# SynTags - Web Interface for Syntactic and Semantic Annotation

**Atanas Atanasov**
Sofia University “St. Kliment Ohridski”
atanasow@gmail.com

## Abstract

This paper presents a web tool for syntactic and semantic annotation and two of its applications. It gives the linguists the possibility to work with corpora and syntactic and semantic frames in XML format without having computer skills. The system is OS and platform independent and could be used both online and offline.

## 1. Introduction

This paper presents an online system for syntactic and semantic annotation. Initially it was developed as a support tool for student theses in syntax and thereafter it was upgraded and used as data processing tool in linguistic research of the prepositional phrases in predicative position in contemporary Bulgarian.

The core of the system is written in XML - it is built on the basis of XForms. In order to be accessible online, it is installed on eXist-db server (http://exist-db.org/), which supports XForms, XQuery etc. It is created using a modified version of AgenceXML’s XSLTforms (http://www.agencexml.com/), which allows browsers to manipulate XForms and has a client-side implementation, preventing server overloading.

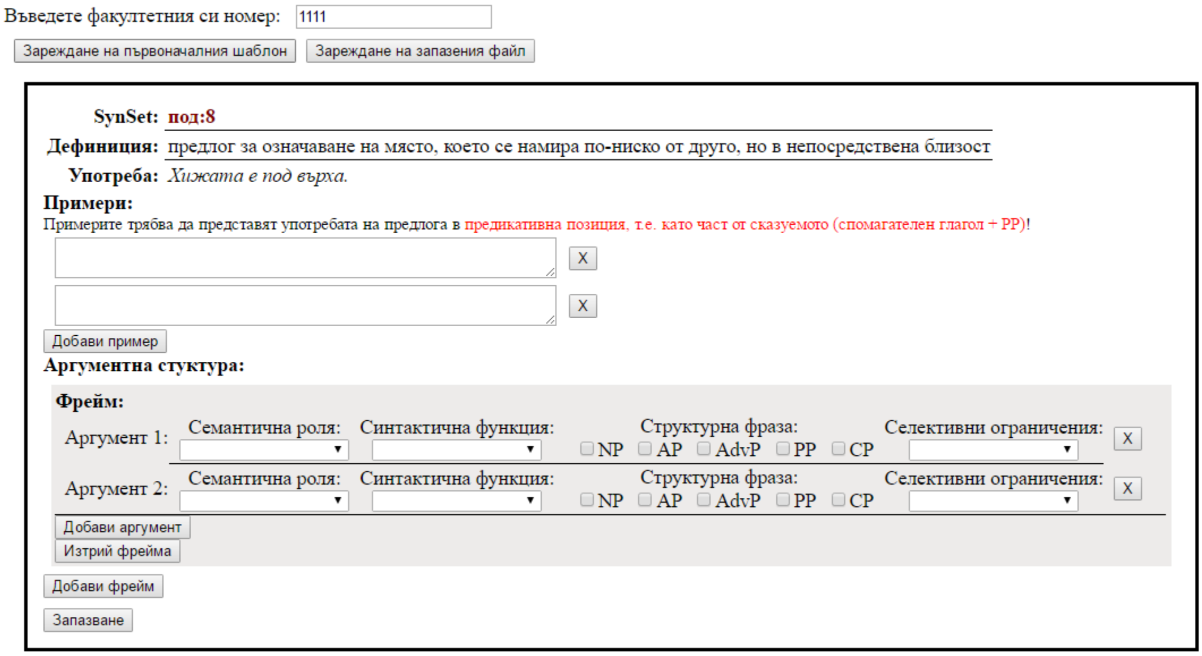
The main advantage of the system is the possibility for the user to fill and save all the data (i.e. to create complicated annotated corpora; to present the argument structure of the predicates and the semantic and subcategorization frame) in xml file without knowing xml or having computer skills at all.

Compared to other existing annotation tools (like Hydra or Chooser for example) SynTags offers a different approach. Unlike Hydra (http://dcl.bas.bg/hydra/), which is a system for browsing and editing wordnet data, SynTags serves a completely different purpose - it uses predefined synsets (that cannot be edited directly from the user interface) and the main goal is to provide an environment for manual presentation of the argument structure of the predicates and the syntactic realization and the semantic properties of these arguments.

