# Fine-grained Classification of Circumstantial Meanings within the Prague Dependency Treebank Annotation Scheme

**Marie Mikulová**
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**

In the contribution, we propose a formally and semantically based fine-grained classification of circumstantial meanings based on the analysis of a large number of valuable examples from the Prague Dependency Treebanks. The methodology and principles of the presented approach are elaborated in detail and demonstrated on two case studies. The classification of circumstantial meanings is carried out for the Czech language, but the methodology and principles used are language independent. The contribution also addresses the question of language universality and specificity through a comparison with English. The aim of this work is to enrich the annotation in the Prague Dependency Treebanks with detailed information on circumstantial meanings but it may also be useful for other semantically oriented projects. To the best of our knowledge, a similar corpus-based and corpus-verified elaborate classification of circumstantial meanings has not yet been proposed in any annotation project. The contribution presents the results of an ongoing work.

**Keywords:** semantics, annotation, circumstantial meanings, form-function relation

## 1. Introduction

*"Language is world as mirrored by form."*
(author's paraphrase of Daneš and Dokulil quote from 1958)

In the last decades, computational linguistics has become increasingly interested in annotation schemes that aim at an adequate description of the meaning of the sentences and texts. Most of these frameworks have a primary focus on verbs and their participants. Extensive databases of verb meanings are being built based on an elaborate classification of the semantic roles of verb participants (e.g. VerbNet (Kipper et al., 2008), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), SynSemClass (Urešová et al., 2022)). Relatively few frameworks, however, have focused on comprehensive accounts of non-participant roles (adjuncts, adverbials, circumstants), though they are very frequent and contribute crucial semantic information to sentences. Our contribution presents a project aimed at a fine-grained classification of the circumstantial meanings. The research question we tackle can be illustrated by the examples in (1).

(1) a. He exercised **in the gym**.
b. He exercised **before the breakfast**.
c. He exercised **hard**.
d. He exercised **on the beam**.
e. He exercised **with the ball**.

Circumstants express a very varied range of meanings. They contribute information about spatial (ex. (1a)), temporal (1b) relations, express the way in which the action is carried out (1c), the means by which it is carried out ((1d) and (1e)), etc. The meanings of circumstants can be described at different level of granularity. In the Universal Conceptual Cognitive Annotation project (Abend and Rappoport, 2013), only one category (*Adverbial*) was established for circumstants, later 7 finer categories were proposed (Wang et al., 2021). In contrast, the Preposition Project (Litkowski and Hargraves, 2021) distinguished 673 meanings of prepositional phrases.[1] Compared to other descriptions, we base our classification on the relation between the meaning and the form by which it is expressed.[2] We aim at such a classification of meanings that are formally rooted in the given language. In our classification, therefore, the meaning of circumstant expressed in (1d) is distinguished from the meaning of circumstant in (1e), because there is a different set of formal means for expressing a static device on which an activity is performed than for tools that are manipulated during the activity (see more in Sect. 4.2).

Another important issue is that of universality and language specificity of the description. It has been confirmed by many studies (e.g. Levinson and Wilkins, 2006) that every language structures the reality in a different way. When studying the relations between the forms and meanings of the circumstants, it is unavoidable to raise the question: To what extent is the set of meanings developed for one language applicable to another one? See Sect. 5 for this issue.

The description of circumstantial meanings presented in our contribution is based on the Czech language material from the Prague Dependency Treebanks (Sect. 1.1)

[1] Other related projects are mentioned in Sect. 6.

[2] The study of the relation between (linguistic) forms and their functions or meanings (terms known from Saussure's structural linguistics (Saussure, 1916) as the relation between "signifié" and "signifiant") is understood as one of the fundamental tasks of linguistics.



*LREC-COLING 2024*, pages 7314–7323

and the proposal of circumstantial meanings is subsequently used to enrich the annotation in these corpora. The methodology (Sect. 3) is grounded on the analysis of large amount of examples that PDT-corpora provide. A formally based description of meanings – reflecting the diversity and variety of the language – is demonstrated on fine-grained classification of temporal meanings (Sect. 4.1) and on circumstants of means (Sect. 4.2). Related work is discussed in Sect. 6.

The contribution presents results of an ongoing work.

### 1.1. The Prague Dependency Treebanks

The Prague Dependency Treebank (PDT) is unique in its attempt to systematically include and link different layers of language including the deep syntax (semantics). From 2006, when the pilot Prague Dependency Treebank was published, various branches of PDT-style corpora of Czech data have been developed on varied types of texts, differing in genre specification. The latest version of PDT is the Prague Dependency Treebank – Consolidated 1.0 (Hajič et al., 2020, Hajič et al., 2020).[3] It is a consolidated release of the existing PDT-corpora of Czech. It consists of four different datasets: written, translated, spoken, and user-generated texts. Altogether, the treebank contains 3,895,348 tokens with manual morphological annotation and 2,771,296 tokens with manual deep syntactic 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.[4] The Prague Czech-English Dependency Treebank (PCEDT; Hajič et al., 2012, Hajič et al., 2012),which is used for our comparative example (Sect. 5), is an annotated Czech-English corpus. The English part consists of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). Czech part was translated from the English source sentence by sentence.

