ForFun 1.0: Prague Database of Forms and Functions An Invaluable Resource for Linguistic Research

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

Abstract

In this paper, we introduce the first version of ForFun, Prague Database of Forms and Functions, as an invaluable resource for profound linguistic research, particularly in describing syntactic functions and their formal realizations. ForFun is built with the use of already existing richly syntactically annotated corpora, collectively called Prague Dependency Treebanks. ForFun brings this complex annotation of Czech sentences closer to researchers. We demonstrate that ForFun 1.0 provides valuable and rich material allowing to elaborate various syntactic issues in depth. We believe that nowadays when corpus linguistics differs from traditional linguistics in its insistence on a systematic study of authentic examples of language in use, our database will contribute to the comprehensive syntactic description.
Keywords: language resource, dependency syntax, semantic labeling, surface form, digital humanities

1. Motivation

What is the difference between location expressions "walk in King Street", "walk on King Street", and "walk along King Street"? Should we use a different preposition when talking about destination rather than about direction and location? Or more precisely, what function does the preposition "on" perform in contrast to the preposition "along" and which forms can express destination? Is the same form used in both spoken and written text? Is there any bias towards one form in translated text? For Czech, the answers can be found in a new database for inspecting thousands of real examples categorized by their *form* (e.g. by a prepositional case) as well as by their deep syntactic *function*.

2. Introduction

In this paper, we present the first version of ForFun, Prague Database of Forms and Functions, as an invaluable resource for different linguistic issues, particularly for the description of syntactic functions and their formal realizations. It takes advantage of several richly syntactically annotated corpora, collectively called Prague Dependency Treebanks (PDTs in the sequel) that have already been developed in Prague. Altogether, the treebanks contain around 180,000 sentences with their morphological, syntactic and semantic annotation. The ForFun database draws on the complex linguistic annotation of these corpora, arranges selected morphological and syntactical annotation into new shape, and offers a user-friendly access to a large resource of real examples.

3. Related Work

There is a wide range of corpora with rich linguistic annotation, e.g., Penn Treebank (Marcus et al., 1999), its successors PropBank (Kingsbury and Palmer, 2002) and NomBank (Meyers et al., 2004); for German, there is Tiger (Brants et al., 2002) and Salsa (Burchardt et al., 2006), and many others. The ForFun database is unique in that it is compiled from four different treebanks of Czech, uniformly annotated using the same scenario, with data coming from text, speech and Internet sources. It offers a really large material with the deep syntactic manual annotation which is well and comprehensibly sorted and easily accessible.

4. Data Resources

The database ForFun is extracted from PDTs. PDTs are the complex linguistically motivated treebanks based on the dependency syntactic theory, which provide interlinked hierarchical layers of standoff annotation. Their annotation scenario is described in detail e.g. in Hajič et al. (2017) and Mikulová et al. (2006).

The **Prague Dependency Treebank** version 3.5[1] (Hajič et al., 2018) is the newest edition of the core Prague Dependency Treebank published in 2006 (Hajič et al., 2006). The data consist of articles from Czech daily newspapers.

A slightly modified scenario was then used for the annotation of the Prague Czech-English Dependency Treebank, the Prague Dependency Treebank of Spoken Czech, and the PDT-Faust corpus. In contrast to the original project of PDT, in these treebanks, the morphological and surface syntactic annotations were done automatically, and the manually annotated deep syntactic layer does not contain annotation of information structure and some other special annotations. However, the annotation of functors (see sect. 5), which we use for building the ForFun database, has been done manually in all four treebanks.

The **Prague Czech-English Dependency Treebank** version 2.0[2] (Hajič et al., 2012), (Hajič et al., 2012) is a manually parsed Czech-English parallel corpus. The English part consists of the Wall Street Journal sections of the Penn Treebank (Marcus et al., 1999). The Czech part, which is used in the database, was manually translated from the English original.

