Lexical Markers on the Informational Structure Level*

Yovka Tisheva, Marina Dzhonova
Bulgarian Language Division
Faculty of Slavonic Languages, St. Kl. Ohridski University, Sofia, Bulgaria
yovka@slav.uni-sofia.bg, marina@slav.uni-sofia.bg

ABSTRACT The collaboration between the linguist and the corpora is of major significance for the corpora building. First of all, thanks to language corpora a new style of theoretical research is established. The linguists study not only their intuition, but can base their analysis on data from different type of corpora. Second of all, as potential users linguist can suggest which language phenomena are representative and what kind of linguistic information could be necessary for them to be put in the corpora. We propose a model for using the data from several Bulgarian corpora when studying the lexical markers of the information structure. We claim that the so-called lexical contrastive focus marked overtly by the particle *samo* ´only´ in Bulgarian is closely related to the sentence prosody. The focus marked by *samo* cannot be unambiguously defined by word order or constituent order. *Samo* can affect the phrases it immediately precedes or follows. Consequently, the information about the sentence prosody is an important means for the recognition of the phrase in the scope of samo which could be NP, PP, AdvP, or VP. In this respect we claim that one of the important features for a spoken corpora to be representative is the availability of the audio data, which provides the linguist with information about the focus structure of the sentence. An analysis based on corpora data and additional information about the sentence prosody can help in elaborating the set of features that should be used when describing exhaustively the language specific phenomena at different layers.

1. Introduction

Grammar books and articles could show various attitudes of linguists towards language corpora. They may start from complete ignorance and end as real obsession. What everybody must admit is the fact that thanks to language corpora a new style of theoretical research is established. The linguists study not only their intuition, but can base their analysis on data from different type of corpora.

There are many requirements for the language corpora, i.e. related with its interface: to be "user friendly", easy to access and manage, free, etc. More important are the requirements concerning its content. The language corpora must be rich, to cover as many language phenomena as possible, to present data from different texts and genres, from different communicative situations. "In building a corpus of a language variety, we are interested in a sample which is maximally representative of the variety under examination, that is, which provides us with as accurate a picture as possible of the tendencies of that variety, including their proportions." (McEnery and Wilson 2001:30)

For the work of the linguists most important is the fact that representativeness and reliability must be intrinsic characteristics of the corpora. In order to develop such kind of corpora more new texts could be added. Enriching the texts included with relevant linguistic information seems to be the second way of developing the corpora. This will enable more linguists to use successfully the data when looking for judgments or examples to test different ideas and analyses. As potential users linguists can suggest which language phenomena are representative, what kind of information could be necessary for them to be put in the corpora, to propose the way for annotation of the texts, etc.

In this paper we propose a model for using the data from several Bulgarian corpora[1] when studying the lexical markers of the information structure, especially the focus, in Bulgarian corpora. Topic and Focus of the sentence could be marked by lexical, morphosyntactic and prosodic means. We claim that the so-called

* We are grateful to K.R. Hauge for his insightful comments and useful suggestions on an earlier version of this paper. Any other errors are, of course, our responsibility.

[1] The available resources for Bulgarian on which our research is based are the following: the corpora of spoken Bulgarian collected by Kr. Aleksova (http://www.hf.uio.no/east/bulg/mat); the Bultreebank corpora of Bulgarian language (www.bultreebank.org); the text archive of Bulgarian jokes collected by K.R. Hauge.

lexical contrastive focus is closely related to the sentence prosody. When a single sentence is uttered the most clear indicator of focus is the major sentence stress. The speaker marks out by prosodic means (kind of emphasis or major tonic) one part of the utterance, which he wishes to be interpreted as informative. The stress marks out a syntactic domain in which focus is expressed. The linguistic means used to mark different parts of information structure are specific for each language. In Bulgarian the focalisers (focusing lexemes which attract the major stress) like *samo* 'only', *daže* 'even', *dori* 'even', question particle *li* etc. can freely associate with different phrases within the sentence. Some of them i.e. *samo* 'only' can take scope in both directions - over the preceding or the following element, and sometimes over a distant element. In this case it is hard to distinguish Focus of the sentence only by studying the position of the focalisers. The sentence prosody and the position of the intonation centre of the sentence are not relevant for the analysis of printed texts. In order to make the analysis of lexical contrastive focus more precise the data from a corpus of spoken Bulgarian also have been used. We argue that the lexical contrastive focus could be defined as a combination of lexical and prosodic means.

