Information Structure and Clitics in TreeBanks*

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

1 Introduction

In this paper we discuss the possibility to insert an additional feature [Information] in the characteristics of the words even at the POS level of annotation. Some lexemes as well as morphosyntactic constructions can play significant role in Topic-Focus articulation. This fact should be taken into consideration during morphosyntactic tagging. There are universal principles based on syntactic and pragmatic features of the constituents specifying the hierarchy of phrases within the sentence. The linguistic means used to mark different parts of formal and of information structure are specific for each language. In this study we will show the relations between features from the aforesaid levels and the possibilities to be associated with particular language means. In Bulgarian illocutionary force marking, e.g. interrogativity is closely related with information structure. The lexemes indicating interrogativity (*li, dali, nali, nima,* etc.) also take part in the topic-focus articulation. One of the language-specific constructions in Bulgarian - clitic doubling is a morphosyntactic way to mark constituents of the Information structure. Furthermore, when adding information about the word order the Information structure has to be taken into consideration. Bulgarian is regarded as SVO language, but postverbal position of the subject and its pro-dropness are quite often observed. Due to the topic-focus articulation the canonical schedule is often transformed to other linear orders.

While many morphologically annotated corpora of English and other languages are now available, syntactically annotated corpora are still rare. From this point, Prague Dependency Treebank (PDT) seems to be a good example how to incorporate Information structure into a corpus of Slavic language. PDT has a three-level structure; a closer look at its middle and highest levels (the analytic level of surface syntax and the tectogrammatical level) will enable us to answer the question whether this model could be applied directly for Bulgarian or not. 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 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. Three sets of attributes are used to label the nodes on this level: the lexical value of the word, the (morphosyntactic) grammatemes and the functors. "In addition […], there is an attribute *tfa* attached to each node, capturing the basic features of the topic-focus articulation of the sentence; the three values of this attribute are T(opic), F(ocus) and C(ontrats)" (Böhmová et al 2000:15). 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

* We are grateful to K.R. Hauge, P. Osenova and K. Simov and the two reviewers for their insightful comments and useful suggestions on an earlier version of this paper. All remaining errors are, of course, our responsibility. Spacial thanks to the BulTreeBank team for supplying our research with data from their text archive.

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 take a different theoretical framework. Our notion of Information structure has more formal orientation and we motivate this in details in the following section. We won't analyse the deep structure of the sentence (as it is in PDT), semantics, context, or discourse, but the overt surface morphosyntactic means indicating different features of the constituents. At this stage we avoid to propose a model for contrastive topic (or focus) in Bulgarian. For our aim we will follow the analysis of Information structure of Bulgarian (see e.g. Avgustinova 1998, Ivančev 1978, Krapova, Karastaneva 1998, Krapova 2001, Nicolova 2001, Tisheva 2001).

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. The word order as a basic formal criterion cannot be used because the constituents (or phrases) within the sentence have relatively free order. It is often the case that Focus precedes Topic. To explain this we need a reliable theory that conceives of syntactic structure a theoretical object representing the hierarchical relations between linguistic elements. When analysing the linear sequence of elements we apply the formal model proposed by Rizzi (1997; 2001).

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. The personal pronoun in Bulgarian possesses two forms for the oblique cases – a clitic and a full form. Some of the Slavic languages also possess both forms, but they are never used coreferentially in the same sentence.

The paper has the following organisation: after a brief description of the syntactic properties of the interrogative particle *li* and the ClLD in Bulgarian (Section 2), their relation with topic and focus is discussed, explaining why it is necessary to add this information simultaneously with the POS tagging (Section 3). We propose (Section 4) an additional feature for the discussed particle (*li*), i.e. the feature [Force] conveying information about the illocutionary force of the sentence, which can be useful at the level of syntactic annotation. Finally, we examine possible ambiguities due to the existence of homonymous forms or multifunctional particles in Bulgarian and propose a way for their disambiguation.

2 Information Structure. Notion of Topic and Focus

Languages offer different means of syntactic expression for a given predication. Since they are part of 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 language, because the Information structure encoded by certain sentence form 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).

The interest in how the language structures may function "transmitting" different sorts of information was leading for the scholars from Prague School of linguistics. Under the approach to grammar taken by the Prague School three levels are distinguished: grammatical structure of the sentence; semantic structure of the sentence; the level of organisation of the utterance – "this level makes it possible to understand how the semantic and the grammatical structures function in the very act of communication" (Firbas) (quoted from Lambrecht 1994:10). The extra-linguistic means of organising utterance (e.g. rhythm, intonation, the order of words and of clauses, some lexical devices, etc.) act at the third level as well. This approach probably accounts for the concept of three-level structured corpus (as Prague Dependency Treebank).