It has more in common with Chooser (http://dcl.bas.bg/chooser-2/), but SynTags is not that powerful in semantic mark-up of elements (it is not connected to the whole wordnet database) as the aim is not the creation of semantically annotated corpus, in which all the words are connected to the corresponding synset. The annotation level in the sentences represents the argument positions, so it is more similar to the one used in the Berkeley FrameNet annotation tool (https://framenet.icsi.berkeley.edu/fndrupal/annotation_tool), but SynTags also provides an option to add and edit the framenet data as well as the subcategorization frames (both discussed more detailed in chapters 3.2 and 3.3).

## 2. Application in student theses

The first beta version of the software was tested as a tool for creation of student theses and it was implemented in e-learning system giving the students the possibility to work online on every browser without need to install XML editors or any other apps. The interface of this first working version looks like this:



Въведете факултетния си номер: 1111

Зареждане на първоначалния шаблон | Зареждане на запазения файл

**SynSet:** **под:8**

**Дефиниция:** предлог за означаване на място, което се намира по-ниско от друго, но в непосредствена близост

**Употреба:** *Хижата е под върха.*

**Примери:**

Примерите трябва да представят употребата на предлога в предикативна позиция, т.е. като част от сказуемото (спомагателен глагол + PP)!

**Аргументна стуктура:**

**Фрейм:**

Аргумент 1: Семантична роля: Синтактична функция: Структурна фраза: □ NP □ AP □ AdvP □ PP □ CP Селективни ограничения:

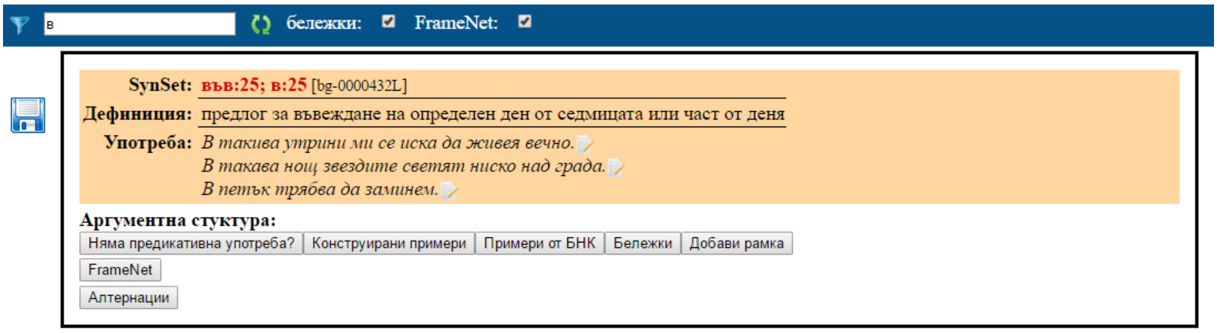
Аргумент 2: Семантична роля: Синтактична функция: Структурна фраза: □ NP □ AP □ AdvP □ PP □ CP Селективни ограничения:

Figure 1

The students have to excerpt the corresponding examples from the Bulgarian National Corpus (BNC) and try to present their argument structure and the semantic relations between the arguments of the predicate. All the data loaded and saved in the browser is actually in XML format, visible for the professors, but not for the students. All the data visible in the web Xform will be discussed in details in the next chapter.

## 3. Application as an annotation tool for PPs in predicative position

After the successful try-out, the system was upgraded with more complex functions, the most important of which is the possibility to annotate the examples and to bind their arguments with the syntactic and semantic frames. Here is a screenshot of the main interface:



в

бележки: ☑ FrameNet: ☑

**SynSet:** **във:25; в:25** [bg-0000432L]

**Дефиниция:** предлог за въвеждане на определен ден от седмицата или част от деня

**Употреба:** *В такива утрини ми се иска да живея вечно.*
*В такава нощ звездите светят ниско над града.*
*В петък трябва да заминем.*