The paper has the following organisation: after a brief description of the notion of information structure and focus in particular (Section 2), we propose (Section 3) a model for analysis of the lexical contrastive focus with *samo* 'only' in Bulgarian simple sentences. The use of the focaliser *samo* in complex sentences is discussed in the following part (Section 4). Finally (Section 5), some special cases are studied.

2. **Information Structure. Notion of Focus**

Languages offer different means of syntactic expression for a given predication. Since they are part of the grammar system of the given language they are specific for each language. What differs the options afforded by the particular grammar is not due to the morphosyntax only. Syntactic expressions also differ significantly in their communicative value. When analysing the Information structure we have to deal simultaneously with formal and communicative aspects of the language, because the Information structure encoded by certain sentence forms shows the mental status of the speakers and the hearers. The difficulties encoded in the analysis of the Information structure component of the grammar are reflected in certain problems of terminology. "Among the labels which have been used are FUNCIONAL SENTENCE PERSPECTIVE, used by scholars of the Prague School of linguistics, INFORMATION STRUCTURE OR THEME (Halliday), INFORMATION PACKAGING (Chafe), DISCOURSE PRAGMATICS and INFORMATICS (Vallduvi)" (Lambrecht 1994:2).

For the analysis proposed here the definition of Information structure given by Lambrecht is adopted: "INFORMATION STRUCTURE: That component of sentence grammar in which propositions as conceptual representations of states of affairs are paired with lexicogrammatical structures in accordance with the mental states of interlocutors who use and interpret these sentences as units of information in given discourse contexts. The information structure of a sentence is the formal expression of the pragmatic structuring of a proposition in a discourse" (Lambrecht 1994:5).

When talking about Information structure in the context of corpora Prague Dependency Treebank (PDT) is a good example how to incorporate the data from different language levels, especially in the corpus of Slavic language. PDT is part of Czech national corpus and has been developed through the approach to grammar established by the Prague School of Linguistics. Three levels are distinguished: grammatical structure of the sentence; semantic structure of the sentence; the level of organisation of the utterance. The extra-linguistic means of organising utterance (e.g. rhythm, intonation, the order of words and of clauses, some lexical devices, etc.) act also at the third level. PDT has a three-level structure; the highest level of annotation for PDT is characterised as the "level of linguistic meaning." "Syntactic tags on the tectogrammatical layer (TGTSs) capture the deep (underlying […]) structure of the given sentence, i.e. its dependency based syntactic structure proper" (Hajičová 2000:49). Two main concepts are important for the analysis of Topic/Focus Articulation (TFA): contextual boundness - corresponding to the ideas of given and new information, well-known from the works of Prague-school structuralists; communicative dynamism - corresponding to the surface word order. In the tectogrammatical structures TFA is

represented by the left-to-right order of the nodes denoting the communicative dynamism and by the index, attached to the verb to denote whether it is contextually bound or non-bound. The definition based on the framework of Functional Generative Description (developed by Sgall, formalisation done by Petkevič) are as follows: Topic is a non-contrastive contextual bound node which always has a lower degree of communicative dynamism than its governor; Focus is a non-bound node; Contrast is a contrastive contextual bound node.

In this paper we will not analyse the deep structure of the sentence (as it is in PDT), semantics, context, or discourse, but lexical means indicating different features of a constituent called contrastive Focus. The claim that communicative dynamism corresponds to the surface word order, hence the place of the word (or node) could be significant for its information value cannot be adopted directly for the analysis of Bulgarian data because the focalisers could have different positions in the sentence. Although Bulgarian is regarded as SVO language, postverbal S or its pro-droppness are quite often observed. SVO order is often transformed to other linear orders due to the information structure of the sentence. Another reason not to use the theoretical framework of PDT is due to the grammar systems. Bulgarian differs significantly from other Slavic languages. For our work it is important to mention some of these differences: Bulgarian has lost its declination system (unlike Czech which has a rich inflectional system and cases can easily give information about syntactic functions); Bulgarian possesses a group of language-specific particles that mark interrogativity as well as focusation; Bulgarian has developed a formal marker for topicality, i.e. the Clitic Left Dislocation (ClLD), expressed by the clitic doubling constructions.

