Lexicon and Grammar in Bulgarian FrameNet

Svetla Koeva
Department of Computational Linguistics, Institute for Bulgarian Language
52 Shipchenski prohod, Sofia 1113, Bulgaria
E-mail: svetla@dcl.bas.bg

Abstract

In this paper, we report on our attempt at assigning semantic information from the English FrameNet to lexical units in the Bulgarian valency lexicon. The paper briefly presents the model underlying the Bulgarian FrameNet (BulFrameNet): each lexical entry consists of a lexical unit; a semantic frame from the English FrameNet, expressing abstract semantic structure; a grammatical class, defining the inflexional paradigm; a valency frame describing (some of) the syntactic and lexical-semantic combinatory properties (an optional component); and (semantically and syntactically) annotated examples. The target is a corpus-based lexicon giving an exhaustive account of the semantic and syntactic combinatory properties of an extensive number of Bulgarian lexical units. The Bulgarian FrameNet database so far contains unique descriptions of over 3 000 Bulgarian lexical units, approx. one tenth of them aligned with appropriate semantic frames, supports XML import and export and will be accessible, i.e., displayed and queried via the web.

1. Introduction

The Bulgarian FrameNet (http://dcl.bas.bg/BulFrameNet.html) represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units (the pairing of a word (either a single word or a multi-word expression) and word sense). The main objectives of the Bulgarian FrameNet project at this stage are as follows:

- To build up an extensive database comprising particular values of language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units, i.e., to enlarge the Bulgarian valency lexicon SynText (Koeva, 2004);
- To assign the appropriate abstract semantic descriptions from FrameNet semantic frames (Fillmore, 1982; 1985; Fillmore et al., 2003) to the language-specific representations in the Bulgarian valency lexicon;
- To develop a large annotated corpus illustrating the combinatory properties – fully annotated with respect to the FrameNet frame elements and language-specific syntactic realisations.

The Berkeley FrameNet project (Baker et al., 2003; Johnson et al., 2002; Ruppenhofer et al., 2006) is one of the most significant linguistic approaches based on frame semantics and supported by corpus evidence. Frame semantics describes word meaning in terms of underlying conceptual structures. These are encoded in the form of frames, i.e., schematic representations of stereotyped situations capturing a certain amount of background (real-world) knowledge. Recently, creation of FrameNets for several languages other than English has been started, based in general on the major assumptions and generalisations of the Berkeley FrameNet, reporting as well some independent results (Burchardt et al., 2006); Ohara et al., 2004; Subirats & Petruck, 2003; among others). To the best of our knowledge, Bulgarian is one of the few Slavic languages in the multilingual FrameNet family, although some research on Czech and Slovenian has been reported (Benešová et al., 2008; Lönneker-Rodman et al., 2008). The development of FrameNets different from English, the so called 'multi-lingual' FrameNet, is expected to reveal cross-language dependancies, similarities and differences and make prominent language-specific and language-independent typological idiosyncrasies.

The Bulgarian FrameNet (BulFrameNet) database (Koeva, 2008) so far contains unique descriptions of over 3,000 Bulgarian lexical units, approx. one tenth of them aligned with appropriate semantic frames, supports XML import and export and will be accessible, i.e., displayed and queried via the web.

2. Bulgarian FrameNet structure in brief

A lexical entry in BulFrameNet consists of a lexical unit, a semantic frame from the English FrameNet expressing abstract semantic structure, a grammatical class, defining the inflexional paradigm, a valency frame describing (some of) the syntactic and lexical-semantic combinatory restrictions (an optional component), and (semantically and syntactically) annotated examples (Koeva, 2008). The meaning of a lexical unit is expressed by the triplet of lemma, unique word sense (if the sense is already defined in the Bulgarian WordNet[1], the respective WordNet ID is linked to) and annotated examples.

A semantic frame is a conceptual structure that describes a particular type of situation, object, or event, along with its participants and props, which are referred to as frame elements (Ruppenhofer et al., 2006). The frames are provided with a name, a definition, and a semantic type, and contain frame elements with a name, a definition, a semantic type, with a specification of the coreness status, and frame-internal relations between frame elements.