The **Prague Dependency Treebank of Spoken Czech** version 2.0[3] (Mikulová et al., 2017b), (Mikulová et al., 2017)

[1] http://ufal.mff.cuni.cz/pdt3.5
[2] https://ufal.mff.cuni.cz/pcedt2.0/
[3] https://ufal.mff.cuni.cz/pdtsc2.0

contains slightly moderated testimonies of Holocaust survivors from the Malach project corpus[4] and dialogues (two participants chat over a collection of photographs) recorded for the EC-funded Companions project.[5]
The **PDT-Faust** corpus is a small treebank containing short segments (very often with vulgar content) typed in by various users on the `reverso.net` webpage for translation.

5. Functions and Forms

An exploration of what formal means (forms) are used for expressing various syntactic functions is one of the main tasks in syntax. The approach "from function to form" (corresponding to generation in computational linguistics) is the basic one. The reversed process – "from form to function" (corresponding to analysis) – describing conditions in which a partial form has the given function and not another one is also not omitted in syntactic research.
The basic semiotic relation between the function and form (known from the Saussure´s structural linguistics (Saussure, 1916) as the relation between "signifié" and "signifiant") is in the PDTs framework (called the Functional Generative Description, see Sgall et al. (1986)) perceived as a relation between two language layers. Concerning the relation between syntactic functions and forms, we deal with the surface layer (for forms) and deep syntactic layer (for functions). The deep syntactic layer of PDTs represents the most complex linguistic annotation that combines syntax and semantics in the form of semantic labeling, co-reference annotation, and argument structure description based on a valency lexicon. The types of the (semantic) dependency relations are represented by the *functor* attribute attached to all nodes. Functors are classified according to different criteria. The basic subdivision is based on valency. The valency criterion divides functors into argument functors and adjunct functors. There are five argument functors: 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, spatial, causal, etc. Other relations such as e.g. relations between the members of coordination or between parts of multi-word expressions, are also labeled by functors. A shortened list of functors is presented in Table 1. For a full list of all dependency functions and their descriptions and labels see (Mikulová et al., 2006). The theoretical description of the valency theory and deep syntactic functions (as developed originally in the theoretical framework of Functional Generative Description and then applied in PDTs) is summarized mainly by Panevová (1974; 1998; 1999).
The lower layers of PDTs contain surface syntax and morphological annotation. Among others they contain information about the formal realization of sentence units (e.g., POS, grammatical cases) in the form of morphological tags assigned to all tokens.
There is a one-to-one correspondence between the tokens at the morphological layer and the nodes at the surface syntactic layer. But there is no such clear correspondence between the nodes at the surface syntactic layer and the deep syntactic layer. The nodes of the deep syntactic layer represent semantic units, i.e. one node for each content word together with its auxiliary words such as prepositions, conjunctions or auxiliary verbs. For example, the prepositional phrase "on street" is represented by one node with the lemma "street". To preserve the original information, nodes on the surface layer are explicitly referred to from this node. Thus there are two links from the node "street" to the surface layer: to the noun "street" and to the preposition "on". These links allow to combine information from different layers of the corpus. We take a big advantage of this linking in building the ForFun database.

Spatial functors		**Causal functors**	
`LOC`	where?	`CAUS`	cause
`DIR1`	where from	`AIM`	aim
`DIR2`	which way?	`CNSC`	concession
`DIR3`	where to?	`COND`	condition
		`INTT`	intention
Temporal functors			
`TWHEN`	when?	**Coordination relations**	
`TSIN`	since when?	`CONJ`	conjunction
`TTILL`	till when?	`ADVS`	adversative
`THL`	how long?	`CSQ`	consecution
`TFHL`	for how long?	`CONFR`	confrontation
`THO`	how often?	`DISJ`	disjunction
`TPAR`	during what time?	`GRAD`	gradation
`TFRWH`	from when?	`REAS`	reason
`TOWH`	to when?	`APPS`	apposition
Functors for manner		**Other functors**	
`MANN`	manner	`ACMP`	accompaniment
`CPR`	comparison	`INTF`	intensifier
`CRIT`	criterion	`BEN`	benefactor
`DIFF`	difference	`RHEM`	rhematizer
`EXT`	extent	`RSTR`	attribute

Table 1: Shortened list of functors (the total number is 66).

