent), ignoring the special label has a larger impact (+0.02 overall, but up to +0.178 for one of the two worst performing tasks).[15]

#### 7.4.3. Category-wise Kappa

To calculate the agreement for a category, we calculate Cohen's $\kappa$ for each pair of annotators for the category and take the average value as the result. Tab. 5 shows the highest scoring values for spatial meanings. By category we here understand potentially both the annotated labels including their preference order but ignoring their special labels. Including two-value annotations, the entire table has 439 rows. The highest-scoring two-value category (`domain,work`) has a score of $\kappa = 0.111$. To calculate the expected agreement for each annotator pair, there are two possibilities: to base the distribution of each annotator on the overlapping data only, or to use all the annotated data by each annotator. The difference in the final $\kappa$ is always less than 0.001.

[15] Although annotation is performed by tasks and comparing them helps us refine the categories and clarify guidelines, the coefficient $\alpha$ per task does not have a scientific value because, for each task, the number of possible forms, the number of its instances, and the number of possible categories varies widely.

| Category | $\kappa$ |
|---|---|
| `value` | 0.878 |
| `event` | 0.808 |
| `inside` | 0.783 |
| `below` | 0.770 |
| `outside` | 0.766 |
| `function` | 0.765 |
| `behind` | 0.757 |
| `front` | 0.757 |
| `placings` | 0.753 |
| `level` | 0.739 |
| `opposite` | 0.721 |
| `selection` | 0.705 |
| `work` | 0.697 |
| `domain` | 0.696 |
| `within-person` | 0.666 |
| `foreground` | 0.635 |
| `above` | 0.593 |

Table 5: Category-wise Kappa (the highest scoring values for spatial meanings).

#### 7.4.4. Confusion Matrices

For each task (i.e. functor and form(s)), group of tasks, or the whole spatial domain we plot a table similar to a confusion matrix (see Fig. 3 in Appendix A). This gives us clues for the distinction which subfunctors are well defined and understood and what are the common sources of disagreement among annotators.

## 8. Handling Label Disagreement

The in-depth analysis of the results reveals three major groups of cases arising from the multiple annotations:

(i) complete agreement (100% consensus),

(ii) recurrent disagreement patterns, and

(iii) indeterminate annotations—odd or seemingly random label combinations.

We now develop strategies for how to present the differing outcomes (i)–(iii) in the final dataset. We avoid simple aggregation or majority voting, as such approaches would not accurately reflect the reality of meaning annotation and leads to significant information loss and uncertain ground truth labels in applications with high label variance (cf. Uma et al., 2021; Plank, 2022). Instead, we aim to preserve the distinction between cases where annotators showed clear agreement on a subfunctor and cases where the choice of a semantic category was inherently ambiguous, with no clear consensus among annotators.

We conducted a small **experiment** to test whether additional rounds of simultaneous multiple labelling would lead to stronger agreement. We selected 50 random cases that showed zero agreement in the original annotation annotated by small number of annotators (see Sect. 7.3) and

| - | Majority: yes | Majority: no |
|---|---|---|
| Twice: yes | 25 | 8 |
| Twice: no | 2 | 15 |

Table 6: Experiment. “Twice” means the winning label was at least two times more frequent than the second most frequent one; “Majority” means the winning label was selected by more than a half of the annotators.

had them annotated by 14 annotators. The results are in some respects quite interesting, although not particularly convincing for drawing broader generalizations (see Tab. 6). For each case, we compared the frequencies of the two most frequently assigned subfunctors. They were never the same, i.e. there was always a “winning” label. Of the 50 sentences, in 33 cases the winning label occurred at least two times more often than the second one. 27 winning labels were higher than 7 (i.e. selected by the majority), but only a half of the cases shared the two characteristics. The intrinsic difficulty of these cases is well illustrated by one annotator’s remark: *I must admit that this task made me doubt almost every case, and I often could not determine what seemed most appropriate.* Examples from the experiment are shown in (9)–(13), with the original ambiguous annotation and the annotation obtained in the experiment.