The grammatical class is constituted by the categories of paradigmaticity, inherent transitivity (for verbs only), aspect (for verbs only), and inflectional type.

The valency frame (applicable to some word classes only) is a combination of syntactic structures (zero, one, or more) that are associated with the lexical unit. The

[1] http://dcl.bas.bg/BulNet/general_en.html

syntactic structures list all possible combinations of syntactic categories, grammatical functions and other attributes that specify lexical-semantic and syntactic realisation of arguments.

The annotation component provides for a given lexical unit (syntactically and semantically) annotated examples taken from the (extended) Bulgarian National Corpus.

The progress in the development of the BulFrameNet is given in Table 1.

Number of lexical units (LU)	3032
LUs with aligned semantic frames	271
LUs with syntactic annotation	2853
LUs with semantic annotation	258
LUs with valence frames	2853

Table 1: The BulFrameNet statistical data.

3. Development methodology

Some efforts were invested in developing the Bulgarian valency dictionary before its redesigning to follow (to a large extent) the English FrameNet model. Applying the so-called expand model (Vossen, 1999) for developing multi-language lexical resources would probably result in faster achievement of the targeted goal – a Bulgarian FrameNet. Nevertheless, we decided to keep and enlarge the extensive valence information we had encoded so far and to follow a kind of merge model.

The four main areas of activity involve: lexical units identification and grouping, corpus annotation, valence frames development, and alignment with English semantic frames. Of course all components of the development of BulFrameNet can be viewed as mutually related and not strictly ordered. For example, the identification of the semantic frame which a particular lexical unit evokes might impose validation and further specification of the word sense definition given so far. In many cases the syntactic and semantic annotations might provoke the redefinition of a target word sense, reformulation of a particular valency frame or reassignment to a particular semantic frame.

The entire process of developing the Bulgarian FrameNet is manual, although supported by some applications: for automatic annotation, frequency calculation, Bulgarian WordNet processing, consistency validation, etc. The basic activities in developing BulFrameNet are briefly described in this section.

3.1. Lexical units identification and grouping

The prospective words for inclusion are selected from a frequency list of noun, verb and adjective lemmas extracted from a fully morpho-syntactically annotated text archive with close to 500 million words[2]. Relative weights are assigned: 0.3 to verbs, 0.2 to nouns, and 0.1 to adjectives, in order to achieve a better balance. Further, additional weight is given to the lemmas found in the Bulgarian WordNet (BulNet) according to the number of corresponding word senses encoded so far. The initial frequency list is then reordered taking into account the assigned weights. The resulting topmost lemmas in the frequency list reflecting the general vocabulary of Bulgarian have priority rendering them appropriate for inclusion in the lexicon.

For each lemma selected, the entire set of corresponding sense definitions are extracted from the Bulgarian WordNet. The following cases are possible: an appropriate sense definition is available; an appropriate sense definition is not available, and in that case the new sense will also be included in the BulNet structure; senses have to be merged or split, which indicates that the existing BulNet sense definitions are either too fine-grained or two vague and have to be revised. The WordNet synonymy sets allow to work simultaneously with all lexical units related with a synonymy relation of equivalence.

The frequency of a given sense, based on the Bulgarian semantically annotated corpus (approx. 150,000 words manually annotated with BulNet senses) is also assigned to the lexical unit, for example the lexical unit ходя ('to walk') with a BulNet based definition 'придвижвам се пеш, като повдигам и премествам краката си един след друг' (in English WordNet 'use one's feet to advance; advance by steps'):

```
<word lemma="ходя">
<def source="WN_based" definition="придвижвам се пеш">
<id frequency="0,4‰">"="ENG20-01849285-v</id></def>
```

A large number of examples illustrating modern Bulgarian are explored to specify the correct word sense. The identified lexical units are grouped into sets according to their semantic descriptions. Grouping is facilitated by the automatic extraction of data accumulated in the Bulgarian valency lexicon so far: i.e. lexical units belonging to one and the same grammatical type, having one and the same syntactic structure, etc.