For the analysis proposed here we take the definition of Information structure given by Lambrecht: "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).

2.1. The Notion of Topic

There is no accepted definition for Topic and not even full agreement on the intuitions of what counts as topic. To refine the definition of topic is beyond our goals. We take one of the standard interpretations of topic - that of pragmatic aboutness. In this conception, topic is what the sentence is about. Other properties, like being background or given information, are not considered inherent to topics. This definition is rather general that's why some of the properties by which sentence topics can be identified in various languages could be mentioned. J. Jacobs (2001) describes Topic-Comment configuration as a prototype category, which has the following dimensions: information separation of Topic; predication; addressation and frame-setting. As one of the means to organise or classify information exchanged in the communication the topic also has pragmatic properties.

The formal properties of the topic are more important for our analysis than its semantic or pragmatic values. Topic can show different degree of (formal) syntactic integration into the rest of the sentence, from full integration (e.g. the topic is a verb argument) via loose integration (the topic is realised outside of the clause, but is co-indexed with an element in the clause). There are different methods of topic marking: by lexical or morphological elements, by certain types of syntactic structures, or by intonation. In some languages topics are identified by special markers, like Japanese *wa*, Korean *nŭn* (see Lee 2000), Turkish *ise* and *de* (see Primus 1993:881). Although the syntactic encoding of topic in natural languages varies in several ways there is a strong tendency topic to occur at the periphery of the sentence. There is a left-right asymmetry in the dislocation types available across the languages – topics to occur at the left periphery of the sentence (this position is perhaps grammaticalised) in Basque, Modern Greek, Russian, Turkish, while Ojiwba, Malagasy are topic-final languages (see Primus 1993). As we had pointed out, this approach was generalised by PDT when identifying topic and focus. Topics as contextual bound nodes with a lower degree of communicative dynamism than their governor are on the left part of the dependency tree.

The identification of the topic by linear order alone is not the most reliable one. Wh-words also occur on the left side of the sentence as will be shown when analysing Bulgarian wh-questions in Section 4.2., but they are associated with focus phrase. In some cases topic could be found at the right periphery of the sentence and this is well illustrated when clitic doubling phenomenon is analysed in Section 3.

2.2. 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 (see Primus 1993). This definition is much closer to our main goal – to distinguish grammatical realisations of focus in the sentence from these of its topic.

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.]
The kids broke the window.

(2) - Koj sčupi prozoreca?
Who broke the window?
- DECATA sčupiha prozoreca. / DECATA go sčupiha.[1]
The kids broke the window. / The kids broke it.

When a single sentence is uttered the indicator of focus is a single major accent (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. We don't generalise the role of sentence accentuation in the process of focusation because the stress marks out a syntactic domain in which focus is expressed. Prosodic emphasis on a constituent is not what matters as a means of marking the constituent as Focus, but rather "the existence of a separate Focus projection, different and lower than Topic" (Krapova 2001). We won't discuss here the sentence prosody or position of the intonation centre of the sentence because they are not relevant for the analysis of printed texts.

The focus of the sentence is determined by a set of grammatical factors, the prosody is just one of them. Lexical and syntactic means of focus marking also are important considerations. The focusing operator Y may be overt as with such particles as *only, even* in English, and the contrastive negation – *not John, but Peter*. We are going to argue that focusation is closely related with interrogativity and question particles in Bulgarian play certain role also on the level of Information structure. 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-marking accent is a syntactic constituent. Question particle *li* in Bulgarian which we analyse in Section 4 attracts the focal stress and assigns it to the constituent (phrase) formed with the element preceding the particle.

2.3. Location of Topic/Focus Features

Our view on focus and topic is closer to the concept that they are formal features carried by certain phrases within the sentence. We don't generalise the word-order criterion (Left-to-rights nodes corresponding to the communicative dynamism), but it should be taken into consideration that topics prefer the periphery of the sentence (left or right peripheries). Much resent work focuses on analysing the semantico-pragmatic properties of the constituents and evidences for their hierarchy. It has been noted in various studies that the order of constituents on the left periphery of the sentence is constrained. Rizzi (1997) proposes an analysis of CP domain according to which the latter has a finer structure and should be split into several projections. The complementiser system is "the interface between a propositional content (expressed by the IP) and the superordinate

[1] Focus of the sentence is noted by capital letters, Topic is marked by "T" after the word or phrase.