We examine which groups or combinations of subfunctors are most frequently involved in disagreements. Since some subfunctor were introduced intuitively and experimentally, without strong grounding in linguistic form (cf. Sect. 5), it is essential to verify whether annotators agree on them to a reasonable degree; those with consistently low agreement should be reconsidered. We apply two strategies: removing a label (or redistributing its instances across other labels) and introducing an intermediate label. Candidates for **removal** include `actinfo` and `media` (cf. ex. (9)), which are often confused with each other as well as with several other subfunctors, including `work` and `inside`—labels that have some of the highest category-wise Kappa scores (cf. Tab. 5).

The confusion matrices (7.4.4) reveal the recurrent disagreement patterns. Among the combinations with the highest number of occurrences are: `inside` / `institute`, `ingroup` / `institute`, `event` / `institute`. However, these subfunctors exhibit high agreement (cf. Tab. 5). For such borderline cases, it may be more appropriate to introduce an **intermediate label** rather than forcing annotators to choose a single label or removing certain labels altogether (cf. (10) and (11)). See more details in Appendix A.

The analysis also showed that there is a relatively large number of cases with indeterminate or noisy label combinations, where none of the well-defined labels fit neatly, or where labels overlap. Such cases may need to be treated as fuzzy. For instances where agreement remains inconclusive, we consider introducing a **fuzzy label** to better capture the uncertainty in meaning. Cf. (12).

Disagreements can be resolved by assigning different weights to labels obtained from multiple annotations. As well as reducing the weight of alternative labels added by annotators, lower weights can be assigned to annotations from annotators with low IAA (Sect. 7.4.1) and to early annotations.

(9) ***V** informačním **bulletinu** se na str. 3 píše: ...*
‘**In** the information **bulletin**, on p. 3 it says: ...’
Orig.: 1 work / 1 media / 1 inside / 1 actinfo
Exp.: 10 work / 3 media / 1 inside

(10) *Odehrál jsem **v NHL** čtyři dobré sezony.*
‘I played four good seasons **in the NHL**.’
Orig.: 2 event / 2 institute
Exp.: 7 event / 4 institute / 3 ingroup

(11) *Neudělal nic, aby své návrhy prosadil, a tak byly **v parlamentu** poraženy.*
‘He did nothing to push his proposals through, and so they were defeated **in the parliament**.’
Orig.: 1 institute 1 ingroup
Exp.: 9 institute / 5 ingroup

(12) *Lidé klečeli **v písku** a hledali kamínky.*
‘People knelt **in the sand** and looked for stones.’
Orig.: 1 inside / 1 among
Exp.: 8 inside / 4 among / 1 coulisse / 1 skip

(13) *V prvních letech byl **v** projekční **kanceláři**.*
‘In the early years he was **in** a design **office**.’
Orig.: 1 institute / 1 inside/abstract / 1 domain
Exp.: 14 institute

Low IAA does not necessarily indicate errors but may reflect divergence from the majority, introducing inconsistency. Annotators’ judgments improve with experience, becoming more confident and consistent; later annotations are therefore more reliable, as in (13), where all annotators eventually reached agreement.

The adjustments discussed here—removing labels with consistently low agreement, introducing intermediate and fuzzy labels, and assigning lower weights to certain annotations—could help bridge the gap between rigid taxonomies and the gradient nature of linguistic meaning.

## 9. Conclusion

We report ongoing work on annotating fine-grained circumstantial meanings in the Prague Dependency Treebank framework, focusing on spatial expressions. Our large-scale, multi-annotator annotation reveals patterns of agreement and disagreement, highlighting the inherent vague-

ness of linguistic meaning and the influence of context and world knowledge. Proposed strategies—weighted annotations, intermediate, and fuzzy labels—preserve this nuance, reflecting the true complexity of meaning annotation. This work also honors Petr Sgall, whose Functional Generative Description laid the foundation for multilayered, form–meaning-oriented approaches that continue to guide Czech linguistic annotation.