2.1. **The Notion of Focus**

The definition of focus as "The semantic component of a pragmatically structured proposition whereby the assertion differs from the presupposition" (Lambrecht 1994:213) is a definition of a semantico-pragmatic category. If the Focus is "the information that is asserted about the topic" (Buranová 2000), we will face a new problem: not all sentences have a topic, therefore focus cannot simply be defined as the complement of topic. It will be the same if we assume that focus has to do with the conveying of new information, and that all sentences convey new information, all sentences must have a focus. For our analysis it is more convenient to define focus as a relation between a constituent X and a focusing operator Y. The non-focused part of the sentence is the background of Y. The focusing operator Y may be overt as with such particles as *only, even* in English, and the contrastive negation - *not John, but Peter* (see Primus 1993).

One way of testing focus is to look at pairs of wh-questions and their answers - the focus of the answer will correspond to the wh-phrase from the question (2). In some cases like in (1) the whole sentence is focus, because as a unit it corresponds to the question.

(1) - Kakvo stana?
What happened?
- [DECATA SČUPIHA PROZORECA.][2]
The kids broke the window.

(2) - Koj sčupi prozoreca?
Who broke the window?
- DECATA sčupiha prozoreca. / DECATA go sčupiha.
The kids broke the window. / The kids broke it.

The focusation is also closely related with interrogativity; question particles and wh-words in Bulgarian play certain role on the level of Information structure (cf. Krapova 2001, Tisheva 2001, Tisheva and Dzhonova 2002). Focusing may affect categories like short noun phrases or clausal nodes. Within different frameworks, focus has been related to syntactic constituents, to individual lexical items, and to semantic entities (see Taglicht 1993). We adopt the idea that the unit (phonological entity) which takes the focus-

[2] Focus of the sentence is noted by capital letters.

marking accent is a syntactic constituent. The focus particle *samo* 'only' in Bulgarian which we analyse in the present paper attracts the focal stress and assigns it to the constituent (phrase) formed with the element over which the particle has its scope.

For the analysis proposed it is more reliable to adopt the idea that Topic and Focus of the sentence are formal features connected with or carried by certain phrases in the sentence. Rizzi (1997) proposes an analysis of formal structure of the sentence.[3] The latter has a finer structure and should be split into several projections or phrases. The complementiser system is "the interface between a propositional content (expressed by the IP) and the superordinate structure (a higher clause or, possibly, the articulation of discourse […]). As such, we expect the C system to express at least two kinds of information, one facing the outside and the other facing inside" (Rizzi 1997:114). He establishes four phrases (projections) in the CP field:

... Force ...(Topic)...(Focus).... Fin IP (from Rizzi 1997)

The phrase on the left boundary determines the sentence type or the illocutionary force of the sentence - declarative or interrogative. On the right boundary is the verb, namely the specifications concerning its "finiteness" - person, number, tense, mood etc. Between these two boundaries Rizzi includes a system of phrases related to the Information structure represented by the interaction of Topic and Focus. It is important to point out that "topic-focus system is presented in a structure only if "needed" […]. If topic-focus system is activated, it will inevitably be "sandwiched" in between force and finiteness, as these two specifications must terminate the C system upward and downward, in order to meet the different selectional requirements" (Rizzi 1997:119). Later a more detailed scheme was proposed, including a special position for Interrogative Force:

Force Top* Int Top* Focus Mod* Top* Fin IP (from Rizzi 2001:21).

For the analysis proposed in this paper it is important to point out that Focus is unique, while there could be more than one Topic in the sentence. As for the mutual ordering of these phrases, there could be Topic phrases before or after Focus.

2.2. The notion of contrastive focus

After the brief overview of the concepts of information structure within the different frameworks, we focus our attention on one particular type of focus, i.e. the lexical contrastive focus. For our aims we will follow the analysis of lexical contrastive focus proposed by Dryer (1994), Kiss (1998), König (1991) and for Bulgarian - Nicolova (2001). The characteristics of focus particles and lexemes are already investigated for some other languages (for English cf. König 1991, Dryer 1994, for German - Buring and Hartmann 2001, for Hungarian - Kiss 1998).