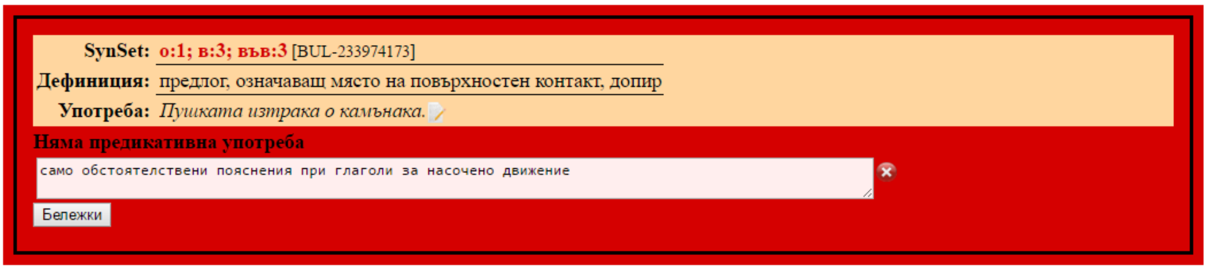
**Аргументна стуктура:**

Figure 2

The header of each Synset contains the main information from the Bulgarian Wordnet - the literals (with the corresponding sense number), the ID, the definition and the usage given in BulNet (where it’s applicable). This information is manually copied from BulNet 3.0 (http://dcl.bas.bg/bulnet/) in a pre-process XML file. The user has the possibility to make some personal notes for every one of the usage examples.

Below the Wordnet block there is an “Argument structure” section containing several other options: “No predicative usage”, “Constructed examples”, “Examples from Bulgarian National Corpus”, “Notes”, “Add frame”, “FrameNet” and “Alternations”.

The first one is used for those prepositions that could not be used as predicatives. When pressed it deletes all the information already entered (if there's any) and eliminates all the other options in the Synset. The Synset window is colored red and only one textbox that remains in it is about free text description for the reason why the preposition cannot be a part of a predicate (for example 'only attributive usage'). Also there's an option to add some additional notes. The delete button next to the textbox reverts the Synset interface to the initial state - the user can again add and edit examples, frames etc.



SynSet: о:1; в:3; във:3 [BUL-233974173]

Дефиниция: предлог, означаващ място на повърхностен контакт, допир

Употреба: *Пушката изтрака о камънака.*

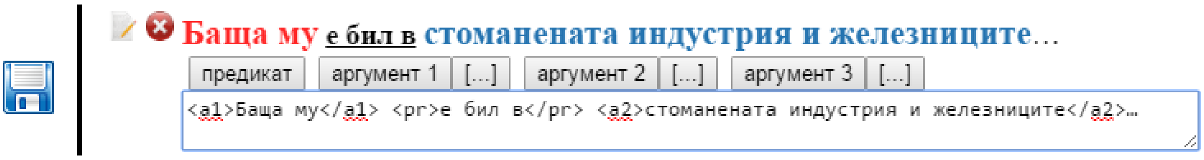
Няма предикативна употреба

само обстоятелствени пояснения при глаголи за насочено движение

Figure 3

### 3.1. Corpora annotation

In order to provide evidence that the analysis is correct, every particular sense should be illustrated by as many examples as possible. In this case it is advised (following the principles stated in Koeva et al. 2008) that at least five examples should be given for every Synset. Pressing one of the next two buttons ('constructed examples' and 'examples from BNC') triggers an interactive text area, where after the example is entered, it could be annotated with the help of the buttons above the box.



Баща му е бил в стоманената индустрия и железниците…

<a1>Баща му</a1> <pr>е бил в</pr> <a2>стоманената индустрия и железниците</a2>…

Figure 4

When a part of the text is selected, pressing a button wraps the selection in XML tag. The first one puts <pr>...</pr>, which marks the predicate (in this specific usage it actually marks a part of the predicate - the auxiliary verb and the preposition, interpreted here as the core of the predicate). The next buttons mark the arguments if they are explicit (e.g. <a1>Той</a1>) or their position if they are implicit (e.g. <a1>[...]</a1>). Any changes in the textbox appear above in real time presenting the data formatted in different style depending on the corresponding XML annotation. The styled text is interactive - clicking on it shows or hides the edit window for the example. Also saving the document makes all the edit text boxes disappear.

For each example there's also an option to add or delete a note or the whole element.

