On the Valency Frames of type Subject-Predicate in Bulgarian

Petya Osenova
Division of Bulgarian Language
Sofia University "St. Kl. Ohridski"
petya@bultreebank.org

Abstract

The paper presents some observations on the semantic constraints of the intransitive subjects with respect to the predicates they combine with. For these observations a valency dictionary of Bulgarian was used. Here two clarifications are to be made. First, the intransitive predicates are viewed in a broader perspective. They combine true intransitives as well as intransitive usages of transitive verbs. The complexity comes from the modeling of these verbs in the morphological dictionary. Second, the semantic constraints that are considered here, are limited to a set of semantic roles and build on the lexicographic classes of verbs in WordNet.

Keywords: intransitive verbs, semantic constraints, subject, lexicographic classes, valency dictionary

1. Introduction

The aim of this paper is to describe some of the syntactic and semantic varieties within the valency frames of type subject-predicate in Bulgarian with the help of a data-driven valency dictionary. The valency dictionary that was used here is the one built over the syntactically annotated corpus BulTreeBank (Simov et al., 2005). Also, some general semantic constraints were available over the grammatical role 'subject'. These semantic constraints include a set of basic semantic roles and general concepts. My aim is to exploit them for the construction of a more formalized and more detailed set of semantic roles in the future.

Let me first briefly introduce the valency dictionary for Bulgarian as described in (Osenova et al., 2012). The data-driven valency lexicon covers the verbs in the syntactically analyzed corpus of Bulgarian — BulTreeBank. It adopts a representation of the surface syntactic structure, and consists of constraints in the form of coarse ontological labels and semantic roles. The process of valency lexicon creation underwent several steps. First, all the verbs were extracted together with the sentences they have been used in. Then they were lemmatized and sorted by the lemma marker. A default valence frame was inserted that presents an example predicate with its core arguments: a subject (SUBJ), a direct object (DIROBJ) and an indirect object (INDOBJ). Since the default valence frame obviously cannot match all the real frames, a manual checking was performed afterwards for the purposes of frame repair and validation.

Here I am interested in frames that have only one grammatical role — Subject. The other roles might by anything but direct object, because it is the well-known marker of transitivity. In principle, the verbs of interest should be the intransitive ones only, i.e. verbs that do not have a direct object. However, since the used valency dictionary followed the surface realizations of the verbal arguments in the corpus, the intransitive verb group is actually wider.

Note that verbs with clausal objects are considered intransitive. The intransitive group includes also transitive verbs that underwent de-tranzitivization under various circumstances (for example, reflexivization, de-causativization, lexical shifts, etc.) and thus can be used as intransitives. In this paper I will not

dwell into the specifics of all these processes that result in intransitive verb usages. I will just mention that different frameworks view these processes in various ways.

It should be noted that the dictionary presentation of verbs, especially the ones with the de-tranzitivizing particles се 'se' and си 'si' as well as the ones with optional arguments, is not trivial. On the one hand, this is due to the fact that grammar and dictionary have complex common interfaces that cannot be fully represented neither in the grammar, nor in the dictionary only. Thus, such a representation needs intermediate levels. On the other hand, there is no ideal way to deal with optionality of the arguments in discourse. Hence, far from trivial is also the relation between the dictionary representation and the text realization of these cases. Not surprisingly, there is vast literature on the specifics of се se- and си si-verbs together with the related phenomena on morphological, syntactic and semantic levels. See (Nitzolova, 2017), (Koeva, 1998), (Petrova, 2014) among others.

The paper is structured as follows: in the next section some more details about the valency dictionary are given. Section 3 outlines my observations on the distribution of certain types of subjects per semantic roles/types of predicates. Section 4 concludes the paper.

2. Valency Lexicon in Brief

The principles behind the valency lexicon are as follows: as mentioned above, the valence frames were kept to the surface syntax. However, the verb usage has been encoded only in active voice[1]. The verbs in perfective and imperfective aspects were encoded as separate lemmas following one of the two linguistic views within the Bulgarian grammar literature. The other one considers them as forms of the same lemma.

