The Relation of Form and Function in Linguistic Theory and in a Multi-layer Treebank

Eduard Bejček Eva Hajičová Marie Mikulová Jarmila Panevová
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{bejcek,hajicova,mikulova,panevova}@ufal.mff.cuni.cz

Abstract

The aim of our contribution is to introduce a database of linguistic forms and their functions built with the use of the multi-layer annotated corpora of Czech, the Prague Dependency Treebanks. The purpose of the Prague Database of Forms and Functions (ForFun) is to help the linguists to study the form-function relation, which we assume to be one of the principal tasks of both theoretical linguistics and natural language processing. We will also demonstrate possibilities of the exploitation of the ForFun database.

1 Introduction

The study of the relation of (linguistic) forms and their functions or meanings is one of the fundamental tasks of linguistics, with important implications for natural language understanding. As Katz (1966, p. 100) says, to understand the ability of natural languages to serve as an instrument to the communication of thoughts and ideas we must understand what it is that permits those who speak them consistently to connect the right sounds with the right meanings. This, however, is obviously not an easy task as the relation between form and function is a many-to-many relation. At present, the availability of richly annotated corpora helps the linguist to analyze the given relation in its variety, and it is a challenging task to provide linguists with useful tools for their study.

One of the most useful types of corpora for this task are treebanks based on a stratificational (multi-layer) approach, where the form-function relation may be understood as a relation between units of two layers of the system. The aim of our contribution is to introduce a database of language forms and their linguistic functions built with the use of the multi-layer annotated corpora of Czech, the Prague Dependency Treebanks (PDTs), with the purpose to help the linguists to study the form-function relation. We offer a new tool ForFun which gives a possibility to search in a user-friendly way all forms (almost 1 500 items) used in PDTs for particular functions and vice versa to look up all functions (66 items) expressed by the particular forms.

The research question we follow by constructing the database and the new tool can be illustrated e.g. by the example of the Czech preposition *po* + Locative case of a noun (translated to English as *along, on, about, at, ...* + noun) in Figure 1. The blue colour indicates the forms, the pink colour the functions, identified in the PDTs by the functors attached to the nodes representing the given item (see below Section 2).[1] The prepositional case *po* + Locative (see the inner circle) may express the following eight functions (see the middle circle): `TWHEN` (when), `THL` (how long), `ORIG` (origin), `MEANS`, `MANN` (manner), `EXT` (extent), `DIR2` (direction which way), `DPHR` (idiomatic meaning). Each of these functions, in turn, may be expressed by a number of forms (see the outer circle) one of which is *po* + Locative. Thus for example, the function labelled `THL` (how long) may be expressed by an adverb, or Accusative of a

[1] Throughout the paper, we use the term functor for the label of the type of the dependency relation between the governor and its dependent; in the dependency tree structure representing the sentence on the deep (underlying, tectogrammatical; see Section 2) layer this label is a part of the complex label attached to the dependent node. The term prepositional case is used for a combination of a preposition and a noun or a nominal group in a morphological case. In the figures and tables, morphological cases are indicated by numbers, i.e. 2 for Genitive, 3 for Dative, 4 for Accusative, 6 for Locative, 7 for Instrumental. When the noun or nominal group is not accompanied by a preposition, we use the term prepositionless case.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 56–63,
Prague, Czech Republic, January 23–24, 2018. Distributed under a CC-BY 4.0 licence. 2017.

Keywords: form-function relation, prepositions, functors

noun (prepositionless case), or prepositional cases *za* + Genitive, *za* + Accusative, *po* + Accusative, and, of course, by the already mentioned *po* + Locative.