3.2. Corpus annotation

The BulFrameNet project is based on the evidence offered by the 320 million words of the Bulgarian National Corpus[3] plus 33 million words' worth of separately acquired texts that were included to illustrate some rare or more specific senses. The whole corpus is automatically annotated for sentence and word boundaries, parts of speech and detailed grammatical information. Because of the size of the Bulgarian National Corpus, only parts of it are manually annotated: 300,000-plus words for parts of

[2] The text archive is compiled by the Department of Computational Linguistics team (Institute for Bulgarian Language at Bulgarian Academy of Sciences) along different tasks carried out at the Department involving the processing of large-volume language data.

[3] http://ibl.bas.bg/en/BGNC_en.htm

speech and 150,000-plus words for word senses, using the respective tagsets and annotation schemata (Koeva et al., 2006).

A number of the examples used for the identification of lexical units are selected for further annotation. At the stage of valency dictionary development at least five of the selected examples are annotated for the instances of the target lexical unit, arguments phrase types and grammatical functions. Although the goal is to show how valency patterns are instantiated in arbitrary texts, preferences are given to the examples where the lexical unit's arguments receive explicit realisation. The logical position of the implicit arguments is marked accordingly, for example the dropped subject in a sentence with the lexical unit ходя ('to walk'):

<example source="http://search.dcl.bas.bg/ - Български национален корпус"> Той забърза повече, искаше му се не <NPext type="NPimpl" /> да <P>ходи</P>, а да тича. </example>

When a given lexical unit is aligned with a semantic frame, a manual semantic annotation of frame elements for targets identified as evoking a particular semantic frame is also performed – by means of attaching frame elements labels to the appropriate sentence constituents, for example the frame element label **Self_mover** to the external argument of the lexical unit ходя ('to walk'):

<example source="http://search.dcl.bas.bg/ - Български национален корпус"> Той забърза повече, искаше му се не <NPext FE="Self_mover" type="NPimpl" /> да <P>ходи</P>, а да тича.</example>

The annotation schema at the frame semantic level follows those used in the English FrameNet project. Dubious cases are solved by calling in a second opinion from at least one more annotator.

To summarise, the annotation can be defined as a Framenet-style annotation providing for a given lexical unit (at least five) annotated examples taken from the (extended) Bulgarian National Corpus. Thus, the number of annotated examples associated with a given semantic frame depends on the number of identified lexical units that evoke it. The main role of the annotated corpus is to represent the semantic and syntactic structure of modern Bulgarian – language-specific lexicalizations and syntactic realisation, as well as to illustrate abstract semantic frames.

3.3. Bulgarian valency lexicon

The identified lexical units (verbs) are supplied with language-specific descriptions of their combinatory properties from the Bulgarian valency lexicon.

The basic modules of the Bulgarian valency lexicon specify the grammatical class and syntactic (valency) frame of respective lexical unit. Syntactically annotated examples are provided as well. The appropriate value selection of language specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units results in creation of / integration to the respective grammatical class and valency frame.

3.3.1. Grammatical class

The grammatical class is described by the set of morpho-syntactic properties of the lexical unit. For the specification of grammatical class attributes and their values (different for particular word classes) are defined. For example each verb is specified as personal, impersonal or third personal according to its person paradigm; as transitive or intransitive (transitive and intransitive verbs may be further distinguished according to their lexical properties to build a compound word with a "reflexive" particle or with an "accusative" or "dative" pronominal clitic); as imperfective, perfective, secondary imperfective, bi-aspectual, imperfectiva tantum or perfectiva tantum according to its aspect. Each lemma is assigned with an unambiguous formal description of its inflectional paradigm. For example the grammatical class of the lexical unit ходя ('to walk') is described as personal, imperfective, intransitive and belonging to a particular inflectional type:

<morph aspect="несвършен" transitivity="непреходен" person="личен" inflection _type="V+P+IM+IN,16 />

The chosen value has an impact on the possible sets of other values – i.e., personal verbs can be either transitive or intransitive, but impersonal and third personal are only intransitive with the exception of the accusativa tantum class; as well as on the possible sets of syntactic structures – i.e., personal verbs can have NP subject, third personal verbs – NP or S subject, impersonal verbs – no subject.