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 projections in the CP field:

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

The projection on the left boundary encodes features determine the sentence type or the illocutionary force of the sentence. On the right boundary is a projection, which carries the characteristics of the verb, namely the specifications concerning its "finiteness". Between these two boundaries Rizzi includes a system of projections 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 of CP domain was proposed, including a special node for Interrogative Force. There could be Topic phrases lower than Focus, because Topic is recursive.

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

When looking for a formal model of sentence structure we have to take account for the one proposed by Benincà (2001:62):

Disc(ourse)P > ForceP > TopP > FocP> FinP

For the analysis provided here we adopt the idea that finiteness is a core IP-related characteristic (Rizzi 1997:115), hence finite forms have to be situated in a position lower than the FocP and cannot move to a specifier position (i.e. Spec of FocP). Focalised elements are in the Spec of FocP. This position is banned for the verb. The second guiding idea is that there are topics higher than TopP (and higher than Force). This projection Benincà calls DiscourseP.

Intonation and word order are not reliable means to distinguish Topic and Focus. Rizzi pointed out some differences between topic and focus which could be used as formal tests. For our analyses very important are the following: there is a unique structural focus position, but a clause can contain as many topics as are the constituents. Only Topic can involve a resumptive clitic within the comment (ClLD); Wh operator in many questions is compatible with a topic in a fixed order (Top>wh). In Section 3 we analyse clitic doubling in Bulgarian as typical way to mark Topic of the sentence and provide more data about the mechanism of ClLD in Bulgarian. The connection between Bulgarian interrogativity markers (e.g. particle *li,* wh-words) and Topic-Focus configuration is discussed in Section 4.

To summarise this section, although various definitions of Topic and Focus could be proposed – as discourse-related, semantic or pragmatic categories etc. it is more important their formal features to be shown. Topic and Focus phrases correspond to special positions (projections) in the sentence. Elements marking Force interplay with both Topic and Focus. The phrases encoding sentence type and Information structure have hierarchic mutual ordering. Therefore, when analysing the Information structure of the sentence one can easily start from revealing this hierarchy.

3 Clitic doubling constructions in Bulgarian and Topicality

As it was mentioned in Section 1, Bulgarian differs from the other Slavic languages in many aspects. Important for our purposes is that Bulgarian has lost its declension system and possesses different forms for nominative, dative and accusative only in the paradigm of the pronouns (cf. Franks, King 2000:52). Although the native speakers do not consider these forms as case forms but rather as separate lexemes (Pašov 1978), here we use the term Nom, Acc, and Dat pronoun meaning

subject, direct and indirect object. For convenience we propose a model of the personal pronoun system in Bulgarian.

Person / Case	Nominative	Accusative		Dative	
	tonic pronoun	tonic pron.	clitic	tonic pron.	clitic
1p Sg	az	mene	me	na mene[2] / mene	mi
2p Sg	ti	tebe	te	na tebe / tebe	ti
3p Sg	toj (m), tja (f), to (n)	nego (m, n) neja (f)	go (m, n) ja (f)	na nego / nemu (m, n) na neja / nej (f)	mu (m, n) i (f)
1p Pl	nie	nas	ni	na nas / nam	ni
2p Pl	vie	vas	vi	na vas / vam	vi
3p Pl	te	tjah	gi	na tjah / tjam	im

The tonic pronouns *mene* and *tebe* have shortened variants *men* and *teb*. The Dat tonic pronouns have two forms – a synthetic (the form without preposition) and an analytic form (generated from the *na*-preposition + respective Acc tonic pronoun). Synthetic forms for first and second person Sg are still in use. The rest of the synthetic forms are very rare in contemporary Bulgarian and are usually substituted with the respective analytic forms.

Full pronominal forms, nouns, and PPs can be topics in Bulgarian. Clitic pronouns lack a proper accent, so they cannot function as topics. Further we investigate the formal markers for topic in Bulgarian and the position it occupies in the sentence. Recall that in the literature on Information structure most authors propose two positions for Topic: Left dislocation/detachment and Topicalisation (Lambrecht 1997), Left hanging-topic and Left dislocation (Primus 1993).