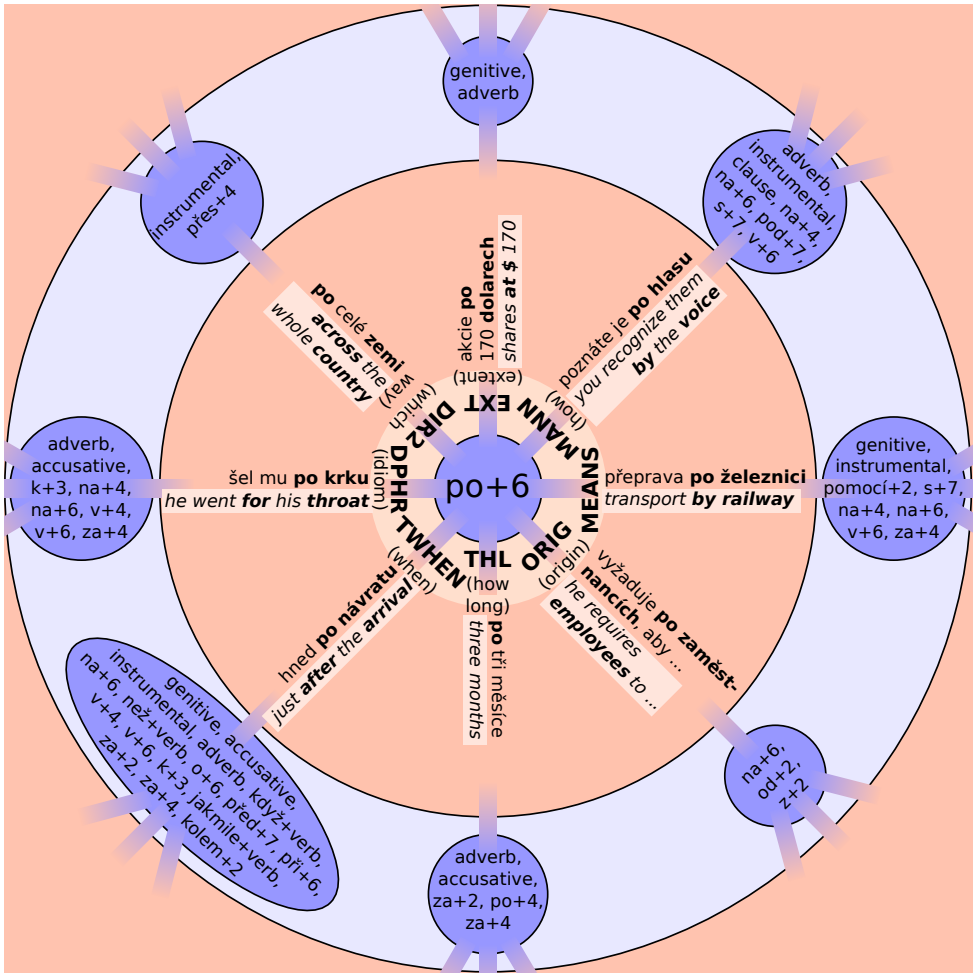

Figure 1: Many-to-many relation between forms and functions: prepositional case *po* + Locative.

2 Multi-layer Architecture of Prague Dependency Treebanks

PDTs (on which our ForFun database is based) are complex linguistically motivated treebanks based on the dependency syntactic theory of the Functional Generative Description (see Sgall et al. 1986). The original annotation scheme has the following multi-layer architecture:[2]

- **morphological layer**: all tokens of the sentence get a lemma and a (disambiguated) morphological tag,
- **surface syntax layer** (analytical): a dependency tree capturing surface syntactic relations such as subject, object, adverbial; a (structural) tag reflecting these relations is attached to the nodes as one component of their (complex) labels,
- **deep syntax layer** (tectogrammatical) capturing the semantico-syntactic relations: on this layer, the dependency structure of a sentence is a tree consisting of nodes only for autonomous meaningful units (function words such as prepositions, subordinating conjunctions, auxiliary verbs etc. are not represented as separate nodes in the structure, their contribution to the meaning of the sentence is captured within the complex labels of the autonomous units). The types of dependency relations are captured by means of the so-called functors.

[2]The PDTs annotation scenario is described in detail in Mikulová et al. (2006) and Hajič et al. (2017).