The grammatical class specifications ensure exact correspondence between lexical unit and its word form instances and provide part of the conditions necessary for the identification of possible alternations.

3.3.2. Valency frame

The valency frame consists of one or more syntactic structures that differ in the syntactic realisation of the arguments, but are identical with respect to their number and basic semantic features encoded. The syntactic structure itself uniquely defines the number of arguments with values for the following attributes: syntactic category (including allowed prepositions, and complementizers, if any), lexical explicitness of the phrase, grammatical function and semantic restrictions (selectional and lexical-semantic).

Thus the valency frame encodes information for the language-specific restrictions in lexicalization and syntactic realisation, i.e., different syntactic categories (Shte blagodarya [PP **na Maria**] – 'I will thank [NP **Maria**]'); different rules for implicit usage (**[dropped subject]** Iskam da blagodarya na Maria – '**I** would like to thank Maria'); different grammatical functions ([Subject **Maria**] [Object **mi** ('I, accusative clitic')] lipsva – '[Subject **I**] miss [Object **Maria**]') and so on.

Detailed information about the structure, tagset and annotation schema adopted in the Bulgarian valency lexicon is given in Koeva (2008). A few examples will be mentioned here to demonstrate some of the representation principles followed.

It has been taken into account that the noun and prepositional phrases in Bulgarian can be obligatorily explicit (Shte izpiya **vodata** – 'I will drink up **the water**'), implicit, or their explicitness might depend on other sentence constituents: for example S argument depending on the explicitness of the NPcl argument (Marzi [NPcl **me**] [S **da cheta**] – 'I do not feel like reading'), PP argument depending on the explicitness of other PPcl argument (Domachnya [PPcl **mi]** [PP **za tyah**] – 'I miss them').

Prepositions are grouped in synonymous sets if they are interchangeable in a context, for instance the prepositions върху, над, по ('on, upon') introducing 'a subject of an activity'. Observations show that different prepositional synonymous sets might satisfy different selectional restrictions. Since prepositions belong to a closed word class, the appropriate prepositional synonymous sets are chosen from a predefined list where multiple selections are allowed. A similar approach is used to specify words introducing clauses.

Semantic restrictions are viewed as properties that specify how word classes are able to associate with a given argument in a given position. Semantic restrictions are divided in two groups – selectional and lexical-semantic restrictions. Selectional restrictions are considered as a classification based on the general type of the denoted concept. Lexical-semantic restrictions, on the other hand, show which lexical-semantic classes of words are allowed to co-occur with a target lexical unit. The accepted values of the selectional restrictions are *concrete*, *abstract*, *animate*, *inanimate*, *human*, *non-human, agentive, non-agentive* and are organised in a directed graph. Some selected synonymous sets from the Bulgarian WordNet (their conjunctive or disjunctive combinations) are used to specify the particular lexical-semantic restrictions.

For example the valency frame of the lexical unit ходя ('to walk') consists of one syntactic structure with NP subject that can remain implicit; the selectional restrictions are satisfied by NP specified as *sentient*, while lexical-semantic restrictions are either *human* or *animate able to walk*:

```
<Vframe>
   <SynStr>
      <NP expl="не" function="подлог">
      <SemR sel="одушевено" ls="човек | животно"/></NP>
   </SynStr>
</Vframe>
```

The semantic restrictions should be valid cross-linguistically – only their lexicalization, grammatical and syntactic realisation are language-specific.

In the English FrameNet a semantic type is assigned to a frame element – it constrains the semantic type of the syntactic phrase that corresponds to the frame element (Ruppenhofer et al., 2006). There is a rather limited set of semantic types for frame elements. On the one hand, not all frame elements are assigned with a semantic type. Until now no contradiction has been observed between FrameNet semantic types and semantic restrictions in the Bulgarian valency lexicon.