In the theoretical framework proposed by Kiss (1998) the contrastive focus again is considered a separate type of focus. She uses the term *identificational focus* and lists a number of characteristics that differentiate it from the *information focus* (known also as presentational or wide focus). She defines the identificational focus as follows: "An identificational focus represents a subset of the set of contextually or situationally given elements for which the predicate phrase can potentially hold; it is identified as the exhaustive subset of this set for which the predicate phrase actually holds" (Kiss 1998:245). There are some aspects in which the identificational focus differs from the informational focus. Important for our investigation are the following:

- The identificational focus expresses exhaustive identification;

- Information focus merely marks the nonpresupposed nature of the information it carries;

- The identificational focus does, information focus does not, take scope (Kiss 1998:248). While studying

[3] More precisely the complementiser system, e.g. words like English *that*. In Bulgarian words like *da*, *če*, *deto* are considered to be complementisers.

the functional aspects of focus particles König (1991) takes into account the fact that focus particles have their own scope. "The scope of a particle can roughly be described as the semantic counterpart of that part of a sentence that is relevant for spelling out that contribution [...] a general account of the meaning of focus particles must be based on a representation in which the focus and the scope of a particle are clearly distinguished" (Koenig 1991:31). The focus and its scope may coincide, but not always.

Dryer uses a similar framework for his investigation. He points out that *focus particles*, like *only* and *even*, unlike free focus, involve differences in truth-conditional meaning. He considers that there is no grammatical link between focal accent and the constituent which is associated semantically with *only*, i.e. the narrow focus; the situation is somewhat different with VP-focus or S-focus, which he calls *broad focus* (Dryer 1994: 2).

The functional aspects of the focus particles (*focalisers*) in Czech were analysed for the annotation of the information structure in PDT. "The prototypical syntactic position of a focalizer is that of a dependent of a verb node […] it is possible to specify the scope of the focalizer as the whole subtree subordinated to the verb (where […] the verb itself is included)" (Buranová 2000:4). According to this the scope of the focalizer is always to the right. The use of focus particle *only* in Bulgarian shows that restriction concerning the direction in which the focalizer may take scope does not exist.

Bulgarian data are analysed by Nicolova (2001). According to her the *contrastive focus* is phonetically the most prominent element of the sentence and it is pronounced with emphasis. It can be divided into two formal types: *suprasegmental contrastive focus* (*free focus* according to Dryer 1994) and *lexical contrastive focus* (*bound focus* in Dryer 1994). The latter is marked by suprasegmental/prosodic means and in addition with certain lexemes that function as operators (Nicolova 2001:77). For our investigation we adopt her definition of contrastive focus and analyse *samo* as marker for the in Bulgarian.

2.3. The lexical contrastive focus with *samo* 'only' in Bulgarian

Our investigation concerns the lexical contrastive focus with *samo*. There are a number of investigations on the *only*-focus in English. In his study on the focus particles König (1991) classifies *only* as exclusive, non-scalar particle. Nicolova (2001) also considers the Bulgarian counterpart of *only* - *samo* an exclusive focus particle.

Our analysis of *samo* in Bulgarian led us to the assumption that it is not possible to generalise one intrinsic feature for the particle. The existence of homonyms in the language or the polifunctionality accounts for the possibility of using *samo* in many constructions, phrases or clauses with different functions. It is important to describe all contexts where *samo* could be found in order to predict the possible ambiguities. We believe the detailed description of empirical data will benefit the work on developing a methodology for their disambiguation.

There are few types of approaches in respect to *samo* in Bulgarian. The grammars treat *samo* as adverb, specifying the sentence modality (GSBKE 1998:242). From the other side, *samo* can function as a derivation particle in some compound conjunctions and expressions (*ne samo no i* 'not only…but'). The information structure of the sentences containing those expressions is still not investigated. They form a close set of words and must be included in the lexicon as multiwords. Here we list those expressions*: ne samo ... no i, samo če; samo kogato, samo zaštoto, samo kâdeto, samo ako, samo kato; ot samo sebe si, samo i samo da*. Most of these expressions (except for *ot samo sebe si*) partake in the clause type marking. For the future work on adding the syntactic information in the corpora our proposal is to include these phrases in the lexicon and to formulate a rule for the clause type marking.