When analyzing circumstants, we use the Prague Database of Forms and Functions – ForFun 1.0 (Mikulová and Bejček, 2017; Mikulová and Bejček, 2018) which is extracted from the PDT-corpora and arranges their formal and semantic annotations in a tool that provides the possibility to search, in a user-friendly way, all forms used in the PDT data for a particular meaning (functor attribute, see Sect. 2) and conversely to look up all functors expressed by a particular form.

## 2. Circumstants in the Prague Dependency Treebanks

We understand circumstants as those sentence components that provide information about the time, place, manner, reason, or other circumstances related to the events (or to the properties and entities). They are primarily expressed by prepositional (and sometimes prepositionless) phrases, as well as by adverbs and subordinate clauses.[5] Circumstants are typically adjuncts, i.e. they are not classified as arguments of predicates. However, some circumstants can have properties of arguments, blurring the boundaries between the two categories. As a result, there is often a divergence of opinions among different approaches when it comes to determining whether specific circumstants should be viewed as participants (arguments) or non-participants (adjuncts; cf. also in Sect. 6). The list of circumstants we work with can be found in Table 1.

| | | | |
|---|---|---|---|
| **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 | aim | TFHRW | from when |
| CNCS | despite what | TOWH | to when |
| COND | conditions | | |
| **Manner and other functors** | | | |
| MANN | manner | EXT | extent |
| ACMP | accompanion | MEANS | means |
| BEN | beneficiary | REG | regard |
| CPR | comparison | RESL | result |
| CRIT | criterion | RESTR | restriction |
| DIFF | difference | SUBS | substitution |

Table 1: PDT functors for circumstants

### 2.1. Two-level Semantic Classification

In the project, we work with a two-level semantic classification of circumstants: a basic coarse-grained classification into so-called *functors* and a fine-grained classification into *subfunctors* (both based on the FGD theory and first described in Panevová, 1980). Functors are defined by the questions we ask about specific circumstances, such as "where", "to where", "when", "since when", "how", "why", etc. (see Table 1). Functors describe circumstantial meanings only as generalized categories and, from the perspective of the description 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 “by what means the exercise was performed” cannot be answered by a circumstant of time (TWHEN) as in ex. (1b), this question is answered by a circumstant of means (MEANS), which may have different partial sub-meanings (subfunctors), cf. ex. (1d) and (1e).

[3] https://hdl.handle.net/11234/1-3185

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

[5] Here we focus mainly on circumstants expressed by prepositions.

## 2.2. Multi-layer Annotation Scheme

In the multi-layer PDT annotation scheme, the issue of circumstants is reflected at the highest deep syntactic layer called tectogrammatical. At this layer, the tree-like dependency representation is conceived of as a linguistically structured meaning of the sentence. The *functor* attribute (as a types of the semantico-syntactic relations) is attached to all nodes. In the current stage of the PDT representation, no subcategorization into *subfunctors* is provided and it is now the subject of our project.

Figure 1 demonstrates the necessity of annotating subfunctors: the semantic difference between the sentences *He exercised before breakfast* and *He exercised after breakfast* is captured solely by the value of the subfunctor (`before`, `after`). Without this annotation, the representations of the two sentences are identical.

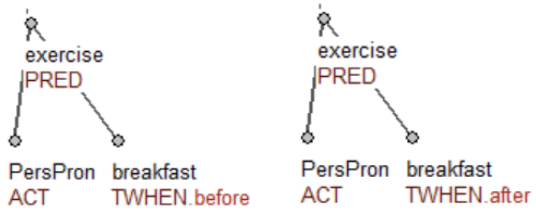


Figure 1: PDT representation of the sentences: *He exercised before breakfast* and *He exercised after breakfast*.

The lower layers of PDT contain surface syntax and morphological annotation, among other they contain information about the formal realization of examined circumstants (i.e. their part-of-speech, preposition used, grammatical case). There is no one-to-one correspondence between the nodes at the lower layers and the tectogrammatical one. The nodes of the tectogrammatical tree represent semantic units, i.e. one node for each content word together with its auxiliary words such as prepositions, conjunctions, etc. E.g., the prepositional case *after breakfast* is one node with the functor `TWHEN` (cf. Fig. 1). To preserve the original information, the respective nodes at the lower layers (i.e. node for the noun *breakfast* and node for the preposition *after*) are explicitly referred to from this node. These links allow to combine information from different annotation layers.

# 3. Fine-grained Classification of Circumstants

The fine-grained classification of circumstants is based first of all on a detailed analysis of the rich language material presented in the PDT corpora via ForFun database (Sect. 1.1). The starting point for the research is a set of (coarse-grained) functors for circumstants (Table 1) verified by manual annotation of 175,000 sentences in PDT-corpora. We study all formal realizations for each functor and we examine which sub-meanings the forms express and compile a set of subfunctors.