3.4. Aligning with English semantic frames

A frame semantic description of a lexical unit identifies the frame which is evoked by a given word sense and specifies the ways in which frame elements are realised in language structures (Johnson et al., 2002). A given frame is associated with a set of words (verbs, nouns, or adjectives) or expressions that evoke it and a set of semantic roles (frame elements) corresponding to the participants and props in the designated prototypical situation.

For example, the set of semantically related Bulgarian verbs сека, цепя ('to cut'), отсичам ('to hew'), отсека ('to cut down'), насека ('to chop up'), насичам ('to chop'), отсека, посека ('to cut down'), съсичам, посичам ('to slay'), разсека ('to cut down'), разсичам ('to slay'), режа, нарязвам ('to slice'), нарежа ('to cut up'), прережа ('to cut off'), прерязвам ('to cut'), отрежа ('to cut off'), отрязвам ('to cut'), разцепвам ('to split'), разцепя ('to split up') evoke the frame **Cutting** with the definition:

An Agent cuts an Item into Pieces using an Instrument

The same frame, **Cutting**, is evoked by the English verbs *carve*, *chop*, *cube*, *cut*, *dice*, *fillet*, *mince*, *pare*, *slice*, etc. and their translation equivalents in other languages in which the concept is lexicalized.

It is accepted that the frame part of the FrameNet database has the highest degree of inherent equivalencies across languages, i.e., to a great extent frames are presumably language-independent (Ruppenhofer et al., 2006). We also assume that the semantic frames are in general cross-linguistically valid, because they describe conceptual structures. The differences between languages appear in different lexicalization patterns, understood broadly as different lexical and syntactic structures. On the basis of the experience in developing FrameNets for languages other than English (Ellsworth et al., 2004; Lönneker-Rodman, 2007; Padó, 2007) one of the following options is possible: some English frames can be used without any changes; some English frames have to be slightly modified to cover Bulgarian language-specific data (it is stated that the degree of interpretability is inversely related to the typological distance from English); or a new frame has to be introduced (either because it has not been defined yet or because there are language-specific concepts that represent different cultural features).

There are some known approaches which attempt to translate manually the annotated sentences from the English FrameNet framework, and there are some other approaches that involve creating new proto-frames which might cover either not defined or language-specific senses

(Burchardt et al., 2006). Most of the non-English FrameNets follow the expand approach, i.e., by importing frame names and definitions from the original Berkeley FrameNet and manual annotation of existing corpora for each language (Ohara et al., 2004; Subirats & Petruck, 2003).

Conditions for cross-lingual equivalence between frames have not been formally defined yet. It is stated that cross-lingual relations between frames should be regarded as strict equivalence only in such cases where all attributes and frame elements, as well as frame element relations, are the same in both languages (Lönneker-Rodman, 2007).

Our experience in the developing of Bulgarian WordNet shows that it is very useful to assign monolingual resources to a kind of interlingua. If we assume the frame part of FrameNet as "interlingua" (Boas, 2005), the conceptually difficult and technically important questions are how the interlingua structure will be modified on the basis of language-specific data. An indirect mapping (following the Global WordNet practice (Vossen, 2004) may use an intermediate index, represented separately from any of the individual languages, and the mapping may contain information about language-specific language features.

Basically, Bulgarian lexical units are aligned with the appropriate semantic frames from the English FrameNet – namely, we decide for a lexical unit whether it evokes one of the already constructed semantic frames. Semantic frames, vice versa, are linked to appropriate sets of Bulgarian lexical units. The cross-lingual equivalence to the following components of the frame structure is verified: the frame provided with its name, the definition, and the frame semantic type; the frame elements provided with their names, the frame elements definitions, the frame elements types, the frame elements coreness status; and the internal relations between frame elements. Outside the scope of the correspondences established for the time being remain frame-to-frame relations, translation equivalence with the English lexical units and English annotated sentences.