What is important for our analysis is that *samo* may function as focus particle, overtly marking the focus element in the information structure of the sentence /e.g. the *narrow* focus/ (Nicolova 2001). In this respect it is important to mention that *samo* is an operator whose scope is the phrase it is adjoined to.

As it was already mentioned, the focusing features of *samo* in Bulgarian have already been investigated. Why than it is important to have a separate investigation of the lexical contrastive focus with *samo* in respect to the Bulgarian corpora? First of all, it is not possible to postulate a simple rule for the position

samo may occupy in Bulgarian sentence. Second of all, we attempt to apply our investigation in corpora of colloquial speech. This raises some problems due to the specificity of this type of communication. As McEnery and Wilson (2001:44-45) point out "In speech there is no explicit punctuation: any attempt at breaking down spoken language into sentences and phrases is an act of interpretation on the part of the corpus builder. One basic decision which needs to be made with spoken data is whether to attempt to transcribe it in a form of orthographic sentences or whether to use intonation units instead, which are often, though not always, coterminous with sentences. There also follows from the intonation unit/sentence decision the issue of whether to attempt to further punctuate spoken language or leave it at the sentence level without punctuation." In the Bulgarian corpora available on the Internet the intonation units are signed by longer or shorter pauses. But it is important to mention that Bulgarian punctuation is not based on intonation principles, so there is no possibility to use the same rules we use for printed texts. We propose the prosody, particularly the primarily sentence stress, to be added in spoken corpora as a distinctive feature for the contrastive focus. In our opinion this is important for spoken corpora to be representative.

In Bulgarian the focus particle *samo* can have scope over NP, PP, AdvP, VP, or part of XP. Usually it has scope over the following element, but it is often the case that the focus particle has scope over a preceding or a distant element (cf. Nicolova 2001). The focused part of the sentence can move freely, because Bulgarian is a free word order language. Nicolova gives the following examples:

(3) Petra običa [samo NEGO]F / [NEGO samo]F.

Petra only loves him (John)

(4) Petra [samo NEGO]F običa.

(5) [Samo NEGO]F običa Petra.

In these three examples the focus particle refers to the personal pronoun *nego* and marks it overtly as focus. This phrase could be situated at the end (3), at the beginning (5) of the sentence or to be sentence internal (4). When the phrase is at the end of the sentence there is possibility *samo* to precede or follow the pronoun *nego*. In both cases *samo* has scope over *nego*.

We analyse the properties of *samo* in three types of syntactic contexts: in simple sentences, in compound sentences and as part of some syntactic constructions.

3. *Samo* in the simple sentence

3.1. *Samo* + NP

Here we will investigate the possibility *samo* to scope over different types of phrases. In examples (3-5) *samo* takes scope over NP (*nego*), expressed by a personal pronoun. We provide additional data with *samo* having scope over NPs in order to study the different possibilities. In Bulgarian *samo* may take scope over the preceding or the following element. Thus in (6a.) *samo* focuses the noun immediately following it in the sentence *romani* 'novels', but there is an option for *samo* to follow this noun, without any change in the sentence meaning, nor in its information structure. In (6b.) the focus again is *romani* and it is overtly marked by *samo*.

(6) a. Četa samo ROMANI.

I read only novels.

(6) b. Četa ROMANI samo.

I read only novels.

When the focused noun is not the final phrase of the sentence, there are different possibilities for the scope of *samo*. In example (7a.) *samo* precedes the noun phrase *krastavici* 'cucumbers' and takes scope over it. But when the focus particle stands after *krastavici* it is more likely the focused constituent to be the phrase following the particle. In (7b.) that is the PP *za salatata* 'for the salad'.

(7) a. Kupih samo KRASTAVICI za salatata.

I only bought cucumbers for the salad.

(7) b. Kupih krastavici (samo) ZA SALATATA (samo).

I only bought cucumbers for the salad.

We may conclude that the position of *samo* in Bulgarian sentence is important for the identification of its scope. When NP occupies the sentence final position *samo* could precede or follow the phrase and in both cases it takes scope over it (6). *Samo* could only precede the NP in its scope when it occupies a sentence internal position (7a.).