Functors (66 in total) are classified according to different criteria. The basic subdivision is based on the the valency criterion, which divides functors into the argument functors and adjunct functors. There are five arguments: Actor/Bearer (ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF). The repertory of adjuncts is much larger than that of arguments. Their set might be divided into several subclasses, such as temporal (TWHEN for "when?", TSIN for "since when?", TTILL for "till when?", THL for "how long?", THO for "how often?", etc.), local (LOC for "where?", DIR1 for "where from?", DIR2 for "which way?", DIR3 for "where to?"), causal (such as CAUS for "cause", AIM for "in order to", COND for "condition", etc.), and other adjuncts (MANN for general "manner", ACMP for "accompaniment", EXT for "extent", MEANS for "means or instrument", INTF for "intensifier", BEN for "benefactor", RSTR for "attribute", etc.). For a full list of all dependency relations and their labels see Mikulová et al. (2006).

For the ForFun database, we use the annotations of the nodes on the deep syntactic layer and their counterparts on the morphological layer, which has made it possible to retrieve the relations between functions (expressed on the deep level by functors) and forms and vice versa.

3 List of available Prague Dependency Treebanks

For Czech, the following four treebanks are now available, each of them contains data of a different source: the Prague Dependency Treebank 3.0,[3] the Prague Czech-English Dependency Treebank 2.0,[4] the Prague Dependency Treebank of Spoken Czech 2.0,[5] and the PDT-Faust corpus.[6]

	PDT 3.0	PCEDT 2.0	PDTSC 2.0	Faust	Total
Tokens	833 195	1 162 072	742 257	33 772	2 771 296
Sentences	49 431	49 208	73 835	3 000	175 474

Table 1: Volume of data in Prague Depencency Treebanks

It is obvious (see Table 1) that the Prague Dependency Treebank family provides rich language data for our purpose, i.e. for the study of the relation of forms and their functions since every content word there is assigned one of those 66 functors. Altogether, the treebanks contain around 180 000 sentences with their morphological, syntactic and semantic annotation.

4 Prague Database of Forms and Functions

ForFun 1.0—Prague Database of Forms and Functions—is a rich database of syntactic functions and their formal realizations with a large amount of examples coming from both written and spoken Czech texts. Since the database is extracted from the PDTs (see Section 3), it takes over the list of syntactic functions as well as the terminology (they are called *functors*). ForFun is provided as a digital open source accessible to all scholars via the LINDAT/CLARIN repository.[7]

4.1 Design

We have already mentioned that in general the relation between forms and functions is a many-to-many relation. As such, it has to be explored from both sides: a given form has several functions and any of these functions may again be realized by several forms (the given one among them). When such relations

[3] https://ufal.mff.cuni.cz/prague-dependency-treebank
In the PDT 3.0 (see Hajič et al., 2006, Bejček et al., 2013), the data consist of articles from Czech daily newspapers.

[4] https://ufal.mff.cuni.cz/pcedt2.0/
In the parallel PCEDT 2.0 (see Hajič et al., 2012), the English part consists of the Wall Street Journal sections of the Penn Treebank (Marcus et al., 1993), and the Czech part, which is used in ForFun, was manually translated from the English original.

[5] https://ufal.mff.cuni.cz/pdtsc2.0
The PDTSC 2.0 (see Mikulová et al., 2017b) contains dialogs from the Malach project (https://ufal.mff.cuni.cz/cvhm/vha-info.html, slightly moderated testimonies of Holocaust survivors) and from the Companions project (http://cordis.europa.eu/project/rcn/96289_en.html, two participants chat over a collection of photographs).

[6] PDT-Faust is a small treebank containing short segments (very often with vulgar content) typed in by various users on the reverso.net webpage for translation.