## 3.1. Formal and Lexical Nature of Fine-grained Classification

We establish only such sub-meanings (subfunctors) for which there is support in the structure of the language. For the expression of temporal and spatial sub-meanings, a system of formal means is fixed in most languages (cf. different temporal relations (`TWHEN-after` and `TWHEN-before`) between *postupovat* 'to proceed' and *návrat* 'arrival' expressed by prepositions *po* 'after' and *před* 'before' in ex. (6) and (7). Also +/- opposition is usually clearly expressed by formal means (e.g. the +/- sub-meanings within the `BEN` functor: *pro návrh* 'for the proposal' - *proti návrhu* 'against the proposal').

To distinguish sub-meanings within circumstants such as manner, means and other, the language often does not have special formal means and "relies" on the lexical meaning of the words used. Within these circumstants, a limited number of formal means are used for various meanings, which can be classified in varying depth mainly on the basis of lexical meaning; cf. the `tool` and `material` sub-meanings (of `MEANS` functor) expressed by the same form in the ex. (2) and (3). However, it turns out that even within these "lexical" functors, a set of typical forms (not interchangeable with another set) can be established for the partial sub-meanings. E.g. the `MEANS-tool` meaning can be expressed (in Czech) by preposition-less instrumental case (ex. (3)), by prepositions *s+7* 'with' (4) and *pomocí* 'with the help of' (5) and for the `MEANS-material` meaning, there is a different set of forms (cf. the cell labelled `material` in Table 3).

(2) *Mazal chleba* ***máslem*** (Instrumental case)
'He spread the bread (with) **butter**.'

(3) *Mazal chleba* ***nožem*** (Instrumental case)
'He spread the bread (with) **a knife**.'

(4) *Mazal chleba* ***s nožem***.
'He spread the bread **with a knife.**'

(5) *Mazal chleba* ***pomocí nože***.
'He spread the bread **with the help of a knife**.'

## 3.2. Methodology and Principles Used

Our methodology is based on the assumption that there is no one-to-one relation between the meaning represented by the functor-subfunctor combination and its formal realization: one form is used for expressing more meanings and one meaning can be expressed by using various forms. The three principles are applied in our analysis:

- principle of form substitutability (3.2.1),
- principle of analogy/systematicity (3.2.2),
- principle of generalizability/separability (3.2.3).

### 3.2.1. Principle of Form Substitutability

The crucial criterion used is the principle of form substitutability postulated in formal semantics (Peregrin, 2003). When deciding which forms are synonymous and thus can be described by the same subfunctor, we test whether the forms are substitutable in different contexts and how the meaning is influenced by the substitution.

| Subfunctor | Forms | Examples |
|---|---|---|
| at | *v+6* 'in / at', *na+4* 'on / at', *o+4* 'about', *v době*, *v čase* 'in time of' | *Většinou se scházíme **o** Vánocích.* 'We usually get together **at** Christmas.' |
| begin | *začátkem*, *na začátku*, *zkraje* 'at the beginning of' | ***Na začátku** roku přišlo 34 cizinců.* '**At the beginning of** the year, 34 foreigners came.' |
| end | *koncem*, *na sklonku* 'at the end of'; *ke konci* 'towards the end of' | ***Ke konci** války se narodila moje sestřenice.* '**Towards the end of** the war, my cousin was born.' |
| middle | *uprostřed* 'in the middle of'; *v polovině*, *v půlce* 'in the half of' | *Pojedou tam **v polovině** prázdnin.* 'They will go there **in the half of** the holidays.' |
| turn | *na přelomu*, *na předělu* 'at the turn of' | *Průzkum proběhl **na přelomu** června a července.* 'The survey took place **at the turn of** June and July.' |
| moment | *v okamžiku*, *v momentu*, *ve chvíli* 'at moment of' | *Poprvé ho viděla **v okamžiku** svatby.* 'She saw him **at the moment of** marriage firstly.' |
| during | *během*, *v průběhu* 'during'; *zatímco* 'while' | ***Během** pobytu ochutnal místní gastronomii.* '**During** the stay he tasted local meals.' |
| around | *okolo*, *kolem* 'around' | *Postřelil ho **kolem** půlnoci neznámý muž.* 'He was shot by an unknown man **around** midnight.' |
| between | *mezi* 'between' | *Nejvíc telefonátů je **mezi** 16. až 18. hodinou.* 'Most phone calls are **between** 4 and 6 p. m.' |
| outside | *mimo´*, *vně* 'outside' | *Vstoupili do budovy **mimo** pracovní dobu.* 'They entered the building **outside** working hours.' |
| distr | *po+6* 'in+ plural of nouns' | ***Po** večerech doma brečela.* 'She cried at home **in** the evenings' |
| after | *po+6* 'after' | *Nájemné za byt se **po** zateplení zvýší.* 'The rent for the flat will increase **after** warming.' |
| justafter | *jakmile* 'as soon as'; *ihned po* 'immediately after' | ***Jakmile** přijde telegram, jedete.* '**As soon as** the telegram arrives, you are on way.' |
| before | *před* 'before'; *než* than'; *v době než* 'in time than' | ***Před** tréninkem si popovídal s útočníkem.* 'He had a chat with the striker **before** training.' |
| justbefore | *k+3* 'towards', *ihned před* 'just before' | *Už bylo **těsně před** půlnocí a zvonil telefon.* 'It was **just before** midnight and the phone rang.' |