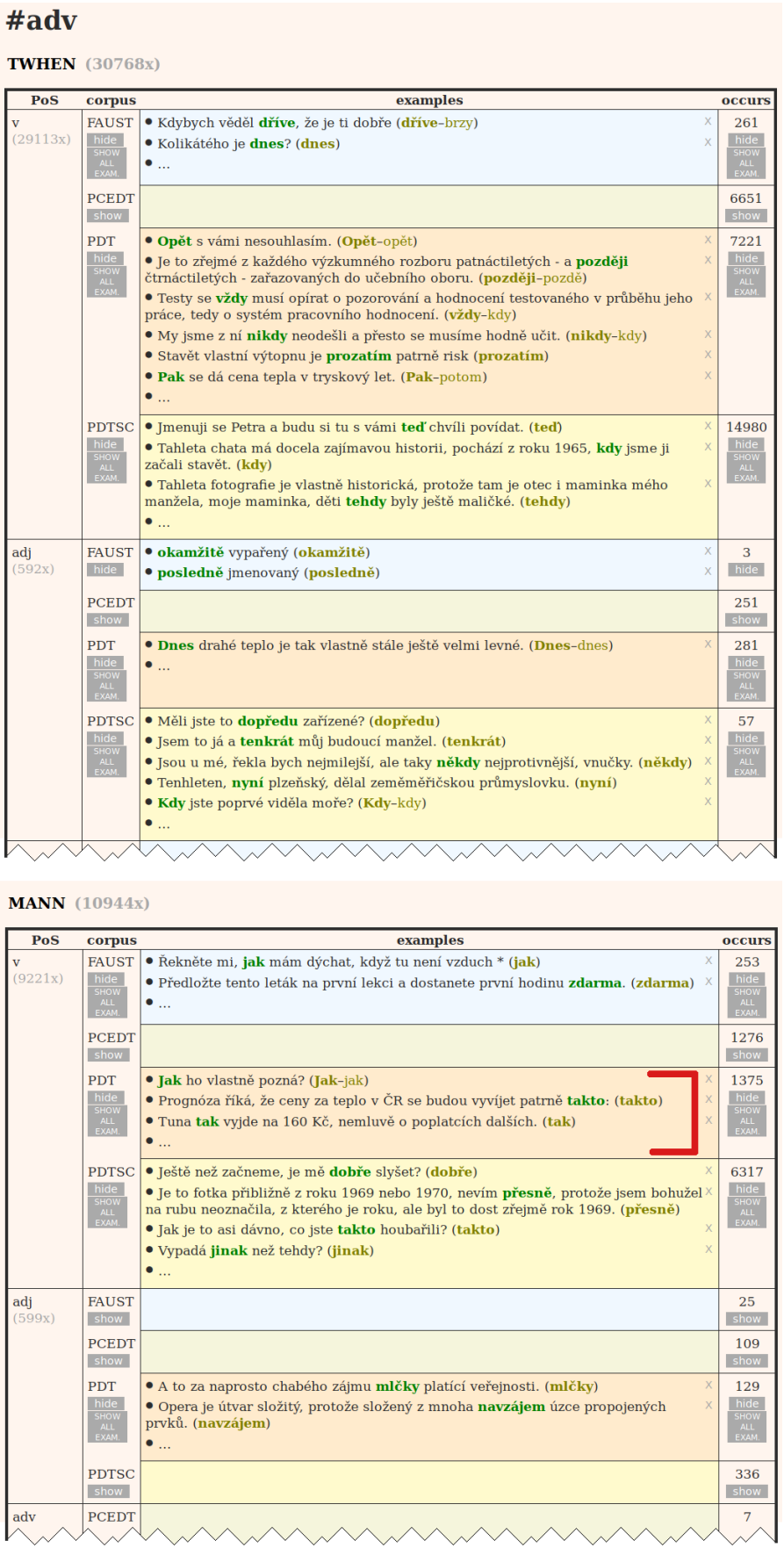
6. ForFun 1.0

Prague Database of Forms and Functions 1.0 (ForFun 1.0) 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. The database is extracted from PDTs (see Sect. 4) and it is provided as a digital open source accessible to all scholars via the LINDAT/CLARIN language resource open repository.[6]

6.1. Design

In language, one form can usually represent various functions, and one function can have several forms. Thus the distinction between form and function is a useful way to tackle two main syntactic approaches: "from function to form" and "from form to function".
The ForFun database is split in the same manner into two interconnected but reversed sets (cf. Figure 1 and Figure 2).