For the analysis of personal pronouns it is important to mention the investigation of word order and clitics in Bulgarian made by Avgustinova (1998). She examines the replication phenomena (i.e. the clitic doubling constructions) and its relation with the focus on the clause level. Her theoretical model of the Information structure is important for the description of this phenomenon in Bulgarian. But we cannot use it in the annotation of topic and focus in Bulgarian treebank, because treebank normally includes printed texts, where the information about the intonation contour is not added (and cannot be added, since in the written language other devices sign the emphasis in the sentence). Neither we can use the subdivision of topic into tale and link since in this case the information about the context and the semantics should be preliminarily added to the corpora.

For our purposes it is more appropriate to use the formal model of the left periphery proposed by L. Rizzi (1997, 2001) and P. Benincà (2001) for the description of the ClLD topic. Thus we could investigate not only the topic or the focus in Bulgarian, but also their interaction with Illocutionary force, wh-words, focus particles and complementisers and thus to propose a formal model for annotation of the topic and focus in treebank.

Following Benincà (2001), we consider that "[t]wo different constructions may produce a marked Topic or Theme: Left Dislocation (LD) and Hanging topic (HT)" (Benincà 2001:45). Here we focus on the LD because HT is more typical for Bulgarian colloquial speech and is not a part of the Left periphery, i.e. doesn't interact with the FocusP and the ForceP. In Italian "with LD the entire argument appears on the left – including any prepositions. A resumptive pronoun is obligatory with direct and partitive objects, optional in other cases (impossible if the type of argument has no appropriate clitic). If present, the clitic agrees with the Topic in gender, number and case" (Benincà

[2] Only PPs formed with *na* preposition are considered dative objects while the other prepositions form only indirect objects, thus the latter cannot be part of a clitic doubling construction.

2001:45). When the resumptive clitic is present we have a ClLD construction in Bulgarian. The clitic agrees with the dislocated element (NP, na-PP or a pronoun) in number, case, person (when the dislocated element is a personal pronoun), and in gender (only in 3pSg):

(3) Na nasT ni lipsvat precizni ikonomičeski organizacii.
To weAcc weDat lack3Pl precise economical organisations
We lack precise economical organisations.

In (3) the full Dat form of the personal pronoun for 1pPl (*na nas* - analytic form that replaces the synthetic *nam*) is doubled by the respective Dat clitic form for 1pPl (*ni*). The construction is ClLD and marks the topic of the sentence (i.e. *na nas*). The clitic agrees with the full form in number, case, and person. When there is no resumptive clitic, we have only a LD topic:

(4) TovaT OTBRANA li e?
ThisNom defence Q is?
Is this a defence?

The topic in (4) is *tova* (this), because it precedes the focus *otbrana* (defence) (marked explicitly by the focus particle *li*). The topic *tova* cannot be doubled, since there is no resumptive clitic for Nom in Bulgarian. Therefore the topic is a LD, not a ClLD. The neutral word order in declarative sentence is:

(5) Tova e otbrana.
This is a defence.

When the focus should be emphasised, the topic can occupy the final position. This configuration could be called a *low topic*:

(6) OTBRANA li e tovaT?
Defence Q is this?
Is this a defence?

The possibility the topic to be lower than the focus reflects grammar differences relevant for the investigation of some language-specific properties of Bulgarian. In other Slavic languages like Czech the contextually bound nodes normally precede contextually non-bound nodes (cf. Buranová 2000), i.e. the topic in (6) should precede the focus.

The condition for a PP to be doubled in Bulgarian is the presence of a *na* preposition (7)

(7) a. Na meneT mi e mâčno za nego. b. *Na mene e mâčno za nego.
to IAcc IDat is sadly for heAcc to IAcc is sadly for heAcc
I feel sorry for him.

In (7) the presence of the na-PP requires obligatory clitic doubling. With other verbs (e.g. *davam* 'to give' *Na Ivan dadoh knigata* 'I gave the book to Ivan') the doubling of the fronted na-PP is not obligatory. When the PP is doubled it signs the topic of the sentence. When no doubling appears the fronted object receives contrastive stress. Hence the clitic doubling constructions are in close connection with the description of the predicates. Such kind of information could be included in the lexicon in treebanks.

If the preposition introducing the indirect object PP is other than *na*, it is not coreferent with the clitic, even if they coincide in case, gender, number and person. In (8) there are two indirect objects – a full pronoun *za nego* (for him) and a Dat clitic *mi* 1pSg (to me) but these two forms are not coreferent and there is no ClLD construction.

(8) Za nego mi e mâčno.
for heAcc IDat is sadly.
I feel sorry for him.

Even if the two pronouns coincide in their grammatical characteristics, e.g. (9), there is no ClLD.

(9) Za nego mu e mâčno.
for heAcc heDat is sadly.
He feels sorry for him.