Table 2: Subfunctors (and selected forms) for TWHEN functor (meaning "when")

E.g., a temporal circumstant expressed by the preposition *po* 'after' is substitutable with the dependent clause with conjunction *až* 'when' only in the case when the subtle meaning is "after the given time" (labelled TWHEN-after); cf. ex. (6) and (8). In the case of temporal distributiveness (cf. ex. (9) labelled TWHEN-distr), the form *po* 'after' cannot be replaced by *až* 'when'.

(6) *Jak postupovat **po** návratu do vlasti?*
'How to proceed **after** arrival to the homeland?'

(7) *Jak postupovat **před** návratem do vlasti?*
'How to proceed **before** arrival to the homeland?'

(8) *Jak postupovat, **až** se navrátíte do vlasti?*
'How to proceed **when** you arrive at home?'

(9) ***Po** večerech doma brečela.*
'She cried at home in the evenings.'

### 3.2.2. Principle of Analogy/Systematicity

We do not establish subfunctors at random, but we apply the principle of systematicity/analogy. This means that if, e.g., the partial meaning of "at the beginning of" is defined (expressed by secondary prepositions *začátkem*, *zkraje* 'at the beginning of'), we examine whether the similar forms can be found for the meaning of "at the end of" (cf. the cells labelled begin and end in Table 2).

### 3.2.3. Principle of Generalizability

There are many preposition-like expressions in the language material, but we only work with those for which other synonymous forms can be found, i.e. we are dealing with groups of forms with a generalizable meaning. E.g., we observe that there are several synonymous formal means expressing the meaning "on behalf of": *jménem*, *v zastoupení* 'on behalf of'. We consider these MEANS circumstants to be a separate category (cf. the cell on-behalf-of in Table 3).

## 4. Two Case Studies

We demonstrate our fine-grained classification of circumstants on the example of determining sub-meanings for the general temporal meaning "when" (TWHEN functor) and for the circumstant of means (MEANS functor). TWHEN is an example of a circumstant with an elaborate system of formal means for different sub-meanings (cf. Sect. 4.1); the analysis of the MEANS functor is more lexical in nature (cf. Sect. 4.2).

### 4.1. Circumstant of Time

Circumstant of time (TWHEN functor) expresses the answer to the question "when". In PDT treebanks (searched via the ForFun database), there is a total of 68,781 examples and about 80 different formal realizations for the TWHEN functor (some of which are obvious annotation errors). The primary prepositions used to express the temporal meaning of "when" divide it into three basic sub-meanings, especially: "at the given time" (at subfunctor), "before the given time" (before) and "after the given time" (after). The other types of temporal sub-meanings add specialized meanings to these three basic categories. The subfunctors for the TWHEN functor are displayed in Table 2. The list of forms is only illustrative and not exhaustive.

As we can see in Table 2, the basic sub-meaning "at the given time" (at) is expressed not only by several primary prepositions (the most common are *v+6* 'in', *na+4* 'at' and *o+4* 'about/at'), but also by the secondary forms. The secondary forms can be substituted for the primary ones without changing the meaning. They contain the word *doba* 'period' or *čas* 'time' and thus convey the meaning of "at the given time" more explicitly. We determine these secondary forms on the basis of the principle of form substitutability characterized above (Sect. 3.2.1). The sub-meanings expressed only by secondary forms (e.g. begin, end, middle, during, outside, turn) are determined on the basis of the principle of systematicity (Sect. 3.2.2) and the principle of generalizability (3.2.3). See more in Mikulová and Panevová, 2021.

### 4.2. Circumstant of Means

Circumstant of means (MEANS functor) expresses, in a broad sense, the answer to the question "by means of what an event is performed or realized". In PDT treebanks, there is a total of 8,619 examples and about 14 different formal realizations for the MEANS functor. The small number of formal means (compared to TWHEN functor) and the large number of different sub-meanings indicate that the sub-meanings are expressed mainly lexically; cf. ex. (10)–(15), where the same means of communication is expressed by six different forms, while the meaning of the circumstant remains the same.

(10) *Hovořili spolu pouze **telefonem**.*
'They only talked to each other **on the phone**.'

(11) *Dohodla jsem si **po telefonu** zvláštní sazbu.*
'I negotiated a special rate **over the phone**.'

(12) *Komunikoval s námi výhradně **přes telefon**.*
'He communicated with us exclusively **by phone**.'

(13) *Hovory se odehrály přes sklo **prostřednictvím telefonu**.*
'The calls were made **via telephone**.'

(14) *Doteď jste komunikovali pouze **pomocí telefonu**.*
'So far you have communicated **using the phone**.'