3.2. *Samo* + PP

In (7b.) *samo* follows the sentence internal NP and consequently cannot take scope over it. The phrase next to the operator is PP, which occupies the sentence final position. *Samo* stands immediately before the PP, hence it takes scope it. The overt marker for the phrase in the scope of *samo* is the major sentence stress. *Samo* may also follow the PP and have scope over it. Here we provide more empirical data in order to examine whether the scope of *samo* depends on its sentence position. In (8a.) the PP *na Ivan* is an indirect object and it is focused by *samo*, which could precede or follow the PP.

(8) a. Obadih se (samo) NA IVAN (samo).

I called only Ivan.

There is one requirement for the position of *samo*. It should be situated immediately before or after the constituent it refers to. If we move *samo* to the beginning of the sentence (8a.) transforms to (8b.) in which *samo* doesn't take scope over the PP, but over the VP. In this respect Bulgarian differs from other languages like English.

(8) b. Samo se obadih na Ivan.

I only called Ivan.

In (8 c.) the focused PP is an adjunct and again it is possible *samo* to precede or follow the phrase. In both cases the particle functions as a focaliser.

(8) c. Šte rabotim (samo) V SRJADA (samo).

We will work only on Wednesday.

There is one restriction concerning the possible positions of the focaliser - it cannot be inserted between the components of the prepositional phrase. This restriction is illustrated by the ungrammaticality of (8 d.) and (8 e.).

(8) d. *Obadih se na samo Ivan.

(8) e. *Šte rabotim v samo srjada.

3.3. *Samo* + AdvP

Further, *samo* could take scope over AdvPs. In (9a.) the AdvP *tri pâti* 'three times' is the focus of the sentence, it occupies the sentence final position and receives the major sentence stress. The overt lexical marker for focus is the particle *samo*, which may precede or follow (9 b.) the AdvP. In both cases it refers to *tri pâti* and acts as a focaliser.

(9) a. Ženila sâm se samo TRI PÂTI. (9) b. Ženila sâm se TRI PÂTI samo.

I've gotten married only three times.

There is one restriction concerning the position of *samo* with AdvP in its scope namely that the particle cannot be inserted into the phrase. Example (9c.) illustrates this restriction. The ungrammaticality of the sentence is a result from the insertion of *samo* between the constituents of the AdvP *tri* and *pâti*.

(9) c. *Ženila sâm se tri samo pâti.

Samo can take scope over a part of AdvP and this part, not the whole AdvP, receives the major sentence stress. In (9d.) the word in the scope of *samo* is *vednâž* 'once'.

(9) d. Scâpi, ot tezi gâbi se jade samo VEDNÂŽ v života!

Darling, one gets to eat mushrooms like these just once in a lifetime!

When *samo* refers to part of the phrase it is possible not only the head, but also other elements within XP to be stressed.

The focus particle can also take scope over numerals as in (9e.). Again there is possibility *samo* to precede or follow the numeral. In both cases the numeral is in the scope of *samo*.

(9) e. Taksata e (samo) 25 $ (samo).

The charge is only/just 25 $.

3.4. *Samo* + VP

In Bulgarian the focus particle *samo* can have scope even over VPs. It is important to differentiate these constructions from those with NP, PP or AdvP in the scope of the particle. When *samo* precedes or follows VP different readings are possible. In (10a.) *samo* precedes the verb *predstavi* 'introduced' and has scope over it. In case that *samo* follows the verb (10b.) there are two possibilities: the particle can have scope over the verb or the following NP *Ivan*. In this respect Bulgarian differs from languages like English. In English if *only* precedes the verb it can have scope over the verb, the NP *Ivan*, or even the PP *na Marija*. In Bulgarian *samo* should immediately precede or follow the XP it refers to. Hence, in (10b.) *samo* cannot have scope over the indirect object *na Marija*.

(10) a. Pavel samo predstavi Ivan na Marija.

(10) b. Pavel predstavi samo Ivan na Marija.

Pavel only introduced Ivan to Maria.

When a clitic is part of the verbal complex as in (10c.) *samo* again has scope over the verb with the option to precede or follow it.

(10) c. Pavel (samo) gi zapozna (samo).

Pavel only introduced them to each other.

It is important to examine the position of *samo* in sentences with compound verb forms. In (11a.) the verb is in future tense and *samo* can precede or follow it. A sentence with *samo* inserted between *šte* and the verb (*sediš*) is ungrammatical as (11b.) shows.