The frame includes only the inner participants (semantically obligatory for the event or situation, presented by the predicate, but might be unexpressed on the surface level) (Pustejovsky, 1998). According to Pustejovsky there are three types of arguments:

- true arguments (obligatory for the predicate on the syntactic level like in 'devour a sandwich')
- default arguments (optional on the syntactic level like in the sentence 'I like reading a book' and 'I like reading.')
- shadow arguments (expressed internally in the lexical semantics of the predicate like in 'I kicked the football [with my leg]'). The prepositional phrase 'with my leg' is presupposed by the verb 'kick', so its explicit realisation is possible only if some additional information is added like in 'I kicked the football with my left leg'.

All these argument types can have also intransitive usages. Note that the Bulgarian subject is considered a default argument in this analysis, i.e. it can be omitted on a regular basis but under certain circumstances. Thus, its explicit or implicit realization, although grammatically possible due to the rich verbal inflection, often depends on specific discourse-related conditions.

Based on the statistics from BultreeBank — (Osenova et al., 2012), the type with an explicit nominal subject that is of interest to me 'Subject (NP) - Predicate' comes third by frequency after the types 'Predicate - Direct Object (NP)' and 'Subject (NP) - Predicate - Direct Object (NP)'.

The construction of the valency frames included also the following steps: extracting examples from the treebank for the corresponding verb; classifying the verb with respect to one of the 15 lexicographic classes in WordNet through the BTB-WN (Osenova and Simov, 2018b); making semantic abstractions over the examples with respect to a general ontology and the transferred typical semantic roles based on VerbNet[2]. Note that the semantic abstractions are still very general and that the set of semantic roles is not exhaustive. It includes the following roles that vary across classes: Agent, Patient, Experiencer, Theme, Goal, Locative, Cause. Also, it much be taken into account that the semantic roles were assigned automatically to the verb arguments and then manually fixed. So, the data is still not completely refined.

[1] With the exception of cases where the predicates do not have active voice.

[2] https://verbs.colorado.edu/verbnet/

The frequencies extracted from the valency dictionary are as follows: from 1928 verbs in the valency dictionary, 520 verbs are intransitive by type or by usage which makes approximately one-fourth of the cases. From them 342 are true intransitives (including intransitive usages) and 178 are de-transivised with the reflexive particle се 'se'.

3. Observations

As already mentioned above, in order to get oriented within the predicate types, the lexicographic classes of verbs from the Wordnet were used. These 15 classes are listed below. Their occurrences in BulTreeBank (215 000 tokens) are given in the brackets according to the information reported in (Osenova and Simov, 2018a):

- verb.communication (283)
- verb.social (222)
- verb.stative (219)
- verb.motion (204)
- verb.cognition (203)
- verb.change (184)
- verb.possession (130)
- verb.contact (97)
- verb.creation (95)
- verb.perception (86)
- verb.competition (63)
- verb.emotion (53)
- verb.body (41)
- verb.weather (14)
- verb.consumption (13)

The total number of the annotated classes is 1907.

The initial semantic restrictions on the nominal groups were based on the SIMPLE lexicon ontology[3]. Below a very small part from it is shown in a simplified flat manner.

```
Person
Organization
Animal
Plant
Physical Object
Artefact (social/cognitive)
    Clothing
Event
Activity
Location
```

[3]http://webilc.ilc.cnr.it/clips/Ontology.htm

From the list of labels, observations were made on the following ones only: Person, Animal, Plant, Artefact and Event. It should be noted that at this stage Organization was subsumed by Person and Activity by Event.