[4] `https://ufal.mff.cuni.cz/cvhm/vha-info.html`
[5] `http://cordis.europa.eu/project/rcn/96289_en.html`
[6] `http://hdl.handle.net/11234/1-2542`



#adv

TWHEN (30768x)

PoS	corpus	examples	occurs
v (29113x)	FAUST hide SHOW ALL EXAM.	• Kdybych věděl **dříve**, že je ti dobře (**dříve**–brzy) x • Kolikátého je **dnes**? (**dnes**) x • ...	261 hide SHOW ALL EXAM.
	PCEDT show		6651 show
	PDT hide SHOW ALL EXAM.	• **Opět** s vámi nesouhlasím. (**Opět**–opět) x • Je to zřejmé z každého výzkumného rozboru patnáctiletých - a **později** čtrnáctiletých - zařazovaných do učebního oboru. (**později**–pozdě) x • Testy se **vždy** musí opírat o pozorování a hodnocení testovaného v průběhu jeho práce, tedy o systém pracovního hodnocení. (**vždy**–kdy) x • My jsme z ní **nikdy** neodešli a přesto se musíme hodně učit. (**nikdy**–kdy) x • Stavět vlastní výtopnu je **prozatím** patrně risk (**prozatím**) x • **Pak** se dá cena tepla v tryskový let. (**Pak**–potom) x • ...	7221 hide SHOW ALL EXAM.
	PDTSC hide SHOW ALL EXAM.	• Jmenuji se Petra a budu si tu s vámi **teď** chvíli povídat. (**teď**) x • Tahleta chata má docela zajímavou historii, pochází z roku 1965, **kdy** jsme ji začali stavět. (**kdy**) x • Tahleta fotografie je vlastně historická, protože tam je otec i maminka mého manžela, moje maminka, děti **tehdy** byly ještě maličké. (**tehdy**) x • ...	14980 hide SHOW ALL EXAM.
adj (592x)	FAUST hide	• **okamžitě** vypařený (**okamžitě**) x • **posledně** jmenovaný (**posledně**) x	3 hide
	PCEDT show		251 show
	PDT hide SHOW ALL EXAM.	• **Dnes** drahé teplo je tak vlastně stále ještě velmi levné. (**Dnes**–dnes) x • ...	281 hide SHOW ALL EXAM.
	PDTSC hide SHOW ALL EXAM.	• Měli jste to **dopředu** zařízené? (**dopředu**) x • Jsem to já a **tenkrát** můj budoucí manžel. (**tenkrát**) x • Jsou u mé, řekla bych nejmilejší, ale taky **někdy** nejprotivnější, vnučky. (**někdy**) x • Tenhleten, **nyní** plzeňský, dělal zeměměřičskou průmyslovku. (**nyní**) x • **Kdy** jste poprvé viděla moře? (**Kdy**–kdy) x • ...	57 hide SHOW ALL EXAM.

MANN (10944x)