Some of the linguistic problems encountered when other FrameNets are developed have been discussed (Boas, 2009): degrees of overlapping cross-lingual polysemy, differences in the lexicalization patterns, measurement of paraphrase relations (words that evoke a given meaning may differ across sentences) and translation equivalence. At the moment, problems caused from linking Bulgarian lexical units to English semantic frames are put on record, but no further steps towards redefinition, refinement or inclusion of proto-frames are attempted.

To summarise, in this stage of the development of the BulFrameNet we adopted the merge model approach – i.e., some components are built independently (lexical units definitions, valency frames, syntactic annotation). The advantages of the merge approach are that it takes into account the semantic distinctions unique to the target language. The shortcomings of the merge model are that some lexical units are left unassigned with a particular semantic frame because the appropriate one has not yet been defined in the English FrameNet.

This is an example of valency for the Bulgarian lexical unit режа ('to cut') mapped to the **Cutting** frame (Table 2) – only core frame elements plus the element Instrument are shown.

Frame element	Syntactic realisations
Agent	NP, subject, implicit; PPот indirect object, implicit; PPcl indirect object, explicit
Lexical unit	rezha, V ('to cut')
Item	NP, direct object, implicit; PPот indirect object, explicit usage dependent from the NP object; NPext, subject, implicit
Pieces	PPна, indirect object, explicit usage dependent from the NP object; NP, direct object, explicit
Instrument	PPc, indirect object, implicit; NPext, subject, explicit

Table 2: Realisation table of the verb режа ('to cut').

The same frame elements occur in the following valency patterns (Table 3):

Agent	Item	Pieces	Instrument
(NPext)	(NP)	(PPна)	(PPc)
(NPext)	(PPот)	NP	(PPc)
(PPот)	(NPext)	(PPна)	(PPc)
PPcl	(NPext)	(PPна)	(PPc)
	NPext	(PPна)	(PPc)
	(NP)	(PPна)	NPext

Table 3: Valency patterns of the verb режа ('to cut').

4. Alternations and (Bulgarian) FrameNet

The representation of alternations is an important question when dealing with verbal lexical units. An alternation, roughly, describes a change in the realisation of verbal arguments with respect to a postulated initial form. The alternations are regular language structures – if and only if the neutral structure satisfies particular conditions, described in the appropriate valency frame, the set of the possible diatheses cam be predicted. Verb alternations, as described by B. Levin, do not constitute a homogeneous class. They can be split in several groups:

1) Semantic alternations, called here **diatheses** (altering verb senses by reducing subject semantic relations), for example the neutral alternation Ива реже хляба на филии с нож ('Iva cuts the bread (in slices) with a knife') alters with the Middle alternation Хлябът се реже лесно на филии с нож (The bread cuts easily (in slices) with a knif.') as well as with the Instrumental subject alternation Ножът реже хляба на филии. ('The knife cuts the bread (in slices)').

2) **Semantic alternations** (transforming non-subject semantic relations), for example the Manner Source alternation Ива реже филии от хляба с нож (‘Iva cuts slices from the bread (with a knife)’).

3) S**yntactic alternations** (affecting the syntactic structure only), for example Passive alternation Хлябът е (на)рязан на филии с нож (от Ива) (‘The bread is cut (with a knife) (in slices) (by Iva)’).

4) U**nspecifications** (affecting the implicit realization of arguments), for example Реже хляб на филии с нож. (‘(She) cuts the bread (in slices) with a knife.’).

We consider that there is a clear distinction between diatheses, syntactic alternations and unspecifications (Koeva, 2007). An attempt has been made to define unambiguous criteria for a classification of Bulgarian alternations based on the analysis of overt features as follows: reduction of the semantic relation of the source subject (complement); modification of the semantic relation of the source complements; alteration of the syntactic categories and/or grammatical functions of the source subject (complements); change of the syntactic categories of the source subject (complements), reduction of verb transitivity, derivational linking between the source and the alternate verb lemma (Koeva, 2007).