(15) *Dnes reportérovi **telefonicky** odpovídám.*
'Today I answer the reporter **telephonically**.'

This does not mean that the form of expression of the MEANS circumstant is arbitrary or any of the possible forms; the typical forms can be established for partial sub-meanings. The purpose of the fine-grained classification of means is not an ontological/dictionary classification of means into means of transport, coercion, production, communication, mass destruction, musical instruments, logical means, etc., but a description of linguistically structured meaning: we look for such sub-meanings, for which there is a support in the formal realization.

An overview of the proposed subfunctors for the MEANS functor is in Table 3. The general category of means is divided into four basic subgroups: tool, coulisse, material and medium. The tool subfunctor stands for a tool, prop, aid, or device (but also for a procedure or performance) that is manipulated in order to perform or realize something. Typical forms are: preposition-less Instrumental case, *s+7* 'with', *pomocí* 'with the help of/using'. In all cases of the tool subfunctor, it is possible to replace the used form with the explicit form *pomocí* 'with the help of/using'. Musical instruments represent a special category (tool-instrument subfunctor). They are expressed (among others) by an (Accusative) form *na+4* 'on'.

The coulisse subfunctor stands for a static device on which some activity is performed. Typical forms are the primary local prepositions *na+6* 'on/at', *v+6* 'in', *po+6* 'on/at', which, unlike local circumstants, cannot be replaced by other local sub-meanings (cf. *cvičit na kladině* 'to exercise on the beam'; the expression *cvičit vedle kladiny* 'to exercise next to the beam' has a totally different meaning). Local prepositions are also compatible with some "tools", this means that some tools or devices can be viewed both as a coulisse and as a tool (e.g. *to work on a computer* (coulisse) - *to work with a computer* (tool)). Such an ambiguous concept is also manifested in the case of means of transport (tool-transport). Unlike other tools, they are also expressed in forms typical of coulisse (*na+6* 'on', *v+6* 'in'). It thus stands on the borderline between a tool and a coulisse. A separate category (subfunctor coulisse-obstacle) is an "obstacle" expressed in Czech by the form *o+4* 'against'.[6]

The material subfunctor stands for a substance, a material used to perform or make something happen. The material subfunctor is to be distinguished from tool. In some cases they are expressed by the same form (cf. ex. (16) and (17)). However, while in the tool case, the alternative expression *pomocí* 'with the help of/using' is possible, in the material case it is not (cf.

[6] Whitin the FGD framework, the "obstacle" meaning was previously described as a quasi-valency participant (Lopatková and Panevová, 2005).

non-occurring **kropit pomocí vody* 'to sprinkle with the help of water').

(16) *Kropil zahradu **vodou*** (Instrumental case)
'He sprinkled the garden (with) **water**.'

(17) *Kropil zahradu **hadicí*** (Instrumental case)
'He sprinkled the garden (with) **a hose**.'

Similar to the case of musical instruments, we establish a separate category `material-power` for which the form *na+4* 'on' is fixed in Czech.

The `medium` subfunctor expresses the external mediator of the action. It is explicitly expressed by the secondary form *prostřednictvím* 'via'. We distinguish several separate categories, with which it turns out to be useful to distinguish between animate (`mediator`) and inanimate intermediator (`intermediary`). Other categories are: `medium-transfer`, `medium-language`, `medium-exchange`, `medium-massmedia`, `medium-storage`, `on-behalg-of` (cf. Table 3).

## 5. Multilingual Perspective

In this section, we address the question of universality and specificity in the proposed representation of circumstantial meanings from a multilingual perspective. When we use linguistic form as one of the criteria for determining semantic distinctions, such a criterion inevitably leads to different representations for different languages. It is undeniable that every language structures reality in a different way, and there may be an *„overwhelming diversity, and apparently endless mismatches between any two languages in both the formal coding of distinctions, and semantical basis for them"* (Levinson and Wilkins, 2006). However, certain tendencies can be observed in the relation between forms and their meanings across languages, and cross-language studies help to explore the differences in structuring reality. Describing these differences is valuable for language understanding tasks and significantly contributes to the question of language universality.

Within the project, we exploit the fact that we have at our disposal the manually annotated parallel Czech-English corpus PCEDT (cf. Sect. 1.1). We therefore intend to enhance the circumstantial meanings description by a comparative studies in which we apply the proposed set of circumstantial functors and subfunctors to the English part of the corpus. On the basis of this comparison, we assess the universality and language specificity of the suggested set of circumstantial meanings.