## 10. Limitations

This paper describes an ongoing research project. So far, only a subgroup of spatial circumstants has been annotated, with annotation of time circumstants being next. So far, the project has only targeted Czech, but we are aware other languages might structure the spatial details differently. We also believe that the experience gained here is applicable across languages.

## 11. Ethics Statement

Sixteen student annotators participated in the project. They were compensated fairly for their contributions in the form of monetary payment. To protect their confidentiality, all personal identifiers were removed from the data, ensuring anonymity throughout the research process.

## 12. Acknowledgements

The research reported in the paper has been supported by the Czech Science Foundation under the projects GA23-05238S and by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz).

We would 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.

## 13. Bibliographical References

Simon Dobnik, Robin Cooper, Adam Ek, Bill Noble, Staffan Larsson, Nikolai Ilinykh, Vladislav Maraev, and Vidya Somashekarappa. 2022. In Search of Meaning and Its Representations for Computational Linguistics. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 30–44, Gothenburg, Sweden. Association for Computational Linguistics.

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.

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.

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* . LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic, http://hdl.handle.net/11234/1-5813.

Eva Hajičová, Marie Mikulová, Barbora Štěpánková, and Jiří Mírovský. 2022. Advantages of a Complex Multilayer Annotation Scheme: The Case of the Prague Dependency Treebank. In *Proceedings of the 16th Linguistic Annotation Workshop*, pages 70–78, Marseille, France. ELRA.

Eva Hajičová and Petr Sgall. 1980. Linguistic Meaning and Knowledge Representation in Automatic Understanding of Natural Language. In *COLING 1980: The 8th International Conference on Computational Linguistics*, pages 67–75.

Martin Haspelmath. 2010. Comparative Concepts and Descriptive Categories in Cross-Linguistic Studies. *Language*, 86(3):663–687.

Kenneth Lai and Shira Wein, editors. 2025. *Proceedings of the 6th Workshop on Designing Meaning Representations*. Association for Computational Lingustics, Prague, Czechia.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie

Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the Era of Large Language Models: A Survey on Datasets, Evaluation, Opportunities and Challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 8679–8696, Vienna, Austria.

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. ELRA and ICCL.

Marie Mikulová and Eduard Bejček. 2018. ForFun 1.0: Prague database of forms and functions – an invaluable resource for linguistic research. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan. ELRA.

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 15th Language Resources and Evaluation Conference*, Palma de Mallorca, Spain.

Marie Mikulová, Jan Štěpánek, and Jan Hajič. 2025a. Label Bias in Symbolic Representation of Meaning. In *Proceedings of the 19th Linguistic Annotation Workshop*, pages 142–159, Vienna, Austria. ACL.

Marie Mikulová, Barbora Štěpánková, and Jan Štěpánek. 2025b. From Form to Meaning: The Case of Particles within the Prague Dependency Treebank Annotation Scheme. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2163–2175, Abu Dhabi, UAE. Association for Computational Linguistics.

Jarmila Panevová. 1980. *Formy a funkce ve stavbě české věty*. Academia, Prague, Czechia.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The Communicative Function of Ambiguity in Language. *Cognition*, 122(3):280–291.

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.

Valentina Pyatkin, Daniel Fried, Elias Stengel-Eskin, Alisa Liu, and Sandro Pezzelle, editors. 2024. *Proceedings of the 3rd Workshop on Understanding Implicit and Underspecified Language*. ACL, Malta.

Michael Roth and Dominik Schlechtweg, editors. 2025. *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*. International Committee on Computational Linguistics, Abu Dhabi, UAE.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A Survey of Meaning Representations – From Theory to Practical Utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2877–2892, Mexico City, Mexico. ACL.

Petr Sgall. 1965. Generation, Production, and Translation. In *COLING 1965*.