When the indirect object is not a PP, but a full synthetic pronominal form (*mene* 1pSg or *tebe* 2pSg), which coincides with the respective Acc form, the doubling clitic is always coreferent with the pronoun and is the only disambiguating device in the process of annotation of a corpus. In contemporary Bulgarian the clitic doubling is obligatory when a synthetic Dat pronoun is used (cf. Nicolova 1986:34).

(10) MeneT me udari Ivan.
IAcc IAcc hit3Sg Ivan
Ivan hit me.

(11) MeneT mi e mâčno.
IDat IDat is sadly.
I am sad.

Why we claim that Bulgarian possesses the ClLD construction and that in this respect differs from other Slavic languages? Bulgarian and Macedonian are the only Slavic languages that have developed the clitic doubling construction. Both languages also differ from the Romance Clitic Doubling "in that no preposition is required to case-license the associate as it is in the more familiar Romance type of clitic doubling. According to Rudin, clitic doubling occurs in Bulgarian when the associate noun phrase is both topicalised and specific" (Franks, Kings 2000:53), i.e. in Bulgarian a NP as well as a PP can be doubled.

Although specificity is necessary and sufficient to licensing clitic doubling in Macedonian, Bulgarian requires an additional factor: topicality. "When a specific NP is topicalised and as a result appears in initial position, it must be doubled (97) [cited here as (12)]. So in [12] both the indirect object *na Ivan* 'to Ivan' and the direct object *knigata* 'the book' are topicalised and specific and must be doubled (Franks, King 2000:253).

(12) Na IvanT knigataT az mu ja dadoh.
to Ivan bookDef INom heDAT sheAcc gave1Sg"
[I gave the book to Ivan]

Apparently it is possible two ClLD to appear simultaneously. The na-PP *na Ivan* is coreferent with the Dat, masculine clitic *mu*, and the NP *knigata* – with the Acc, feminine clitic *ja*. The neutral word order (SVO) of this sentence is

(13) Az dadoh knigata na Ivan.
INom gave1Sg bookDef to Ivan
I gave the book to Ivan

The topicalisation of the direct or the indirect object requires its doubling with the respective (i.e. accusative or dative) pronominal clitic.

(14) Na Ivan az mu dadoh knigata.
to Ivan INom heDat gave1Sg bookDef

(15) Knigata (az) ja dadoh na Ivan.
bookDef INom sheDat gave1Sg to Ivan

When the preposed direct/indirect object is not doubled, it receives contrastive stress: *Na Ivan dadoh knigata* – na Ivan is opposed to another entity in the close context, e.g. *to Ivan, not to Pavel, I gave the book.*

According to Sv. Ivančev when the object of the sentence (direct or indirect) is expressed with clitic doubling construction, it is the exposition (i.e. the topic) of the sentence. The exposition could be

initial (in the beginning of the sentence, the full form preceding the clitic) or final (at the end of the sentence, the full form is postponed in respect to the clitic). He uses the following examples:

(16) NegoT go vižda IVAN.
heAcc heAcc sees3Sg Ivan
Ivan sees him.

(17) IVAN go vižda negoT.
Ivan heAcc sees3Sg heAcc
Ivan sees him. (Ivančev 1978:166).

So the doubling construction is sufficient for the identification of the topic. Often occurs that instead of pronoun the direct object is a N/NP:

(18) MarijaT jâ vidja Ivan.
Maria sheAcc saw3Sg Ivan
Ivan saw Maria.

In (18) it is clear that the doubled element, and therefore the topic, is *Maria*, because it agrees with the Acc clitic in number and gender (Sg, Fem) while the subject of the sentence is a masculine name (*Ivan*). The agreement is an important marker to distinguish the Subject and Object in Bulgarian, in case that both are expressed by a NP. The subject in Bulgarian (if it is expressed overtly) agrees with the verb, and the complement agrees with the respective clitic (if it is expressed overtly). There are also sentences where the subject and the object are both masculine and both of them agree with the clitic:

(19) Pavel go vidja Ivan.
Pavel heAcc saw3Sg Ivan
Ivan saw Pavel / Pavel saw Ivan

So we must elaborate other rules to annotate the topic into Bulgarian corpora. For Bulgarian speakers it is clear from the context which of these nouns is doubled (in (19) normally it is *Pavel*) and consequently the topic, but there are other devices applicable for annotation of the Information structure in treebank, i.e. the relation of the topic with other elements of the left periphery.