Here we present only a small example of such a comparison. We compare formal realizations of the corresponding circumstants of means, focusing on the difference between `tool` and `medium` sub-meaning explicitly expressed in Czech by secondary forms *pomocí* 'with the help of' (and its variants, e.g. *s pomocí* 'with the help of') and *prostřednictvím* 'via' respectively. In the Czech part of PCEDT we have searched for circumstants with the `MEANS` functor (depending on a verb) expressed by these prepositions, and then looked for their most frequent equivalents in the English part. The Czech-English pairs of sentences were then sorted out according to the form of the English equivalent (see Table 4).[7]

In spite of the fact that the collected material is not large, certain tendencies can be followed:

**(A)** There are different sets of equivalent prepositions in English for the Czech preposition *pomocí* 'with the help of', which is typically used to express `tool` meaning, and for the preposition *prostřednictvím* 'via', which is typically used to express `medium` meaning. The equivalents for *pomocí* 'with the help of' are mostly the forms *using*, *with*, and *with the help of* (and its variants, e.g. *with the aid of*). The equivalents for *prostřednictvím* 'via' are mostly the forms *through*, *via*, and *by*. Cf. ex. (18)–(20) for the `tool` meaning; and (22)–(23) for the `medium` meaning.

(18) *Obchodníci pracovali **za pomoci** slunečního světla pronikajícího okny.*
Traders worked **with the help of** sunlight streaming through windows.

(19) ***S pomocí** kalkulačky a cenových tabulek prací mohou likvidátoři škod spočítat hodnotu domu na dnešním trhu.*
**Using** a calculator and a unit-price guide, adjusters can figure out the value of a home in today's market.

(20) *Mezi jednotlivými záběry dostanou subjekty **pomocí** magnetického stimulátoru šok do určité oblasti mozku.*
Between flashes, certain areas in subjects' brains are jolted **with** a magnetic stimulator.

(21) *Společnost Trans World Airlines Inc. nabízí **prostřednictvím** společnosti Drexel Burnham prioritní obligace v hodnotě 150 miliónů dolarů.*
Trans World Airlines Inc., offering of $150 million senior notes, **via** Drexel Burnham.

(22) *Federal Home Loan Mortgage Corp. - 250 milionů dolarů v hypotečních cenných papírech REMIC je nabízeno v 11 třídách **prostřednictvím** společnosti Morgan Stanley.*
Federal Home Loan Mortgage Corp. – $250 million of Remic mortgage securities being offered in 11 classes **by** Morgan Stanley.

(23) *V květnu nabídly tyto dvě společnosti **prostřednictvím** své společně vlastněné holdingové společnosti Temple za společnost Sea Containers 50 dolarů za akcii.*
In May, the two companies, **through** their jointly owned holding company, Temple, offered $50 a share, for Sea Containers.

**(B)** The tendency to differentiate between `tool` and `medium` meaning is evident in both languages. The boundaries between the meanings blur when it comes to

[7] We exclude cases where the equivalent in the English sentence is not a prepositional group. However, these cases are not without interest and should be dealt with in the future, cf: *spustit kampaň prostřednictvím pošty* 'lit. to mount campaign via mail' vs. *to mount a direct-mail campaign*. We also exclude cases of annotation mistakes and we do not work with idiomatic and fixed phraseological expressions.

| Subfunctor | Forms | Examples |
|---|---|---|
| tool | Instrumental, *s+7* 'with', *pomocí* 'with the help of' | *Smlouva se uzavírala **rukoudáním**.* 'The contract was concluded **with a handshake**.' |
| tool-transport | Instr, *na+6* 'on', *v+6* 'in', *s+7* 'with', *pomocí* 'with help of' | *Zloděj jezdil **v** pronajatém mercedesu.* 'The thief was driving **in** a rented Mercedes.' |
| tool-instrument | Instr, *na+4* 'on', *do+2* 'to', *s+7* 'with', *pomocí* 'with help of' | *Trubte **na** trumpety.* 'Blow the trumpets.' |
| coulisse | *na+6* 'on', *v+6* 'in', *po+6* 'on' | *Cvičil **na** hrazdě veletoče.* 'He practiced high spins **on** the trapeze.' |
| coulisse-obstacle | *o+4* 'on/against' | *Spálil se **o** kamna.* 'He burned himself **on** the stove.' |
| material | Instrumental, *s+7* 'with' | *Krmil nemocnou dceru **polívkou**.* 'He fed his sick daughter **soup**.' |
| material-power | *na+4* 'on' | *Lokomotiva jezdí **na** elektřinu.* 'The locomotive runs **on** electricity.' |
| medium | *přes+4* 'through', *prostřednictvím* 'by' | *Pronikla jsem k nim **prostřednictvím** numerologie.* 'I got to them **through** numeralogy.' |
| medium-mediator | *přes+4* 'through', *od+2* 'from', *s+7*'with', *prostřednictvím* 'by' | *Obraz byl zakoupen **přes** Sotheby´s .* 'The painting was purchased **through** Sotheby's.' |
| medium-intermediary | *na+4* 'on', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Vstup do zahrady je **na** vstupenku.* 'Entrance to the garden is **by** ticket.' |
| medium-transfer | Instr, Adv, *přes+4* 'through', *v+6* 'in', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Domluvili se **telefonicky**.* 'They agreed **by phone**.' |
| medium-language | Instr, Adv, *v+6* 'in', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Přednáška bude pronesena **anglicky**.* 'The lecture will be given **in English**.' |
| medium-exchange | Instr, *v+6* 'in', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Platil jen **v** dolarech.* 'He only paid **in** dollars.' |
| medium-massmedia | *na+6* 'on', *v+6* 'in', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Sledujete tenis i **na** internetu?* 'Do you also watch tennis **on** the Internet?' |
| medium-storage | *na+6* 'on', *prostřednictvím* 'by', *pomocí* 'with the help of' | *Digitální fotografie pořizuju **na** papír.* 'I take digital photos **on** paper.' |
| on-behalf-of | *jménem*, *ve jménu*, *v zastoupení* 'on behalf of' | *Promluvil jsem **v zastoupení** velitele.* 'I spoke **on behalf of** the commander.' |