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. 2006. Valence jako jádro jazykového systému. *Slovo a slovesnost*, 67(3):163–178.

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

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.

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.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, 51(1):23–72.

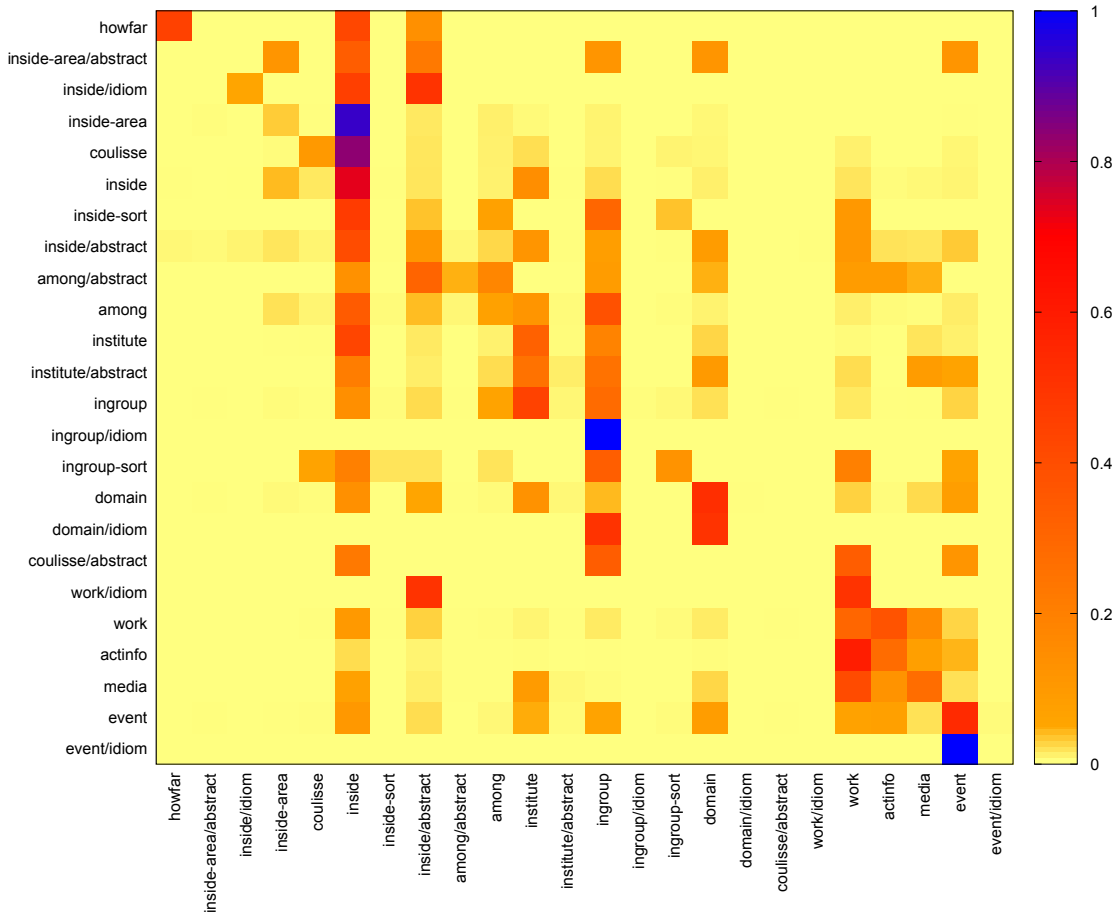


Figure 3: Confusion matrix for the spatial meaning "where" (functor `LOC`) expressed by *v*+6 'in'. There is no golden data, so we instead show how often each subfunctor rivals other subfunctors. The table is normalized by rows.