Under the formal approach chosen here we propose the following model of the clitic doubling construction in Bulgarian:

1. NP (direct object) – resumptive Acc clitic in the sentence;
2. Acc tonic pronoun (direct object) – resumptive Acc clitic in the sentence;
3. na-PP (*na* + NP) (indirect object) - resumptive Dat clitic in the sentence;
4. Dat tonic pronoun or a na-PP (na+Acc tonic pronoun) (indirect object) - resumptive Dat clitic in the sentence;
5. Acc clitic in the sentence - NP (direct object);
6. Acc clitic in the sentence - Acc tonic pronoun (direct object);
7. Dat clitic in the sentence - na-PP (*na* + NP) (indirect object);
8. Dat clitic in the sentence - Dat tonic pronoun or a na-PP (na+Acc tonic pronoun) (indirect object).

The direct or the indirect objects are expressed by two coreferent words in the sentence. This phenomenon was mentioned as one of the formal differences between topic and focus. Due to these differences we propose the Information structure analysis as a tool to explain such constructions. The 1-4 constructions mark overtly the topic of the sentence and present the so-called Clitic Left Dislocation (ClLD). The direct/indirect object occupies the topic node in the Left Periphery of the sentence.

In Bulgarian it is possible the doubled XP to be situated to the right of the sentence, thus postponed in respect to the clitic (5-8 constructions):

(20) Dosta često go risuvaha nego
very often he/itAcc painted3Pl he/itAcc
They often painted him/it.

(21) Toj izpâlni tova, koeto mi obešta na men
he fulfiled3Sg this what IDat promised to IAcc
He did what he had promised me.

The status of these constructions in the Information structure of the sentence is still not unambiguously defined, so we limit our investigation to the ClLD (i.e. constructions 1-4). The rule for annotating clitic doubling constructions at the sentence level is:

> When two pronouns in the sentence are assigned the same case (two Acc or two Dat pronouns) and one of them is a clitic, we have a clitic doubling construction. When the tonic pronoun precedes the clitic, we have a ClLD construction and the tonic pronoun occupies the topic position in the left periphery (constructions 2 and 4).

(22) Bedata e, če nasT ne ni upravljavat patrioti.
troubleDef is that weAcc not weAcc govern3Pl patriots
The trouble is that we are not governed by patriots.

(23) Na menT vsičko mi haresva.
to IAcc everything IDat pleases3Sg
Everything pleases me.

Imoprtant for the identification of the ClLD is the agreement between two pronominal forms in person, gender, case and number. The agreement is also important for the identification of clitic doubling with an NP or a na-PP instead of the tonic pronoun:

(24) ZâbiteT vinagi si gi mijâ sled večerja.
teeth always Refl theyAcc wash1Sg after dinner
I always brush my teeth after dinner.

In this example the verb is in 1pSg, the subject is not overt. The NP *zâbite* (my teeth) is not the subject, because it cannot function as a subject in 1pSg. If it is not a subject it could only be a direct object of the verb *mija* (to brush). On the other side, it agrees with the clitic *gi* in number. The NP precedes the resumptive clitic, so we have a ClLD construction and the NP is a topic in the sentence. The rule is similar when we have a PP as a part of the ClLD construction:

(25) Na drugite decaT im davat igrački.
to other kids theyDat give3Pl toys.
Other kids get toys.

In (25) we have a PP with *na*-preposition and the Dat clitic agrees with it in number only (because in Bulgarian there is only one clitic form for 3pPl). We can also have a focus particle simultaneously with the ClLD:

(26) IvanT MARIJA li go pokani?
Ivan Maria Q heAcc invited3Sg
Did Maria invite Ivan?

The verb in this sentence is in 3pSg. The subject is *Maria*. The NP *Ivan* is a direct object and agrees with the Acc clitic *go* in gender and number. In Bulgarian the pause after the left dislocation is not obligatory, that's why the intonation (or the punctuation reflecting the intonation) is not a disambiguating feature, at least when we investigate printed texts.

4 Bulgarian particle *li* as Focus marker

Bulgarian particle *li* can be used as a force marker in direct yes-no questions and indirect (embedded) questions. It could be also part of some adverbial phrases. Due to the main task – to show the intrinsic features of *li* we will concentrate on analysing direct yes-no questions. Other cases will be just mentioned in order to prove the conclusions from the investigation of yes-no questions. The lexeme *li* is used in several other Slavic languages. In Russian *li* is optional in main clauses, but obligatory in subordinated clauses. In Czech *li* is not a question marker. It occurs only in conditional sentences, contrasted to another sentence. Slovak and Ukrainian lack such particle (see Franks, King 2000). Bulgarian and Macedonian differ significantly form other Slavic languages. In both languages *li* occur regularly in matrix and embedded questions.