One of the problematic question in the FrameNet description is the treatment of semantic alternations (diathesis). It is considered that the semantic alternations of a particular lexical unit belong to the same semantic frame. However, the representation of **Causation** (Ruppenhofer et al., 2006) constitutes an exception. We think that the alternations that involve a determinate reduction of a subject semantic relation have to be treated in a special way.

A syntactic alternation, on the other hand, can be viewed as a relation between a pair of syntactic structures within one valency frame that involves a reordering of the elements that affects their grammatical functions and syntactic categories (Koeva, 2007). The following parameters in the Bulgarian valency lexicon determine the set of syntactic alternations that are allowed for a lexical unit: grammatical class, argumentness (the property of a predicate to incorporate a specific number of variables that correspond to the arguments and their syntactical categories and grammatical functions), and selectional restrictions. For example, if the morpho-syntactic values for a given verb are specified as personal, transitive, imperfective and the syntactic structure is NP subject and NP object, and the selectional restrictions are respectively human/agentive for the NP subject and concrete/inanimate for the NP object, then the syntactic alternation se–passive (without a passive participle) is possible, for example the sentence Онези мъже строят тези къщи (‘Those men are building these houses’) alters with the sentence Тези къщи се строят от онези мъже (‘These houses are being built by those men’). Thus the syntactic frames specify the typological differences in lexicalization patterns between Bulgarian and English or other languages.

5. Support software

Initially we started with a software product specially designed for the Bulgarian valence lexicon – a web-based system called SYNText (SYNtactic dictionary Tool. The system allows fast and easy administration of the attributes and their values inside the linguistic modules and supports XML import and export of the data base.

At the moment we use the Altova Authentic® 2009[4], a free authoring tool that allows to capture, view and edit XML and database content (Figure 1).

Figure 1. The BulFrameNet development environment

The reprogramming of the SYNText system is required (if possible) in order to support new linguistic modules that have been added. The process of semantic annotation suffers from the lack of specialised software such as the FrameNet Desktop (Ruppenhofer et al., 2006), which provides rich semantic information by associating lexical units with semantic frames. Thus one of the important immediate goals is to supply the Bulgarian FrameNet project with appropriate software.

6. Availability

The Bulgarian FrameNet is being developed at the Department of Computational Linguistics at the Institute for Bulgarian Language. It is partially available online.

7. Conclusion

The paper outlines a conceptual model and inner structure of the Bulgarian FrameNet. The target is a corpus-based lexicon giving an exhaustive account of the semantic and syntactic combinatory properties of an extensive number of Bulgarian lexical units. The lexical entries (lexical unit evoking a semantic frame, its grammatical class and valency frame and annotated examples) in Bulgarian FrameNet can be grouped with respect to:

- a particular semantic frame at the cross-linguistical level;
- equivalent sets of morpho-syntactic features;
- equivalent sets of (combinations of) obligatory and allowed environments described by the syntactic categories, grammatical functions, and obligatory explicit usage;
- equivalent sets of (combinations of) obligatory and allowed environments described by selectional and

[4] http://www.altova.com/products/authentic/xml_db_form_editor.html

lexical-semantic restrictions.

Describing different semantic and syntactic environments of a lexical unit, it will be possible to go from the Bulgarian component to the English component (and to components for other languages) to compare how semantic frames are realised differently across languages. Complex interactions between lexicon and grammar should be taken note of and accounted for, in order to understand (and compare) how languages encode the same conceptual frames differently.

The Bulgarian FrameNet is linked to other lexical resources: the Bulgarian WordNet (through word senses), the Bulgarian inflectional dictionary (through inflection types), and the Bulgarian National Corpus (through annotation examples), making them work all together.

8. Acknowledgements

Bulgarian FrameNet is a research project funded by the Scientific Research Fund of the Bulgarian Ministry of Education, Youth and Science (Grant No. ДТК 02 / 53).

9. References

Baker et al. (2003) Baker, C.F., C.J. Fillmore and B. Cronin. The Structure of the FrameNet Database, International Journal of Lexicography, Volume 16(3), pp. 281-296.