PoS	corpus	examples	occurs
v (9221x)	FAUST hide SHOW ALL EXAM.	• Řekněte mi, **jak** mám dýchat, když tu není vzduch * (**jak**) x • Předložte tento leták na první lekci a dostanete první hodinu **zdarma**. (**zdarma**) x • ...	253 hide SHOW ALL EXAM.
	PCEDT show		1276 show
	PDT hide SHOW ALL EXAM.	• **Jak** ho vlastně pozná? (**Jak**–jak) x • Prognóza říká, že ceny za teplo v ČR se budou vyvíjet patrně **takto**: (**takto**) x • Tuna **tak** vyjde na 160 Kč, nemluvě o poplatcích dalších. (**tak**) x • ...	1375 hide SHOW ALL EXAM.
	PDTSC hide SHOW ALL EXAM.	• Ještě než začneme, je mě **dobře** slyšet? (**dobře**) x • Je to fotka přibližně z roku 1969 nebo 1970, nevím **přesně**, protože jsem bohužel na rubu neoznačila, z kterého je roku, ale byl to dost zřejmě rok 1969. (**přesně**) x • Jak je to asi dávno, co jste **takto** houbařili? (**takto**) x • Vypadá **jinak** než tehdy? (**jinak**) x • ...	6317 hide SHOW ALL EXAM.
adj (599x)	FAUST show		25 show
	PCEDT show		109 show
	PDT hide SHOW ALL EXAM.	• A to za naprosto chabého zájmu **mlčky** platící veřejnosti. (**mlčky**) x • Opera je útvar složitý, protože složený z mnoha **navzájem** úzce propojených prvků. (**navzájem**) x • ...	129 hide SHOW ALL EXAM.
	PDTSC show		336 show
adv	PCEDT		7

Figure 1: A screenshot of the form adverb in ForFun. The figure presents only a part of the full response obtained from the ForFun database. Adverb can serve for as many as 55 functions (see also Table 2), two of them (`TWHEN` and `MANN`) are shown here.

Function		Examples	Raw Frequency
TWHEN	When?	*dnes* ‘today’; *hned* ‘immediately’; *pozdě* ‘late’; *nikdy* ‘never’	29113
LOC	Where?	*venku* ‘outside’; *doma* ‘at home’; *všude* ‘everywhere’; *dole* ‘down’	16251
MANN	How?	*krásně* ‘beautifully’; *dobře* ‘well’; *detailně* ‘in detail’; *trpělivě* ‘patiently’	9221
DIR3	Where to?	domů ‘home’; zpět ‘back’; jinam ‘elsewhere’; dovnitř ‘indoors’	4357
EXT	How much?	*příliš* ‘too much’; *vůbec* ‘not at all’; *úplně* ‘entirely’; *trošku* ‘a little’	3815
THO	How often?	*často* ‘often’; *občas* ‘sometimes’; *zřídka* ‘rarely’; *pravidelně* ‘regularly’	3211
THL	How long?	*ještě* ‘still’; *pořád* ‘all the time’	2721
TTILL	Till when?	*doposud* ‘heretofore’; *dodnes* ‘up to now’	762
CAUS	Why?	*náhodou* ‘accidentally’; *právem* ‘by right’	648
DIR1	Where from?	*odtud* ‘thereof’; *zdola* ‘from below’; *zprava* ‘from the right’	503
		⋮	
		another 43 functions	
		⋮	
MEANS	means	*koňmo* ‘on horseback’; *ručně* ‘manually’	52
TOWH	To when?	*nakonec* ‘finally’	4

Table 2: Functions of adverbs. Shortened list (total number of functions is 55) gained from ForFun. For each function some examples with translation and the number of examples in the database is given.

The **“from function to form”** set contains a list of all deep syntactic functions (66 functors altogether). When choosing one function type, the user can search for all forms that may represent that function. (See Figure 2 that shows a result of a search for a function “manner”.) For each function-form relation there are plenty of examples in the form of sentence with the highlighted expression representing the relation. All examples are sorted by various criteria:

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

The number of examples available in the database is always shown for each specified 4-combination (given form, functor, parent word class and source). Either the first ten or all examples are displayed on demand.

The **“from form to function”** set contains a long list (almost 1 500 items) of all formal realizations of particular sentence units that occur in PDTs: prepositionless cases, prepositional cases, subordinated and coordinate conjunctions, adverbs, infinitive and finite verb forms, etc. For any form (see Figure 1 for adverbs), there are again plenty of examples sorted by function, word class of the parent node, and the source of text data, always with the frequency in the data. In both sets, examples can be also filtered by their source, which allows the user to hide e.g. all forms used only in spoken language or use only sentences from written corpora.