The actual data is saved in the xml file in an <example>...</example> element, so the previous example is coded in the following format:

<example>&lt;a1&gt;Баща му&lt;/a1&gt; &lt;pr&gt;е бил в&lt;/pr&gt; &lt;a2&gt;стоманената индустрия и железниците&lt;/a2&gt;…</example>,

creating this way a syntactically and semantically annotated corpus.

### 3.2. Argument structure

When the examples are ready the next button adds the subcategorization frames. Here the linguist has the possibility to add or remove frames and to add or delete arguments in the frames. The number of the arguments depends on the semantic properties of the predicate - they should vary from zero to three.

In the system there are two semantic levels of presentation. The first one is more generalized and it follows the well-known semantic roles in Role and Reference Grammar (Van Valin et al. 1997), where the relations between the predicates and their arguments are presented with the following scheme:

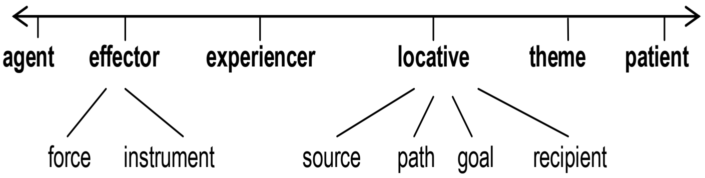
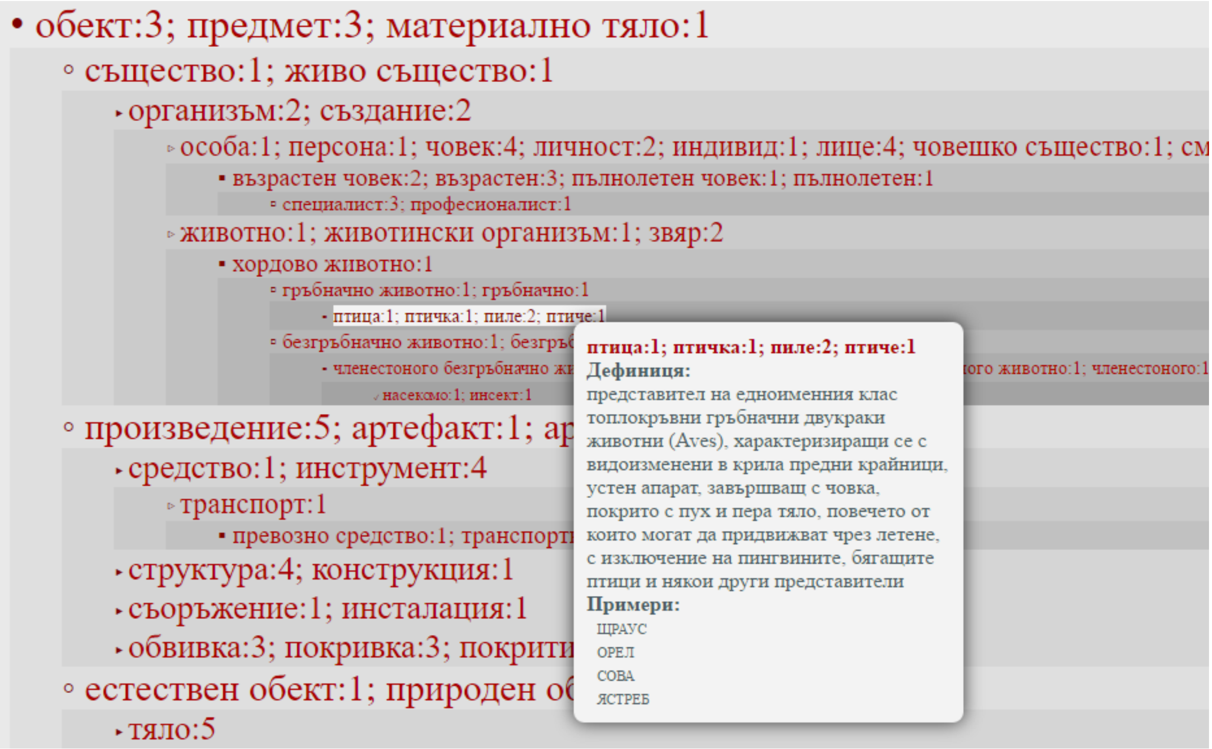


Figure 5

As all the frames in a Synset refer to the same definition they need to have the same number of arguments and in the most cases their arguments should have the same semantic roles. If the semantic roles are different, it means the definition should be divided into parts presenting more accurately the semantics of the predicate.