Table 3: Subfunctors and forms for MEANS functor (meaning "mediated by what?")

an abstract concept, an event, etc. In such cases, formal means are indistinct in both languages. Cf. (24)–(28).

(24) *financovat rozsáhlá převzetí **pomocí** prodeje rizikových obligací*
to finance large takeovers **with the help of** junk bond sales

(25) *Tyto půjčky bývají často refinancovány **prostřednictvím** prodeje vysoce rizikových obligací.*
Such loans are often refinanced **through** the sale of high-risk bonds.

(26) *B. A. T. zamýšlí získat podporu pro své velké papírenské podniky **pomocí** emisí akcií.*
B.A.T aims to float its big paper businesses **via** share issues.

(27) *Získáme je **pomocí** bankovních půjček.*
We'll raise it **through** bank loans.

(28) *Prosazují jednobarevný vzhled **prostřednictvím** kolekcí provedených v černé a bílé.*
They have been advancing the monochrome look **with** collections done in black and white.

This analysis serves as a brief illustration of a more extensive concept. We posit that the analysis of linguistically structured meanings should be undertaken individually for each language, after which the connections between these meanings, along with their corresponding sets of forms, can be explored. These connections between semantic categories can exhibit variations, encompassing complete correspondence, overlap, or subsets.

| Czech form | English form | Number of pairs |
|---|---|---|
| *pomocí* | using | 21 |
| | with | 18 |
| | with the help of | 11 |
| | through | 9 |
| | by | 9 |
| | via | 3 |
| *prostřednictvím* | through | 125 |
| | via | 21 |
| | by | 14 |
| | in | 3 |
| | with | 3 |
| | under | 2 |

Table 4: Czech prepositions for `tool` and `medium` meaning and their English equivalents

## 6. Related Work

In this section, we touch on some of the approaches to circumstantial semantics found in other corpus development projects with detailed linguistic annotations.

In the framework of the **Uniform Meaning Representation** (Van Gysel et al., 2021), the circumstantial meanings are solved on the sentence-level representation which is adapted from the Abstract Meaning Representation project (Banarescu et al., 2013). The representation of a sentence is formally a graph where the nodes represent semantic concepts and edges represent relations (semantic roles). Approximately thirty circumstantial meanings are distinguished through the use of semantic roles: *temporal*, *place*, *goal*, *direction*, *duration*, *reason*, *quant*, etc. (some of them are understood as participants, others as non-participants). The semantic roles roughly correspond to our category of functor and there is no fine-grained classification at the level of our subfunctors. However, in some cases the semantic roles are more granular and correspond to our subfunctors rather than to the functors; e.g., the semantic roles *instrument*, *material*, and *medium* (with *language* subtype) correspond to the subfunctors we defined for the `MEANS` functor (cf. Sect. 4.2). It is not clear by what principle the semantic roles are established; why circumstants in the sentences *I read it in the newspaper* and *She talked to him in French* are distinguished by different labels (*medium* vs. *language*) and in the sentences *He decorated the room in a creative way* and *Lindbergh crossed the Atlantic in the Spirit of St. Louis*, the circumstants are not distinguished (both are labelled as *manner*). Prepositions expressing the subtle semantic distinctions such as *after* in *after the war*, *next to* in *next to me* are captured as a concept which heads the entity concept. Except in few cases, no generalization is made for modifiers of the same meaning.

Under a formal semantic approach the **Parallel Meaning Bank** (Abzianidze et al., 2017) is developed. It is a project of semantically annotated parallel corpus for English, German, Dutch and Italian, built on syntactic parsing with Combinatory Categorial Grammar (Steedman, 2001). The set of the semantic roles (for verb participants) coincide with those of VerbNet (Kipper et al., 2008), circumstantial meanings are described by semantic role labels such as *Destination*, *Goal*, *Instrument*, *Location*, *Manner*, *Material*, *Time*. The set is similar to the one in the UMR project. For some temporal, spatial and quantitative subtle meanings, a set of operators is used, e.g. $X \approx Y$ is for *approximately equal to*, $X \prec Y$ for *temporally precedes*, $X/Y$ is for *spatial above*, etc.