4.1. *Li* in yes-no questions

Bulgarian has a large number of lexical items that indicate yes-no questions: *li, dali, nali, nima, migar, zer, da, da ne bi da, a, e* (see GSBKE 1983:62), but *li* is the one that was investigated much more than the other. It was classified within different theoretical frameworks, but for the analysis here we adopted the formal approach based on concepts of generative linguistics. Rivero (1993), Rudin (1997) argue that the question marker *li* is a complementiser[3]. Since *li* is a clitic, more precisely an enclitic to constituents in SpecCP or C (cf. King 1997, Franks, King 2000:350), it needs a host to form a prosodic word together with it. For our research it is more important to point out that *li* assigns focus features, as well as marks the sentence type. In yes-no questions with *li* the focus position is stable because focus is morphologically marked by the "bound morpheme" *li* (see Motapanyane 1997). In all configurations the [focus] features associate with C and this follows from the lexical properties of the clitic complementiser. When discussing the distribution of *li* we follow the "principle" that *li* is specified for both interrogative and focus features and attracts these features to C (Force/InterrogativityP).

The clitic *li* together with a phonological word forms a clitic phrase through the phonological operation of cliticisation. An XP can move to SpecCP where it hosts the clitic *li*. If a maximal projection moves to SpecCP, it hosts the clitic *li* and it is a focus of the question. The phrase hosting *li* could be a non-complex maximal projection as in (27) and (28)

(27) KNIGATA li donese?
bookDef Q brought2Sg
Was it the book that you brought?

(28) Včera MARIJA li se obadi?
yesterday Maria Q Refl called
Was Maria the person who called yesterday?

or complex maximal projections as in (29). The final position is taken by the verb and that corresponds with Rizzi's analysis of CP domain.

(29) [Novijâ mu ROMAN] li četeš?
newDef his novel Q read2Sg
Is it his new novel that you are reading?

We won't analyse here the possibilities APs to move out of the NPs in Bulgarian as in (30)

(30) NOVATA li kola prodade (ili starata)?
new+the Q car sold3Sg or old+the
Did (s)he sell the NEW car (or the old one)? (from Rudin at al. 1997)

Such examples could be accepted only when the element in front of *li* gets a contrastive reading and get the status of Contrastive focus

[3] We will refrain from discussing here the complementiser status of *li*. At this stage it is more convenient to adopt the theoretical background, proposed within the generativist framework.

(31) TVOJÂT/ZELENIJÂT li telefon zvâni (ili mojât/belijât)?
yourDef/greenDef Q phone rings or mineDef/whiteDef
Is it your/the green phone ringing (or mine/the white one)?

Examples as (31) are restricted to cases of contrast which we are not going to discuss. For the analysis proposed here we will generalise the case when *li* cliticises onto the fronted NP regardless how heavy it is. The focal stress is usually taken by the last element from the NP – its head as in (32):

(32) [Tvojât kolega IVAN] li šte donese knigata?
yourDef colleague IVAN Q will bring bookDef
Is your colleague Ivan the person who will bring the book?

It is possible not only the head, but also other elements within NP to be stressed. In examples like (33) it is the adjective that is focused. From the example in (34-35) we see that the PPs within the NPs can get the focal stress. Last two examples follow the canonical pattern of focusing in *li*-questions: the element immediately preceding *li* is the focus of the question.

(33) [ZELENATA roklja] li kupi?
greenDef dress Q bought
Did you buy the green dress/the green one?

(34) [Novata statija na IVANOV] li četeš?
newDef article of Ivanov Q read2Sg
Is it the Ivanov's new article what you are reading?

(35) [Karta na SOFIJA] li iskate?
mapDef of Sofia Q want2Pl
Is it the map of Sofia that you want?

Li marks the sentence type (force) as well as focusation. When changing the phrase which hosts *li* the focused constituent changes, but not the force – compare (36) and (37)

(36) Včera IVAN li donese knigata?
yesterday Ivan Q brought bookDef
Was it Ivan who brought the book yesterday?

(37) Včera KNIGATA li donese Ivan?
yesterday bookDef Q brought Ivan
Was it the book that Ivan brought yesterday?

If there are no other candidates to host *li*, then the tensed lexical verb moves to C and incorporates with the clitic. This is the way to form neutral non-focused yes-no questions. The verb to which *li* is attached may be preceded by one or more TopicP, as the Rizzi's scheme predicts (cf. Rizzi 2001, cited above in 2.3.)