Benešová et al. (2008) Benešová, V., M. Lopatková and K. Hrstková. Enhancing Czech Valency Lexicon with Semantic Information from FrameNet: The Case of Communication Verbs, Hong Kong, China: ICGL 2008: ICGL 2008 Proceedings of the First International Conference on Global Interoperability for Language Resources pp. 19-25.

Boas (2005) Boas, Hans C. Semantic Frames as interlingual representations for multilingual lexical databases. International Journal of Lexicography 18(4): pp. 445-478.

Boas (2009) Boas Hans C. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. In Boas, H. C. (ed), Multilingual FrameNets in Computational Lexicography: Methods and Applications. Berlin: Mouton de Gruyter. pp. 59-99.

Burchardt et al. (2006) Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó and M. Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In Proceedings of LREC 2006, Genoa, Italy, pp. 969–974.

Ellsworth et al. (2004) M. Ellsworth, K. Erk, P. Kingsbury, and S. Padó. PropBank, SALSA and FrameNet: How Design Determines Product. Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora, LREC-2004, Lisbon.

Johnson et al. (2002) Johnson, C., C. Fillmore, M. Petruck, C. Baker, M. Ellsworth, J. Ruppenhofer, and E. Wood. FrameNet: Theory and Practice. International Computer Science Institute, Technical Report-02009. Berkeley, CA.

Fillmore (1982) Fillmore, C.J. Frame Semantics. In Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co. pp. 111-137.

Fillmore (1985) Fillmore, C.J. Frames and the semantics of understanding. In Quaderni di Semantica 6(2), пп. 222-254.

Fillmore et al. (2003) Fillmore C. J., C. R. Johnson, and M.R.L. Petruck. Background to FrameNet. *International Journal of Lexicography* 16(3), pp. 235-250.

Koeva (2004) Koeva, Sv. Theoretical model for a formal representation of syntactic frames, Scripta and e-Scripta, Vol. 2, Sofia, pp. 9-26.

Koeva et. al. (2006) Koeva, Sv., Sv. Leseva, I. Stoyanova, E. Tarpomanova, and M. Todorova. Bulgarian Tagged Corpora, In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, Sofia, pp. 78-86.

Koeva (2007) Koeva, Sv. Bulgarian Alternations – Lexicon or Grammar? In: Southern Journal of Linguistics 29, pp. 49-76.

Koeva (2008) Koeva, Sv. (ed) Bulgarian FrameNet. Semantic-syntactic dictionary of Bulgarian, Sofia, pp. 104.

Lönneker-Rodman (2007) Lönneker-Rodman, B. Multilinguality and FrameNet. ICSI Technical Report TR-07-001. Berkeley, CA

Lönneker-Rodman et al. (2008) Lönneker-Rodman B., Collin Baker and Jisup Hong. The new FrameNet Desktop: A Usage Scenario for Slovenian. In Proceedings of ICGL 2008, the First International Conference on Global Interoperability for Language Resources, 9-11 January 2008, Hong Kong. pp. 147-154.

Ohara et al. (2004) Ohara, K., S. Fujii, T. Ohori, R. Suzuki, H. Saito, and S. Ishizaki. The Japanese FrameNet Project: An introduction. The Fourth international conference on Language Resources and Evaluation. Proceedings of the Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora", Lisbon, Portugal. May, 2004.

Padó (2007) Padó, S. Translational equivalence and cross-lingual parallelism: The case of FrameNet frames. In: Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages, Tartu, Estonia, pp. 39–46.

Ruppenhofer et al. (2006) Ruppenhofer J., M. Ellsworth, M. R. L. Petruck, and C. R. Johnson. FrameNet II: Extended Theory and Practice ICSI Technical Report.

Subirats & Petruck (2003) Subirats C. & M. R.L. Petruck. Surprise: Spanish FrameNet! In E. Hajicova, A. Kotesovcova & J. Mirovsky (eds.), Proceedings of CIL 17. CD-ROM. Prague: Matfyzpress.

Vossen (1999) Vossen, P. (ed) EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014.

Vossen (2004) Vossen, P. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Linual-Index. International Journal of Lexicography, 17(2), pp. 161–173.