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

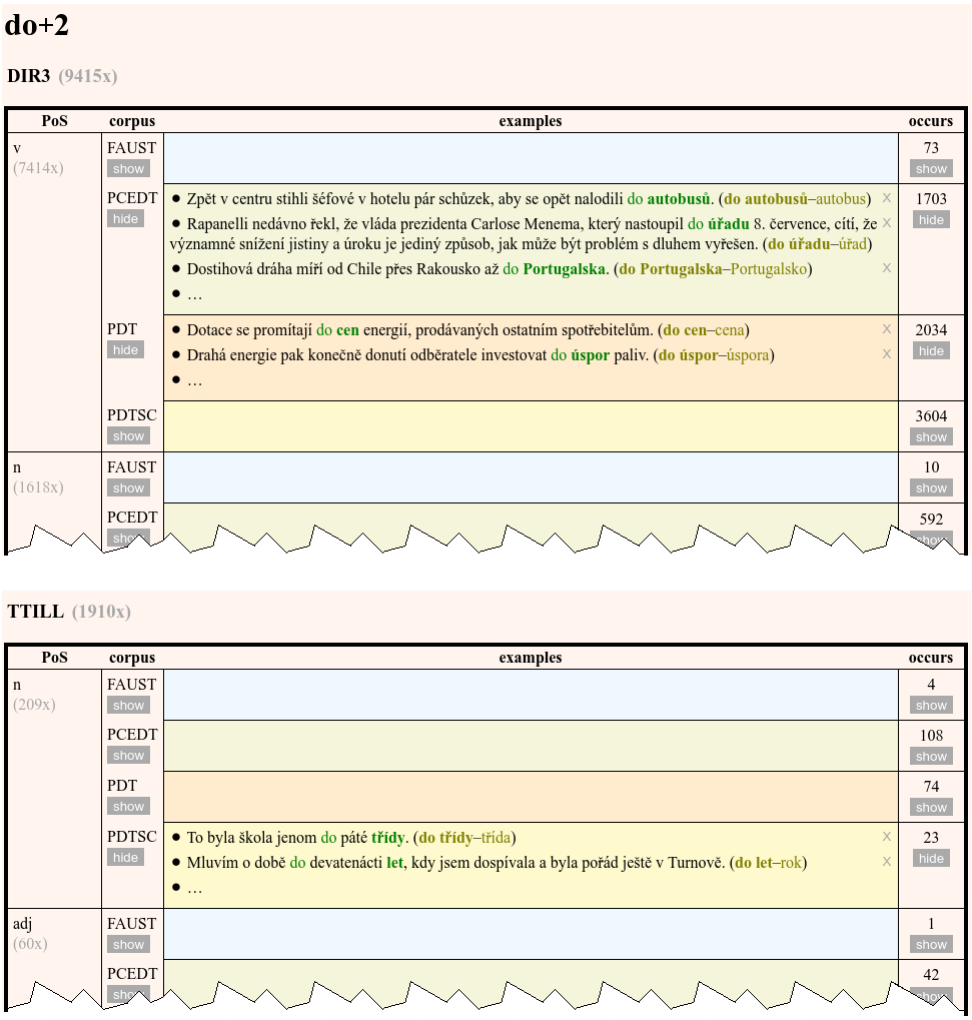

Figure 2: A screenshot of the ForFun web interface: From Form to Function.

have to be explored, ForFun is a perfect choice, since it is designed exactly for this kind of traversing through data.

Although the annotated example sentences are the same, they can be retrieved by asking either for their forms or for their functions. The ForFun database provides two entry points (cf. Figures 2 and 3):

- The user can choose one of almost 1 500 formal realizations of sentence units (i.e. prepositionless and prepositional cases, subordinated and coordinate conjunctions, adverbs, infinitive and finite verb forms, etc.) and obtains all functions it can represent.
- The user can choose one of 66 syntactic functions (i.e. `LOC`, `TTILL`, `CAUS` etc.) and obtains all forms used to express it.

The view can be always switched from a list of forms to a list of functions of one of them and vice versa.

For each form-function relation there are plenty of examples in the form of a sentence with the highlighted expression representing the relation. All these examples are sorted by various criteria:

- the word class of the parent node,
- the particular forms for the function or particular functions for the form, and
- the source of text data (written, spoken, translated texts and texts from internet users).