The other semantic level is directly connected to the Princeton Wordnet synsets and their Bulgarian correlates. The main goal is to present the exact selective restrictions of the core elements. In other words, this is an attempt for more precise description of the semantic properties of the arguments. In a separate XML file are extracted the main concepts from the Wordnet hierarchy - about 65 synsets considered as a “skeleton” and they are dynamic – every time when an argument requiring a synset not included in the file is found, it has to be added. This file is published and accessible online as a HTML page and the user could go to this interactive web page (fig. 6) for a quick reference of the hierarchical relations, definitions, examples and so on. Of course, if more detailed information is needed, the linguist should check the official BulNet/WordNet website.



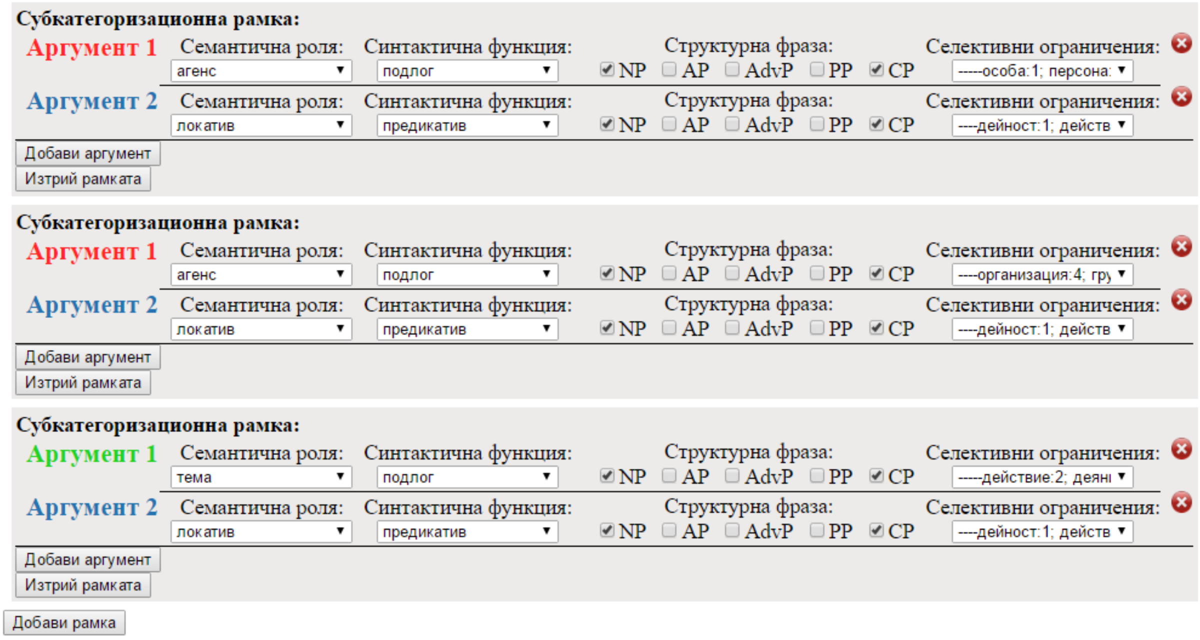
- обект:3; предмет:3; материално тяло:1
    - съществo:1; живо същество:1
        - организъм:2; създание:2
            - особа:1; персона:1; човек:4; личност:2; индивид:1; лице:4; човешко същество:1; см
                - възрастен човек:2; възрастен:3; пълнолетен човек:1; пълнолетен:1
                    - специалист:3; професионалист:1
            - животно:1; животински организъм:1; звяр:2
                - хордово животно:1
                    - гръбначно животно:1; гръбначно:1
                        - птица:1; птичка:1; пиле:2; птиче:1
                    - безгръбначно животно:1; безгръб
                        - членестоного безгръбначно жи ого животно:1; членестоного:1
                            - насекомо:1; инсект:1
    - произведение:5; артефакт:1; ар
        - средство:1; инструмент:4
            - транспорт:1
                - превозно средство:1; транспорт
        - структура:4; конструкция:1
        - съоръжение:1; инсталация:1
        - обвивка:3; покривка:3; покрити
    - естествен обект:1; природен об
        - тяло:5