## A. Appendix

Here, we provided more information on the course and evaluation of the annotation process. As described in Sect. 7.1, the annotation is carried out in smaller batches—tasks. An example of such a task is spatial meanings of "where" (assigned in the data with the functor `LOC`) expressed by *v*+6 'in'. In this task, annotators selected from the group of 14 subfunctors and could optionally assign a special label, `abstract` or `idiom`. In total, 23,522 instances were annotated in this task, with each instance annotated at least three times. The distribution of individual meanings/subfunctors is not proportional: the most frequent subfunctor assigned was `inside` (50,385 occurrences[16]), while the least frequent was `domain/idiom` (1 occurrence).

[16]Each instance was annotated at least three times, each annotator could specify two different values, so the number of subfunctor occurrences is higher than the number of instances.

The confusion matrix in Fig. 3 shows how often, in this task, a given subfunctor (optionally combined with an associated special label, indicated after a slash) was confused with another subfunctor during annotation. The darker a cell, the more frequently the row subfunctor co-occurs with the column subfunctor. Ideally, the darkest cells should lie on the diagonal, indicating agreement in subfunctor(/special label) selection. Dark cells off the diagonal may point to problematic phenomena, but they may also reflect marginal cases. Each such case therefore needs to be evaluated carefully.

The matrix shows, for example, that the `ingroup` subfunctor occurs much more frequently (in absolute terms across all pairs of subfunctors) in combination with `institute` (2,702 occurrences) than with `ingroup` itself (1,694 occurrences). Examples (see (14) and (15)) indicate that in some cases it is indeed difficult to clearly distinguish between the meaning "within a group based on shared interests" (`ingroup`) and "within an institution" (`institute`). For instance, a *committee* can

be understood both as an institution and as a group of people of the same interest. Given that both of these subfunctors also have a substantial number of clear cases where they do not compete with any other subfunctor (cases with 100% IAA), we do not consider it appropriate to merge them. Instead, we find it more suitable to label cases of ambiguity as an intermediate `ingroup-institute` label.

The matrix further indicates that `work`, `media`, and `actinfo` are very frequent competing subfunctors; moreover, `actinfo` co-occurs more often with `work` than with `actinfo` itself. The definitions of these subfunctors have been rather vague from the beginning of the annotation process, and this caused the confusion. Subfunctor `work` refers to a content of a work (e.g., content of a book: *violence in a story*). Subfunctor `media` also involves content (e.g., a program or information) conveyed through a particular medium (*a show on television*). The subfunctor `actinfo` emphasizes the form of communication—the act of conveying information in some way (*in his statement, he said that...*), which, however, may take the form of a written text (which may be mistaken for the `work` meaning) or a television program (which may be confuses with the `media` meaning). Cf. (16), in which all three values were assigned. The `actinfo` and `work` labels compete in (17) and (18). The blurred boundaries suggest that these values should be merged into a single category.

(14) *služba **v** cizím **vojsku***
'service **in** a foreign **army**'

(15) *Zasedal **v** nejrůznějších **komisích**.*
'He served **on** various **committees**.'

(16) *Řekl to **v pořadu** Proč na stanici ABC.*
'He said this **on** ABC's **program** Why.'

(17) ***Ve** svém **článku** citoval odůvodnění trestního stíhání.*
'**In** his **article**, he cited the justification for the criminal prosecution.'

(18) ***Ve** zvláštní **zprávě** ministerstvo oznámilo, že stavební náklady byly....*
'**In** a special **report**, the ministry announced that construction costs were...'

(19) *To bylo jenom **v** nejužším rodinném **kruhu**.*
'That was only **in** the closest family **circle**.'

The `ingroup/idiom` column is empty because, apart from three occurrences with `ingroup`, `ingroup/idiom` does not appear at all. The dark cell for `ingroup/idiom` vs. `ingroup` therefore represents the most frequent (indeed the only) combination for `ingroup/idiom`, but its shading does not necessarily indicate that confusion between them is common. These marginal cases mainly concern idioms (instances with the special label `idiom`) and reveal borderline cases of idiomaticity. (19).