The number of examples available in the database is displayed for each pair form+functor, or functor+word class, each combination functor+form+word class and each specified 4-combination (form+functor+word class+source), see Figures 2 and 3. Either first ten examples or all of them are displayed on demand.

On top of that, examples can be also first filtered by their source, which allows the user to hide e.g. all

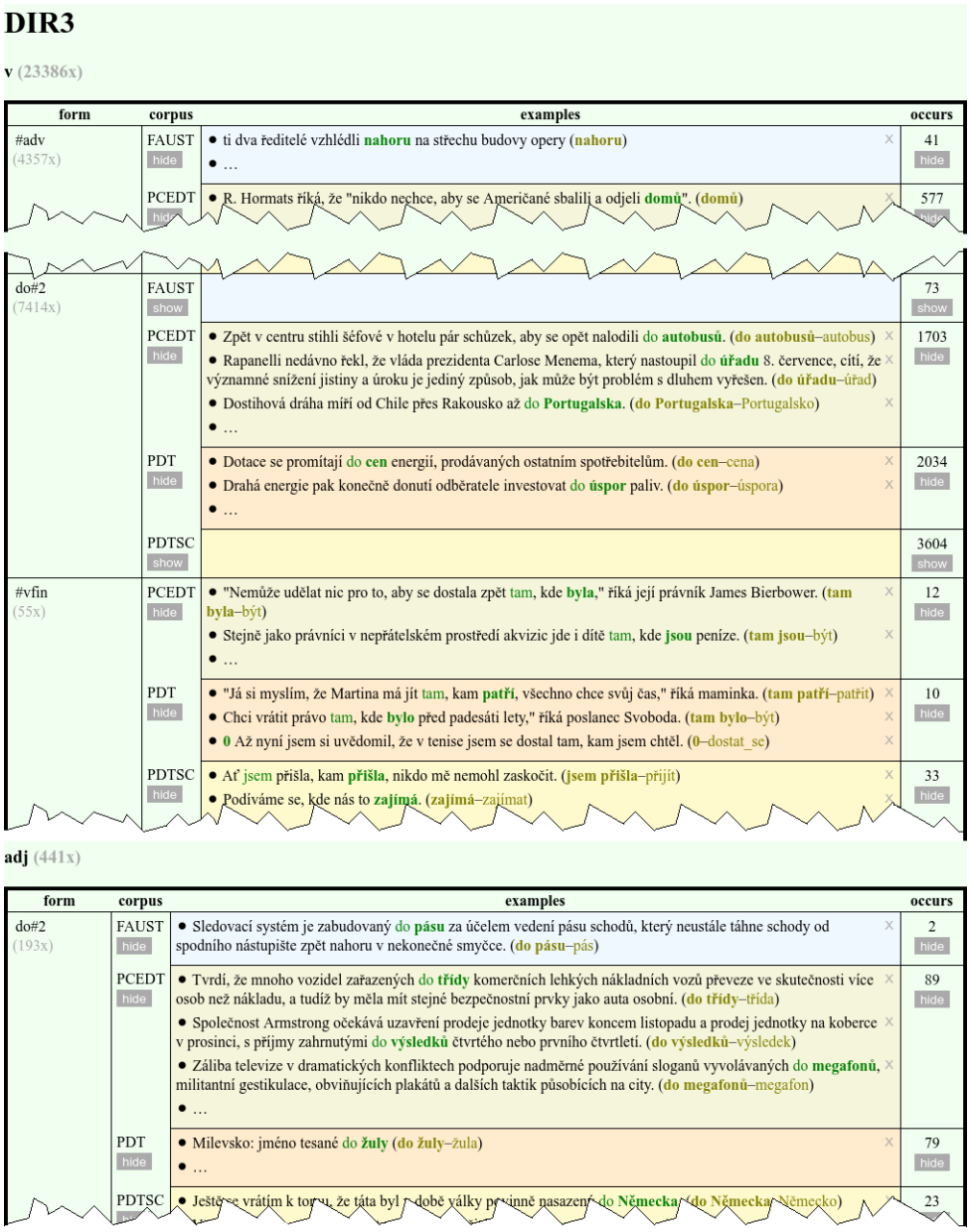

Figure 3: A screenshot of the ForFun web interface: From Function to Form.

forms used only in the spoken language.