An illustration of how the result of user search for all functions of an adverb phrase looks like is given in Figure 1. In the upper part, there are examples of the form “#adv” (meaning either an adverb phrase or an adverb as a word) representing the time expression “when” (i.e. the functor TWHEN); there are 30 768 occurrences available. The occurrences of adverb form are divided according to their syntactic parents (be it a v(erb), adj(ective), adv(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 sample of them is displayed on demand right in the table whereas many others (see the last column for their numbers) stay hidden and can be displayed in a full list.

In the lower part of Figure 1, the same form “#adv” is exemplified in the same style as an expression of manner (i.e. the functor MANN, third most frequent). See Table 2 for functions represented by adverbs other than TWHEN and MANN.

For the opposite direction (“from function to form”), see Figure 2, where (among others) the same sentences (for adverb form) can be found when searching for all representations of the functor MANN (see the sentence *Jak ho vlastně pozná?* and others in both Figures 1 and 2). Other forms are less frequent and include a preposition *na* together with either a genitive, accusative or locative case[7] or a preposition *mimo* with an accusative case etc.; see also Table 4.

6.2. 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 3). Each example is one sentence long.[8] 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 on average. There is also a 4-combination with almost 100 000 examples. Maximum number of examples for a function is 490 000 across all forms and corpus sources (function RSTR). Maximum number of examples for a form is 370 000 (nominative case).

While the average number is high, the median is only two examples. The reason is that there is a long tail of 4-combi-

[7] Morphological cases in ForFun are indicated by numbers, thus forms mentioned above are shortened as *na#2*, *na#4* and *na#6*.

[8] One sentence typically contains many different functions and can be used once for each of its parts.