The truly intransitive verbs as well as intransitive verb usages, show the following distribution of the respective nominal subject types:

- 234 subjects with the label Person
- 38 subjects with the label Event
- 34 subjects with the label Artefact
- 14 subjects with the label Animal
- 9 subjects with the label Plant

It can be seen that the most frequent type is Person, then almost equally often come Event and Artefact. Finally, with the fewest occurrences are Animal and Plant. Again, it should be taken into account that the corpus is mainly news media and partly literature. This fact influences the distribution of the semantic constraints over subjects. However, apart from the fact that Person subjects prevail over the Event and Artefact ones, this observation is not very informative per se. For that reason I focus on the semantic roles of subjects of intransitive/de-tranzitivized verbs within the most frequent lexicographic classes: verb.communication, verb.social, verb.stative, verb.motion and verb.cognition. I will briefly introduce each group according to (Miller et al., 1990).

3.1. Verb.communication Subjects

Verbs of communication are considered as: "verbs of verbal and nonverbal communication (gesturing); the former are further divided into verbs of speaking and verbs of writing [...] verbs referring to animal noises (neigh, moo, etc.) and verbs of noise production and uttering that have an inanimate source and lack a communicative function (creak, screech)." (p. 58). This class is expectedly the most frequent one in our news media corpus.

From 283 verbs 60 are with intransitive usages. This is around one-fifth of the cases. Here come verbs like броя (count), бягам от (avoid, escape), изпитвам (exam), наричам (name), говоря (speak), договарям се (negotiate), etc. Most of the subjects are AGENTS with a constraint persons. This cluster includes also the role of EFFECTOR and other ones that can cause an event, but are not persons. Rarely there occur other types. For example, animals (the verb вия (howl) with a subject wolves); events (the verb гръмна (disclose) where the subject is a scandal, a secret, etc).

Let us look into some of the typical verbs. For example, the verb говоря (speak) has an intransitive usage in one of its senses, namely: make a speech. A person can speak in front of an organization, audience; for some time; from a certain place. A variant of this verb is the perfective one заговоря (start speaking). However, more frequent is its subjectless impersonal usage in se-passive with an indirect object: В града се заговори за нея 'In town-the se.REFL spoke about her' (In the town they spoke about her).

The verb потека (spread, circulate) has as its subject an artefact (THEME): После потекоха компроматите 'Then leaked compromising-material-the' (Then the compromising material was disclosed).

Figure 1 shows an example from the valency dictionary visualized in XML in the CLaRK System[4].

The screenshot shows the verb гръмна (disclose) with a subject скандал 'scandal'. The notations are as follows: 'FD' stands for a Frame in the Dictionary; 'l' encodes the lemma; 'def' gives the definition; 'F' presents the general semantic constraint over the subject which says 'event discloses'; 'FSRL' encodes the semantic role AGENT; 'en' gives the link to this meaning of the verb in Princeton WordNet; 'senses' outlines the Bulgarian definition; 'tok' provides examples from the treebank.

[4]http://bultreebank.org/en/clark/

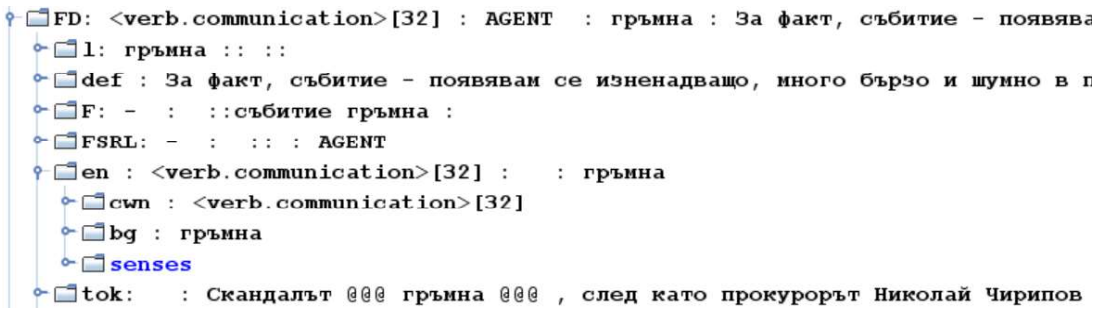

Figure 1: Verb.communication Subjects

3.2. Verb.social Subjects

This group refers to "verbs from different areas of social life: law, politics, economy, education, family, religion, etc. Many have a specialized meaning, restricted to a particular domain of social life, and they tend to be monosemous". (pp. 60–61)