An illustration of how the result of user's search for the functions of the prepositional case *do* + Genitive looks like is given in Figure 2. In the upper part, there are 9 415 occurrences in all PDTs of the form *do* + Genitive representing the functor DIR3. The occurrences of *do* + Genitive are divided according to their heads (be it a v(erb) or a n(oun), see the first column); their distribution within particular treebank is given in the second column followed by real examples from the corresponding treebank. A few of them are displayed on demand whereas many (see the last column) stay hidden. In the lower part of Figure 2, the same form *do* + Genitive in the function TTILL is exemplified in the same style.[8] For the opposite direction "from function to form" see Figure 3, where (among others) the same sentences for *do* + Genitive as the functor DIR3 can be found searching for all representations of the functor DIR3. Other forms include a finite verb (#vfin) or an adverbial.

[8] Figure 2 presents only a part of the full response obtained from the ForFun database for the given query. The other functions of *do* + Genitive (PAT, EXT, EFF and others) are also not included in this shortened sample.

4.2 Volume

The database contains 2.2 million examples altogether for all forms (and the same number from the function point of view), split approx. 3:1 between written and spoken text (see Table 2). Each example is one sentence long.[9] They can be examined from the function side (66 functors) or the form side (1 469 forms). All examples are split into 13.5 thousand of 4-combinations, each with 163 examples in average.

examples from written text	1 608 061
examples from spoken text	593 400
examples altogether	2 201 461
number of functions	66
number of forms	1 469
number of 4-combinations	13 514
avg. examples for a function	33 355
avg. examples for a form	1 500
avg. examples for a 4-combination	163
max. number of examples for a function	490 121
max. number of examples for a form	370 586
max. number of examples for a 4-combination	97 469

Table 2: Volume of the ForFun database

While the average number is high, median is only two examples. The reason is that there is a long tail of 4-combinations used very rarely. These occurences with very low frequencies in the data are one of the main benefits of the large volume of database, but they have to be used carefully. Every result has to be always understood solely as an input for a subsequent research, as the ForFun database may contain errors (caused by annotators as well as speakers/writers) considering its volume.

5 What Can We Find Out about Form-Function Relations in the ForFun Database?

To display the richness of the material we work with, we present several examples connected with the studies of the form-function relation what the user can find out in the ForFun database.

prep.	number	list of functors
na+4	42	ACT ADDR AIM APP ATT BEN CAUS COMPL COND CPHR CPR CRIT DIFF DIR1 DIR3 DPHR EFF EXT ID INTF INTT LOC MANN MAT MEANS MOD ORIG PAT PREC REG RESL RESTR RHEM RSTR SUBS TFHL TFRWH THL TOWH TPAR TTILL TWHEN
v+6	36	ACMP ACT AIM APP ATT BEN CAUS COMPL COND CPR CRIT DENOM DIR2 DIR3 DPHR EFF EXT ID LOC MANN MAT MEANS MOD PAT PREC REG RESL RESTR RHEM RSTR SUBS TFHL THL THO TPAR TWHEN
k+3	34	ACMP ACT ADDR AIM APP ATT BEN CAUS COMPL CPHR CRIT DIR1 DIR2 DIR3 DPHR EFF EXT ID INTT LOC MANN PAR PAT PREC REG RESL RESTR RHEM RSTR TOWH TPAR TSIN TTILL TWHEN

Table 3: The prepositional cases with the highest number of functions.