**птица:1; птичка:1; пиле:2; птиче:1**
**Дефиниця:**
представител на едноименния клас топлокръвни гръбначни двукраки животни (Aves), характеризиращи се с видоизменени в крила предни крайници, устен апарат, завършващ с човка, покрито с пух и пера тяло, повечето от които могат да придвижват чрез летене, с изключение на пингвините, бягащите птици и някои други представители
**Примери:**
щраус
орел
сова
ястреб

Figure 6

The syntactic function of the arguments also should be presented, following the traditional classification: subject, predicative (not an argument of the predicate, as it is a part of it together with the preposition and the copula - it is considered to be an argument of the preposition itself), direct and indirect object, adjunct and small clause.

The following figure illustrates a sysnset’s argument structure presentation:



**Субкатегоризационна рамка:**

Аргумент 1 Семантична роля: агенс ▼ Синтактична функция: подлог ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: -----особа:1; персона: ▼ ☒

Аргумент 2 Семантична роля: локатив ▼ Синтактична функция: предикатив ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: ----дейност:1; действ ▼ ☒

Добави аргумент

Изтрий рамката

**Субкатегоризационна рамка:**

Аргумент 1 Семантична роля: агенс ▼ Синтактична функция: подлог ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: ----организация:4; гру ▼ ☒

Аргумент 2 Семантична роля: локатив ▼ Синтактична функция: предикатив ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: ----дейност:1; действ ▼ ☒

Добави аргумент

Изтрий рамката

**Субкатегоризационна рамка:**

Аргумент 1 Семантична роля: тема ▼ Синтактична функция: подлог ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: -----действие:2; деяни ▼ ☒

Аргумент 2 Семантична роля: локатив ▼ Синтактична функция: предикатив ▼ Структурна фраза: ☑ NP ☐ AP ☐ AdvP ☐ PP ☑ CP Селективни ограничения: ----дейност:1; действ ▼ ☒

Добави аргумент

Изтрий рамката

Добави рамка

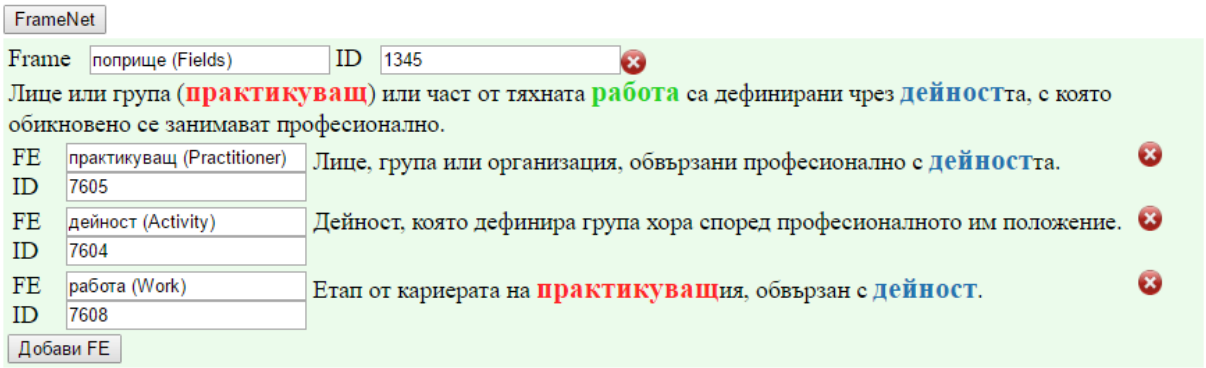
Figure 7

The number of the frames in a Synset depends mainly on the selective restrictions of the arguments. The predicate - representing a real situation - should have a fixed number of core elements, but they could have different realization - syntactic or semantic. The main phrase type (NP, AP, AdvP, PP or CP) have to be chosen for each element. If an argument with the same meaning could be realized as more than one type of structure phrase, there is an option all of them to be presented in the same frame. For example the subject in Bulgarian sentence always could be expressed with NP or with CP and the locative adjuncts can be expressed with AdvP or PP. Since this alternations are consistent there is no need adding a second frame - it is enough to check both in the same frame.