This is the second most frequent type in the corpus. From 222 verbs 40 are with intransitive usages. This also makes approximately one-fifth of the cases.

One of the typical verbs here is действам (act, perform an action). It has three main occurrences: a) as it is: За да действа, човек трябва да говори 'In order to act, person must speak'; b) with a se-particle: Трябва да се действа 'Must to se.REFL act' (One has to act), and c) as an attributive present participle: действаща военна структура, 'acting military structure' (an active military unit). Another typical verb is работя (work). People mostly work at some position, or at some organization, or in some place, for some time, with some device. Here come also verbs as сгреша (sin), служа (serve, do military service), сътруднича (collaborate), etc. The verb справям се (cope, manage) presents either frames without any participants apart from the subject (manage), or with an indirect object (cope with something) and with adjuncts (typically adverbs of manner).

Among the usages there are a number of idioms, such as потъна (fall through, collapse). This holds also for the other verb classes.

In spite of the predominance of the person constraint within the AGENT role there occur also some social verbs whose subject is different. For example, спомагам/спомогна (help). In the following example the subject is an event and the semantic role is not AGENT but a kind of EFFECTOR: Физическите натоварвания ще спомогнат за повишаване на тонуса 'Physical-the exercises will help for increasing of tonus' (Physical exercises will make one fit).

3.3. Verb.stative Subjects

This group includes "for the most part verbs of being and having. Many stative verbs also have non-stative senses that have been placed into other files." (p. 60)

It is the third largest type in the corpus. From 219 verbs approximately one-half exhibits intransitive usages (i.e. around 100). Moreover, in this group the AGENT subject role more often is alternated by the roles PATIENT and THEME.

Concerning the AGENT subjects, person is typical for verbs like гостувам (visit as a guest at some place, organization, event); присъствам (attend), etc.

As for the roles other than AGENT, there is a big variety of semantic constraints, mostly of type THEME. For example, the verb действам (apply, hold) has as its subject some artefact (legal text, legal document, project, contract, etc.): Редът е такъв, откакто действа новият Закон за държавната собственост 'Order-the is such, from-where applies new-the Act for state ownership' (This has been the case since the new Act for the State Property came into force).

Some event can also take the subject role with verbs like бавя се (prolong). For example: Ремонтът на летището се бави 'Renovation-the of airport-the se.REFL late' (The renovation of the airport is delayed). Thus, the semantic role is a THEME. Another example of a THEME role subject is the verb

водя (lead, go) with a subject that is a street, path, road. See: Пътят води към върха 'Road-the leads towards peak-the' (The road leads to the peak). More examples refer to verbs, among which: предстоя (impend), приключвам/приключа (end, stop).

There are cases in which the verb can take as subjects both - AGENT (person, organization, state) and THEME (event, artefact, object, etc.). For example, the verb идвам (follow): На следващо място идва този аргумент 'To next place comes this evidence' (Next comes this argument). Another verb is липсвам (be absent): Липсва добрата алтернатива 'Lacks good-the alternative' (There is a lack of a good alternative). More verbs are: оставам (endure, persist), преобладавам (predominate, loom), принадлежа (belong), служа (serve), etc.

3.4. Verb.motion Subjects

The motion verbs "derive from two roots: move, make a movement, and move, travel". (p. 59)

This is the fourth most frequent group of verbs in the corpus. From 206 verbs 150 are with an intransitive usage. Thus, within this group of typical verbs of moving and acting the intransitives do prevail as expected.