5.1 Multi-functionality of Forms

A rather straightforward use of the ForFun database is to retrieve which functions can be expressed by the particular form. Table 3 contains three prepositional cases with the highest number of functions they

[9] One sentence typically contains many different functions and serves for many examples (once for each of its parts).

express: *na* + Accusative, *v* + Locative and *k* + Dative. The *po* + Locative case from Figure 1 with 32 functions would be the seventh prepositional case in this Table.

5.2 Absolute Frequency of Forms and Functions (in both written and spoken texts)

An observation of frequency has an important place in the description of language because it quantifies linguistic choices made by speakers and writers. Theoretical statements are often of a little value for generalizations about language use unless they can be corroborated by observations of frequency.

For each form and function, ForFun provides information about absolute frequency in all the PDTs as well as in each corpus separately. The users can search quickly and in a user-friendly way which formal means are the most frequent in Czech sentences and which ones are rarely used. (See Table 4 for five most frequent prepositional cases in Czech in comparison with the class of adverbs and the clause with the conjunction *že* [*that*].) They can find out the distribution of a particular function (various arguments or adjuncts) in the sentences. For both forms and functions, they can compare their absolute frequencies in written and spoken texts.

form	occurences
v+6	51 682
na+4	22 444
s+7	19 747
z+2	19 502
na+6	17 870
adverb	93 824
že[that]+verb	26 831

Table 4: The most frequent prepositional cases

5.3 Material for Detailed Linguistic Studies

In addition to valuable statistical data, the ForFun database provides an extremely rich material for detailed linguistic studies of individual language phenomena and for their description and classification. One of the first linguistic studies based on the database is the analysis and subclassification of the original functors denoting space (Mikulová et al., 2017a).

6 Conclusion

The ForFun database has been built as a rich and user-friendly resource for those researchers who (want to) use corpora in their everyday work and look for various occurrences of specific forms or patterns in relation to their syntactic functions etc. but they are not interested or just do not need to deal with various technical, formal and annotation issues. ForFun brings a rich and complex annotation in PDTs based on a sound linguistic theory closer to common researchers. It will be further developed, though it should be borne in mind that it is designed to provide only a limited number of most useful features, rather than a full interface to everything PDTs can offer. There are other complex tools for that[10] and ForFun does not aim to substitute them. In its simplicity and clarity, it is a user-friendly source of examples for various explorations especially in syntax.

Acknowledgments

The research reported in the paper has been supported by the Czech Science Foundation under the projects GA17-12624S and GA17-07313S and by the LINDAT/CLARIN project of Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). This work has been using language resources developed, stored and distributed by the latter project (LM2015071).

[10] E.g. PML Tree Query `https://lindat.mff.cuni.cz/services/pmltq/`, INESS Search `http://clarino.uib.no/iness`, etc.

References

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0. Data, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. *Handbook on Linguistic Annotation*, Springer Verlag, Dordrecht, Netherlands, chapter Prague Dependency Treebank, pages 555–594.

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 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association, Istanbul, Turkey, pages 3153–3160. https://aclanthology.info/pdf/L/L12/L12-1280.pdf.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová, and Zdeňka Urešová. 2006. Prague Dependency Treebank 2.0 (LDC2006T01).

Jerrold J. Katz. 1966. *The philosophy of language*. Studies in languages. Harper & Row, New York.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330. https://aclanthology.info/pdf/J/J93/J93-2004.pdf.

Marie Mikulová, Eduard Bejček, Veronika Kolářová, and Jarmila Panevová. 2017a. Subcategorization of adverbial meanings based on corpus data. *Journal of Linguistics / Jazykovedný časopis* 68(2):268–277.

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 30, ÚFAL MFF UK, Prague, Czech Rep.

Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jan Štěpánek, and Jan Hajič. 2017b. PDTSC 2.0 – spoken corpus with rich multi-layer structural annotation. In *Text, Speech, and Dialogue 20th International Conference, TSD 2017*. Charles University, Springer International Publishing, Cham / Heidelberg / New York / Dordrecht / London, Lecture Notes in Computer Science, pages 129–137.

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.