(11) a. Ti (samo) šte sediš (samo).

You will just stay.

(11) b. *Ti šte samo sediš.

We may generalise that with compound tenses *samo* cannot be inserted between the components of the verb form. It could only precede or follow the verb, in both cases *samo* has scope over the verb.

There are few possible positions for *samo* when it refers to compound verb forms with modals: after the modal (12c.) or after the whole verbal complex (12d.). The restriction is that *samo* cannot precede the modal (*može* in 12c.). When the particle is postponed in respect to the verbal complex (12d.), various readings are possible. It may have scope over the preceding verb *bâde* or the following NP *ofisât na Microsoft*. When it refers to the NP, *samo* can also move after the phrase. In this case it could only be analysed as having the NP in its scope.

(12) c. Tova (*samo) može samo da bâde ofisât na Microsoft.

This could only be the office of Microsoft.

(12) d. Tova može da bâde (samo) ofisât na Microsoft (samo).

This could only be the office of Microsoft.

We may conclude that in Bulgarian simple sentences *samo* can operate over the phrase it immediately precedes or follows. When *samo* modifies VP, different readings are possible.

4. *Samo* in the complex sentence

In the complex sentence *samo* usually stands at the boundary between the matrix and the subordinate clause. *Samo* can refer to the preceding or the following phrase. The prosody is the only means to identify the phrase in the scope of the particle. Thus in (13a.) there are two verbs *iskam* and *da nauča* and *samo* can take scope over each of them. The phrase in its scope receives the major sentence stress.

(13) a. Iskam samo da nauča kak e vljazâl u nas.

I just/only want to find out how he has entered our place.

The analysis of (13b.) is similar. *Samo* stands at the boundary between the two clauses and can have in its scope the VP from the main clause *neobhodimo e* 'it is necessary' or the NP *plešivijat čovek* 'the bald-headed man' from the subordinate clause. The interpretation of the sentence depends on the context and the sentence stress is the overt marker, which shows the word in the scope of *samo*.

(13) b. Neobhodimo e samo plešivijat čovek da izpie flakon ot novoto lekarstvo.

The bald-headed man just needs to consume a bottle of the new medicine.

To summarise this section, in complex sentences the information about the prosody is the only means for the recognition of the scope of the particle *samo*.

5. Special cases

Samo may be used in constructions, phrases or clauses without acting as a lexical marker for contrastive focus. For the extraction of corpus data is important to describe all contexts where *samo* could be found.

Samo combined with imperatives and/or wh-words functions as a special tool forming exclamations. In (14a.) the construction 'imperative (*gledaj* 'watch') + samo + wh-word (*kakvi* 'what3pPl')' marks overtly the sentence type as an exclamation. In this case we cannot analyse *samo* as a focaliser with an element in its scope. Here *samo* is part of a syntactic construction, which is relevant for the identification of the illocutionary force of the sentence. The particle can freely move out of the construction towards the end or the beginning of the sentence (14b.) without changing its function.

(14) a. Gledaj samo kakvi gluposti razpravjat!

Just look what stupid things they are saying!

(14) b. (samo) Gledaj kakvi gluposti razpravjat (samo)!

Another instance for the discussed function of *samo* is (14c.). The difference between (14a.) and (14c.) is that the former is a complex and the latter is a simple sentence. Again the phrase is formed by 'imperative (*viž* 'look') + samo + wh-word (*kakva* 'what3pSgF')' and functions as a force marker. *Samo* can move to the beginning or to the end of the sentence, still functioning as part of the construction and not as a focaliser.

(14) c. (samo) Viž (samo) kakva krasavica (samo)!

Just look what beauty!

Other combinations with *samo* are possible. In (15a.) the expression marking the sentence as exclamation is 'wh-word (*kak* 'how') + samo'. The possibility *samo* to move towards the end of the sentence reveals the fact that the contact position of these two elements is not obligatory.

(15) a. Marija, kak (samo) me izplaši (samo)!

Maria, how you scared me!

The third possible combination (14d.) is 'samo + imperative (*počakaj* 'wait')'. The expression has

pragmatic function only. There are two possible uses of (14d.), depending of the context. *Samo* could be added to the sentence in order to express a threat. Besides *samo* could be used to express a type of politeness. Important means for the correct analysis of the pragmatic meaning of the utterance is the intonation.