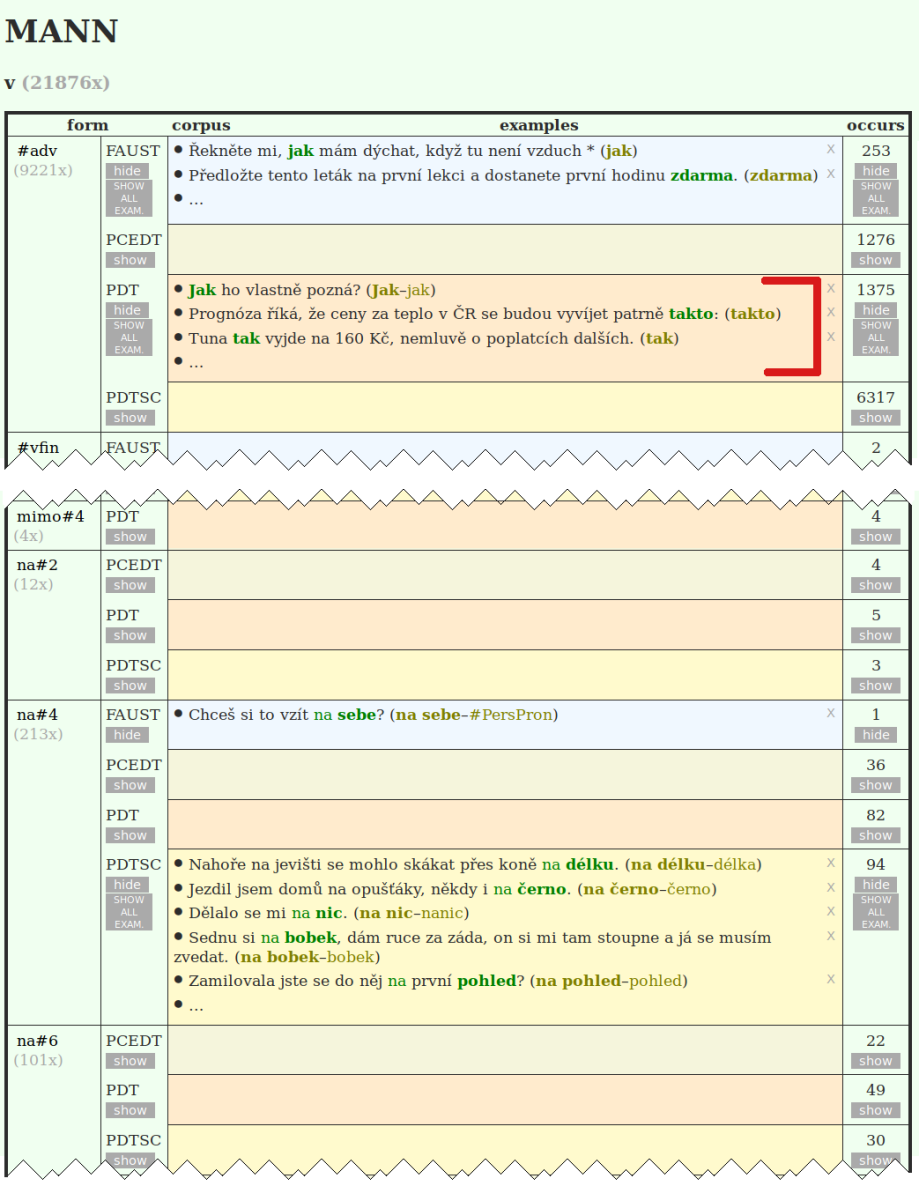


Figure 2: A screenshot of function "manner" in ForFun. The figure presents only a part of the full response obtained from the ForFun database, because the functor MANN is represented by as many as 122 forms (see also Table 4).

nations used very rarely. These occurrences 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 ForFun may contain errors (caused by annotators as well as speakers/writers), especially considering its volume.

7. What is ForFun good for?

Linguistic research is predominantly text-based. Before the corpus era, researchers had to rely on their own excerpts; nowadays, however, a vast amount of supporting material is available in digital form. Such resources are truly valuable only if they are enriched with different layers of linguistic annotation ranging from morphology and syntax to semantics. However, there are many researchers who (want to) use corpora in their everyday work and look for various occurrences of specific words, forms or patterns, syntactic functions, etc. but they are not interested or just do not need to deal with various technical, formal and annotation issues (because they are just researchers in humanities and not so fluent also in technology). Moreover, often if an annotation scenario is based on a sound linguistic theory, it is quite complex and perhaps too complicated for everyday users. Thus, there is a requirement for voluminous and richly linguistically annotated resource which is easy to use. And that is ForFun!

The ForFun database brings the rich and complex annotation in PDTs closer to such everyday, simple use. A rather straightforward use of ForFun is to retrieve which functions can be expressed by the particular form and which forms can express the particular function. To display the richness of the material in the ForFun database we present here two simple examples. Table 2 demonstrates multifunctionality of form (we choose the adverb phrase as an example) and Table 4 demonstrates formal diversity of

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 function	33 355
avg. examples for 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 3: Volume of the ForFun database.

function (we choose "manner" as example, i.e., the functor`MANN`). We can see that the relation between forms and their functions is many-to-many, one form is used for expressing many functions and one function can be expressed using various forms (see also Bejček et al. (2017)).
Besides analysis of the form-function relation, ForFun is user-friendly source of examples for other various explorations in syntax, e.g., valency behavior, coordination/discourse relations, idioms and complex predicates, comparison of written and spoken texts, etc. The first linguistic studies based on the ForFun database analyze subtle meanings of spatial and temporal adverbials (2017a; 2018).

8. Conclusion

We have introduced a unique resource for linguistic studies in syntax: the ForFun 1.0 database. We have demonstrated that ForFun is:

- a simplified interface to PDTs,
- a tool primarily for linguists,
- a database of 180 000 Czech sentences,
- a source of information about syntax,
- a place where 2.2 million examples can be studied,
- a gateway to forms (for a given function),
- a gateway to functions (of a given form).

We believe that nowadays when corpus linguistics differs from traditional linguistics in its insistence on a systematic study of authentic examples of language in use, our database will contribute to a comprehensive syntactic studies.

Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project GA17-12624S. This work has also been supported by the LINDAT/CLARIN project of Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Bibliographical References

Bejček, E., Hajičová, E., Mikulová, M., and Panevová, J. (2017). The relation of form and function in linguistic theory and in a multi-layer treebank. In Jan Hajič, editor, *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 56–63, Praha, Czechia. Univerzita Karlova, Univerzita Karlova.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 969–974. European Language Resources Association.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160. European Language Resources Association, Istanbul, Turkey.