There should be more than one frame when the selective restrictions belong to different categories, e.g. the predicates that require a person (physical entity) or an organization (abstract entity) in the same argument position.

## 3.3. FrameNet

In the system there is also an option to connect the predicate meaning (the synset definition) with the corresponding frame from Berkeley FrameNet Project. As the Bulgarian FrameNet is still in working stage and is not accessible yet, this binding for now is only manual and presents the only the core frame elements. The frame and the frame core elements names and definitions and translated in Bulgarian and aligned with the original data (FameNet 1.6).



FrameNet

Frame поприще (Fields) ID 1345 ☒

Лице или група (**практикуващ**) или част от тяхната **работа** са дефинирани чрез **дейност**та, с която обикновено се занимават професионално.

FE практикуващ (Practitioner) Лице, група или организация, обвързани професионално с **дейност**та. ☒
ID 7605

FE дейност (Activity) Дейност, която дефинира група хора според професионалното им положение. ☒
ID 7604

FE работа (Work) Етап от кариерата на **практикуващ**ия, обвързан с **дейност**. ☒
ID 7608

Добави FE

Figure 8

All the frame elements, the arguments in the subcategorization frames and their realization in the examples are bound to each other and styled the same way (cf. fig. 4, 7 & 8).

In this particular application (for description of predicative PPs) another experimental function is available – presenting the possible substitutions of the auxiliary verb with a lexical verb or the PP with AdvP.

### 3.4. Filtering and search

At the top of the web page there are several filter options. It is possible to search for a literal and display only the synsets containing it, to show or hide the user notes and also to activate or stop the FrameNet functionality.

## 4. Advantages

This are the main pluses of the SynTags system:

- *Universal tool for corpora and syntax frame annotation*. The system can be easily modified (for now only by changing a few lines in the source code) in order to satisfy the needs of any particular linguistic task related to corpus annotation or semantic and syntactic presentation.
- *Easy collaboration*. The tool can be used by many developers working on the same xml database.
- *Easy access*. It is platform and operating system independent - the only requirement is a current web browser.
- *Comfortable user interface*. Not special programming knowledge is required, so everybody could use the tool without having advanced computer skills.
- *Online and offline usage*. The tool is accessible online, but it also could be easily installed locally on a free open source eXist-db server.

## 5. What's next?

- *Optimization for large data processing*. The current version has some issues concerning the processing of very big files, so in the future the efforts will be concentrated mainly on improving the stability and the speed of the system.
- *Adding a more complex search and filter functionality*. Now the system can search only by xPath expressions - it is planned to improve this functionality by adding a full xQuery support.
- *Adding options for advanced user settings*. SynTags currently works with predefined XML and DTD files – the next step will be to give the uses the opportunity to modify them partially from the user interface.
- *Implementation of full FrameNet support*. It was mentioned that the FrameNet data could be entered manually. The future plans include full implementation of FrameNet 1.6 in the system database.

## References

Baker, C. F., Fillmore, C. J. and Lowe, J. B.  (2006). The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics, Volume 1*. Association for Computational Linguistics, 1998, pp. 86-90.

Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fillmore, C. J. and Baker, C. F. (2009). A Frames Approach to Semantic Description. In B. Heine and H. Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.

Koeva, S. (2010). *Balgarskiyat FrameNet 2010*. Sofia, 2010.

Koeva, S., Vlahova, R., Dekova, R., Nestorova, P. and Atanasov, A. (2008). *Balgarskiyat FrameNet. Semantiko-sintaktichen rechnik na balgarskiya ezik*. sastavitel Svetla Koeva, Sofia, 2008.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson C.R. and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Van Valin, R. D., Jr. and LaPolla, R. J. (2007). *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

## Resources

Bulgarian National Corpus: http://search.dcl.bas.bg/

BulNet: http://dcl.bas.bg/bulnet/

Chooser: http://dcl.bas.bg/chooser-2/

FrameNet: https://framenet.icsi.berkeley.edu/fndrupal/

FrameNet Annotation Tool: https://framenet.icsi.berkeley.edu/fndrupal/annotation_tool

Hydra: http://dcl.bas.bg/hydra/

WordNet: http://wordnet.princeton.edu/