(14) d. Samo počakaj malko!

Just wait a little!

From the analysis of the examples in Section 5, we can generalise the possibility *samo* to have pragmatic function or to partake in different constructions, taking part in the force marking of the sentence.

6. Conclusions

The focus marked by *samo* cannot be unambiguously defined by word order or constituent order. In Bulgarian *samo* can affect the phrases it immediately precedes or follows. The information about the sentence prosody is an important means for the recognition of the phrase in the scope of *samo* which could be NP, PP, AdvP, VP or part of XP. We argue that one of the important features for a corpus of spoken language to be representative is the availability of the audio data. The latter is important not only in the investigation of the phonetic aspects of the words, but also in the study of the information structure which reveals the communicative intentions of the speaker. In many cases the prosodic tools marking out the information structure are the only means for accurate disambiguation of the sentences.

It is not unusual copies of the actual recording of colloquial speech to be available in addition to the transcribed data in the corpora (e.g. the Lancaster/IBM Spoken English Corpus). As far as we know, this data is rarely if at all used for the investigation of the information structure of the sentence. It finds application mostly in the instrumental phonetic analysis. For our purposes the existence of the audio data is important and we propose it to become an obligatory part of the corpora of colloquial spoken Bulgarian. Such analysis can help in elaborating the set of features that should be used when describing exhaustively the language-specific phenomena at different layers.

References:

2. Buranová, E., Hajičová, E., Sgall, E. P. 2000. "Tagging of Very Large Corpora: Topic-Focus Articulation." *Proceedings of Coling 2000*, 278-284, Saarbrücken, Germany.
3. Buring, D., Hartmann, K. 2001. "The syntax and semantics of focus sensitive particles in German." *Natural Language and linguistic theory* Vol. 19:229-281.
4. De Kuthy, K. 2002. "The Information Structure of Discontinuous NPs in German." *Proceedings of the 8th International HPSG Conference*, Norwegian University of Science and technology. Frank van Eynde, Lars Hellan and Dorothee Beermann (eds.) CSLI Publications.
5. Dryer, M. 1994. "The Pragmatics of Focus-Association with *only*." (Unpublished paper delivered at the *1994 Annual Meeting of the Linguistic Society of America*).
6. Hajičová, E. 2000. "Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus." *T.A.L.*, Vol.41, n.1, pp. 47-66.
7. Kiss, K. È. 1998. "Identificational Focus versus Information Focus." Language. Vol. 74, 2. 245-272.
8. König, E. 1991. *The Meaning of Focus Particles*. London and New York: Routledge.
9. Lambrecht, K. 1994. *Information Structure and Sentence Form.* Cambridge: Cambridge University Press.
10. McEnery, T., Wilson, A. (eds.) 2001. *Corpus Linguistics: an introduction*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.
11. Nicolova, R. 2000. "Kontrastfocusierung mit Partiklen im Bulgarishen." *Lingwistische*

Arbeitsberichte 75. Leipzig, 103-115.

12. Nicolova, R. 2001. "Za kontrastnija fokus s častici v bâlgarskija ezik." *Sâvremenni sintaktični teorii.* Plovdiv, 76-85.
13. Primus, B. 1993. "Word Order and Informational Structure: A Performance-Based Account of Topic Positions and Focus Positions." *Jacobs, J. et al. (eds.), Syntax. Ein internationales Handbuch zeitgenőssischer Forschung*. Berlin, 880-896.
14. Rizzi, L. 1997. "The fine structure of the left periphery." L. Haegeman (ed.) *Elements of grammar.* Kluwer Academic publishers.
15. Rizzi, L. 2001. "Locality and Left Periphery." (manuscript)
16. Taglicht, J. 1993. "Focus and Background." *Jacobs, J. et al (eds.) Syntax. Ein internationales Handbuch zeitgenössischer Forschung.* Berlin, 998-1006.
17. Tisheva, Y. 2001. "Bulgarian yes-no questions with particles *nali* and *nima*." *Proceedings from FDSL4*. Potsdam (forthcoming).
18. Tisheva, Y., Dzhonova, M. 2002. "Information Structure and Clitics in TreeBanks." *Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol, Bulgaria.