The AGENT role with a person constraint but allowing also other ontological concepts like animal is typical for verbs like бягам (leave, exit), вървя (walk), идвам/дойда (arrive). The generalized AGENT role can combine various constraints: persons/vehicles (plane)/celestial bodies like обикалям (circle); person/artefact/vehicle like потъвам/потъна (sink); person/vehicle/bird like пътувам (travel) or person/event/activity like стигам/стигна (reach): Докъде стигна работата по случая? 'To where reached work-the on case-the?' (What is the status of the work on this case?). Such cases have to be refined with respect to the specific semantic roles. Here come also verbs with restricted subjects other than person AGENT like бия (heart beats), изминавам/измина (time elapses).

3.5. Verb.cognition Subjects

This group includes "verbs denoting various cognitive actions and states, such as reasoning, judging, learning, memorizing, understanding, and concluding". (p. 59)

This is the fifth most frequent group in the corpus. From 203 verbs only 50 are with an intransitive usage which makes one-fourth of the cases. Here the subject roles are labeled exclusively EXPERIENCER. A typical EXPERINCER person subject belongs to verbs like: знам (know, cognize), надниквам/надникна в нещо (get through, sink in), научавам/науча за нещо (learn, hear), мисля (think, judge): Тя го мисли за глупав човек 'She thinks him.ACC for stupid person' (She thinks that he is a fool).

The combination of EXPERIENCER subjects that are persons with oblique participants possessing a GOAL role are verbs like отстъпвам/отстъпя от позиция (abondon, give up): гледам на нещо по някакъв начин (consider): Политиците гледат практично на нещата 'Politicians look practically on things-the' (Politicians view everything from a practical point of view).

4. Conclusions

The paper presents some observations on the combination of certain semantic types/roles of subjects in 5 lexicographic classes with intransitive predicates.

Within these most frequent types the verb.communication and verb.social exhibit predominantly AGENT subjects with a person constraint.

Verb.stative type increases the intransitive frames and also the PATIENT/THEME subject roles. Verb.motion keeps the AGENT subjects as majority similarly to verb.communication and verb.social, but like verb.stative it has prevailing numbers of intransitive frames. The only type among the five most frequent ones in the corpus – verb.cognition – imposes the EXPERIENCER subject role within the group of not so many intransitive cases.

Depending on the verb meaning, the frame can have a more specific or a more general set of semantic constraints/roles. Since the valence dictionary presentation of frames is data-driven, it requires more

work on the proper mappings among the lexical meanings, verb valencies and semantic labels of the verb arguments.

Acknowledgements

This work was partially supported by the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019.

References

Koeva, S. (1998). Reflexive, passive, optative, reciprocal and impersonal verbs in Bulgarian. In *Nauchni trudove na Plovdivskiya universitet – Filologiya, 36, 1*, pages 142–157.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. 3:235–244.

Nitzolova, R. (2017). *Bulgarian grammar*. Berlin: Frank and Timme GmbH.

Osenova, P. and Simov, K. (2018a). Enriching valency frames lexicon of Bulgarian with Semantic Roles. In *Slovanská lexikografie počátkem 21. století. Sborník z konference Praha 20. – 22. 4. 201*, pages 239–246.

Osenova, P. and Simov, K. (2018b). The datadriven Bulgarian WordNet: BTBWN . In *Cognitive Studies, Études cognitives, 18*.

Osenova, P., Simov, K., Laskova, L., and Kancheva, S. (2012). A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2636–2640.

Petrova, G. (2014). Medialni glagoli s refleksivna semantika. In *Nauchni trudove na Rusenskiya universitet – volume 53, series 6.3*, pages 36–40.

Pustejovsky, J. (1998). *The Generative Lexicon*. The MIT Press.

Simov, K., Osenova, P., Simov, A., and Kouylekov, M. (2005). Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation, Special Issue*, pages 495–522.