In incorporating the formal aspect into the compilation of set of meanings, the **Semantic Network of Adposition and Case Supersenses** (SNACS) project (Schneider et al., 2018) is the closest to our project. However, it is primarily focused only on prepositions. In SNACS, the authors aim to create a semantic classification of prepositional phrases. They distinguish 52 so-called supersenses, which are organized into a multi-level hierarchy. At the highest level, circumstances, participants and configurations (noun attributes) are distinguished. The set of labels for circumstances and participants is similar to the projects described above (it is also inspired by VerbNet), with the difference that the supersenses are hierarchically organized (e.g. *Temporal* label is superior to *Time*, *Duration*, *Frequency* and *Interval* labels). A gold-standard corpus STRESULE (of 55,575 tokens) was built with all tokens of prepositions disambiguated. There is also Xposition, an online platform (Gessler et al., 2022) containing annotation guidelines, documentation and annotated corpora across languages, based on the SNACS project. The proposed categories of supersenses again roughly correspond to the functors in PDT, as also compared in Scivetti and Schneider (2023).

Compared to our project, the semantic classification of circumstances in the projects mentioned above is coarse-grained, if compared to the level of our subfunctors. A fine-grained classification is sporadic, not systematic and comprehensive. Compared with the projects mentioned, our approach is characterized especially by the focus on the way how the given language in its structure reflects the reality, and on a detailed corpus-based research. A set of fine-grained circumstantial meanings is the result of a detailed analysis of the corpus material. Categories are not proposed "accidentally" but reflect the structure and meanings formally fixed in the given language. In our project, we benefit from the fact that the coarse-grained disambiguation of meanings has already been manually verified on a large amount of data, and the fine-grained classification can be proposed step by step for each general meaning separately.

## 7. Conclusion

In summary, this contribution introduces a project focused on the fine-grained classification of circumstantial meanings, the proposal of which is subsequently used to enrich the annotation in the Prague Dependency Treebank corpora. Compared to similar projects in computational linguistics, the unique aspect of this project is its focus on the relation between meaning and the forms used to express it. We do not describe the world, we describe the language we use to describe this world. The project is grounded in the analysis of a significant amount of valuable examples provided by the Prague Dependency Treebank framework. The methodology and principles of the presented approach are elaborated in detail. The working process is documented by two case

studies. The fine-grained classification of circumstantial meanings is carried out for the Czech language, but the methodology and principles used are language independent. The contribution also addresses the question of linguistic universality and specificity through a comparison with English. To the best of our knowledge, a similar corpus-based and corpus-verified elaborate classification of circumstantial meanings has not yet been proposed in any annotation project.

## 8. Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA23-05238S. 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).

## 9. Bibliographical 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 (Volume 1: Long Papers)*, 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: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. 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, Volume 1*, 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.

František Daneš and Miroslav Dokulil. 1958. *K tzv. významové a mluvnické stavbě věty*. [On the So-called Semantic and Grammatical Structure of the Sentence]. ČSAV, Prague, Czech Republic.

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, 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.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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

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

Markéta Lopatková and Jarmila Panevová. 2005. Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, pages 83–92. Veda, Bratislava, Slovakia.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical report, Prague, Czech Republic.

Marie Mikulová and Jarmila Panevová. 2021. *Formy a funkce okolnostních určení v češtině. Určení prostorová a časová*. [Forms and Functions of Circumstants in Czech. Spatial and Temporal Determinations]. Charles University, Prague, Czech Republic.

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.

Jaroslav Peregrin. 2003. *Filosofie a jazyk*. [Philosophy and Language]. Triton, Prague, Czech Republic.

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 (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. 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*, Nancy, France.

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.

Mark Steedman. 2001. *The Syntactic Process*. MIT Press, Cambridge, Massachusetts, USA.

Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, and Jan Hajič. 2022. Making a Semantic Event-type Ontology Multilingual. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1332–1334, Marseille, France. European Language Resources Association.

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 (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 96–105, Punta Cana, Dominican Republic. Association for Computational Linguistics.

## 10. Language Resource References

Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Fučíková, Eva and Hajičová, Eva and Havelka, Jiří and Hlaváčová, Jaroslava and Homola, Petr and Ircing, Pavel and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mareček, David and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Novák, Michal and Pajas, Petr and Panevová, Jarmila and Peterek, Nino and Poláková, Lucie and Popel, Martin and Popelka, Jan and Romportl, Jan and Rysová, Magdaléna and Semecký, Jiří and Sgall, Petr and Spoustová, Johanka and Straka, Milan and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Šindlerová, Jana and Štěpánek, Jan and Štěpánková, Barbora and Toman, Josef and Urešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and Žabokrtský, Zdeněk. 2020. *Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic; http://hdl.handle.net/11234/1-3185.

Hajič, Jan and Hajičová, Eva and Panevová, Jarmila and Sgall, Petr and Cinková, Silvie and Fučíková, Eva and Mikulová, Marie and Pajas, Petr and Popelka, Jan and Semecký, Jiří and Šindlerová, Jana and Štěpánek, Jan and Toman, Josef and Urešová, Zdeňka and Žabokrtský, Zdeněk. 2012. *Prague Czech-English Dependency Treebank 2.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic; http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.

Mikulová, Marie and Bejček, Eduard. 2017. *ForFun 1.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic; http://hdl.handle.net/11234/1-2542.