Hajič, J., Hajičová, E., Mikulová, M., and Mírovský, J., (2017). *Prague Dependency Treebank,* Handbook on Linguistic Annotation, pages 555–594. Springer Verlag, Dordrecht, Netherlands.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *Proceedings of the 2th International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. European Language Resources Association.

Marcus, M., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Penn Treebank-3*. Linguistic Data Consortium, LDC99T42, University of Pennsylvania.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). Annotating noun argument structure for nombank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pages 803–806. European Language Resources Association.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., and Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, Prague, Czech Rep.

Mikulová, M., Bejček, E., Kolářová, V., and Panevová, J. (2017a). Subcategorization of adverbial meanings based on corpus data. *Journal of Linguistics / Jazykovedný časopis*, 68(2):268–277.

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

Mikulová, M., Bejček, E., and Panevová, J. (2018). What can we find out about time and space in the forfun database? In *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*

Form	Examples	Raw Frequency
adverb	***Hlasitě*** *se hádali.* 'They was arguing **loudly**.'	9221
v+locative	*Chodil tam* ***ve starých teplákách****.* 'He has been walking there **in old tracksuit**.'	1000
instrumental	*Lupiči mu* ***násilím*** *ukradli hodinky.* 'The robbers has stolen his watches **by force**.'	747
(*jak*)+verb	*Všechno je to tak,* ***jak to má být****.* 'Everything is **as it should be**.'	275
(*tak*) *že*+verb	*V soukromí se tváří,* ***že nevěří ničemu****.* 'In private she pretends **she does not believe anything**'	251
na+accusative	*Zamilovali jste se do sebe* ***na první pohled****?* 'Have you fallen in love to each other **at first sight**?'	213
pod+instrumental	*Hráli jsme* ***pod velkým psychickým tlakem****.* 'We've been playing **under great psychic pressure**.'	150
s+instrumental	*Domácí hráli* ***s nadšením****.* 'Locals have played **with a passion**.'	150
na+locative	***Na jedné noze*** *se tenis hrát nedá.* 'A tennis cannot be played **on one leg**.'	101
po+locative	***Po tmě*** *jsem ráno šel na vlak.* 'I've gone for a train **in the dark**.'	81
	⋮ another 110 forms ⋮	
ve formě+genitive	*publikuje* ***ve formě přehledů*** *a tabulek* 'he publishes **in the form of surveys**'	14
přes+accusative	*obchoduje* ***přes přepážky*** 'she trades **over the counters**'	14

Table 4: Formal realizations of the MANN functor. Shortened list (from more than 100 rows) of the forms that can express manner (functor MANN) gained from ForFun. For each form the number of examples in the database is given.

CRH-2, Gerastree proceedings, pages 133–142. Austrian Academy of Science, Dept. of Geoinformation, Wien, Austria.

Panevová, J. (1974). On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics*, 22(3-40):6–3.

Panevová, J. (1998). Ještě k teorii valence. In *Slovo a slovesnost 59, č.1*.

Panevová, J. (1999). Valence a její univerzální a specifické projevy. In Petr Karlík Zdeňka Hladká, editor, *Čeština - univerzália a specifika. Sborník konference ve Šlapanicích u Brna 17.-18. 11. 1998*. Masarykova univerzita, Brno.

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

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague/Dordrecht.

Language Resource References

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., and Urešová, Z. (2006). *Prague Dependency Treebank 2.0 (LDC2006T01)*. Linguistic Data Consortium.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., and Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). *Prague Czech-English Dependency Treebank 2.0*. Charles

University in Prague, UFAL, LINDAT/CLARIN PID: http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1999). *Penn Treebank*. LDC, ISLRN 141-282-691-413-2.

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Ircing, P., Kolářová, V., Lopatková, M., Mareček, D., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Peterek, N., Romportl, J., Sgall, P., Ševčíková, M., Štěpánek, J., Urešová, Z., and Žabokrtský, Z. (2017). *Prague Dependency Treebank of Spoken Czech 2.0*. Charles University in Prague, UFAL.