(38) VčeraT IvanT donese li knigata?
yesterday Ivan brought3Sg Q bookDef
Did Ivan bring the book yesterday?

In Bulgarian *li* can also cliticise onto non-finite forms - past participles as in (39) and passive participles as shown in (40) The auxiliary *sâm* (*to* be) takes the specification of finiteness: person, mood, subject agreement licensing nominative case, tense. Finiteness phrase could be preceded by focus, but only topic(s) could be after it. As shown in Section 2.3., the finite forms cannot take the focus position. The non-finite forms have the possibility to be focused because they don't form a prosodic constituent with the verb, but with *li*.

(39) MarijaT PROČELA li e knigata?
Maria readPastPartF Q is bookDef
Has Maria read the book?

(40) VečerjataT PRIGOTVENA li e?
dinnerDef donePassivePartF Q is
Is the dinner ready?

Li is considered to be one of the clausal clitics in Bulgarian. This term is used to refer the clitics that locate their host within the clause, not within a phrase (see Tomić 1997), such as negative particle *ne*, future particle *šte*, as well as present forms of auxiliary *sâm* (to be) and pronominal sets of clitics. When a clitic cluster is formed *li* is in front of pronominal clitics

(41) *Mu li go dade/ *Maria mu li go dade
heDat Q itAcc gave/ Maria heDat Q itAcc gave

(42) *Si li izmi zâbite/ *Petâr si li izmi zâbite
refl Q wash teethDef/ Petar Refl Q wash teethDef

Li still cannot take the first (initial) position within the sentence. It has to be hosted by a stressed word which after the operation of cliticisation gets the focal stress

(43) MARIJA li mu go dade?
Maria Q heDat itAcc gave
Is Maria the one who gave it to him?

With respect to prosody forms of auxiliary *sâm* (to be) as parts of compound tenses in Bulgarian differ significantly. The forms for present tense are clitics, as shown in (44), but the past tense forms of auxiliary are not clitics (45), so they can host *li*

(44) *Si li pročela knigata
are2Sg Q readPPart bookDef

(45) Marija beše li pročela knigata?
Maria had3Sg Q readPPart bookDef
Had Maria read the book?

Under some circumstances the normally clitic forms of auxiliary *sâm* (to be) can also be treated as non-clitics, taking word stress and appearing in the non-clitic positions. In the following examples the form of auxiliary *sâm* (to be) takes the first position in the sentence, hence it acts as a stressed word and could host *li*:

(46) E li tova rešenie naj-dobroto?
is Q this decision the best
Is this decision the best one?

(47) Zatova pitam onija, koito sa porâčali i napisali tazi objava za glavata mi: E li tova smârtna prisâda za mene....
That's why I am asking those who ordered and wrote this notice for my head: is this a death sentence for me…

Bulgarian has a special auxiliary verb *bih* for conditional forms. They are constructed from the auxiliary verb *bih*, declined for person and number, plus the past active participle, declined for gender and number. The forms of this auxiliary are not clitics, hence they could host *li*. The forms of *bih* are specified for Finiteness, hence they cannot be focused.

(48) Bi li mi podal solta?
could2Sg Q IDat givenPPsrt saltDef
Could you pass me the salt, please?

Although the negation particle is one of the clausal clitics in Bulgarian, it has to be analysed separately from other clitics. Bulgarian *ne*, the Negation head, has a special property of stressing the following element, causing *li* to invert prosodically past the prosodic word consisting of NE plus the following, stressed element (Rudin et al 1999). In the following examples the first clitic – personal

pronoun or reflexive, is stressed by the negation particle. *Li* follows the prosodic word consisting of [NE + stressed clitic].

(49) NE MU li se obadihte?
Not heDat Q refl called2Pl
Haven't you already call him?

(50) NE SE li obadihte na Petâr?
Not Refl Q called2Pl to Petar
Haven't you already call Petar?

The element in front of [NE + stressed clitic] constituent is in the canonical position for Topics – beginning of the sentence, preceding the FocusP

(51) Marija NE MU li go e dala?
Maria not heDat Q itAcc is given
Hasn't Maria already given it to him?

From the examples analysed here we can conclude that the position of *li* in Bulgarian depends on two factors: phonological (the stress) and the structure of the phrase which hosts *li*. Two different prosodic phenomena are relevant for the placement of *li* in Bulgarian: stressing of clitics after the negation particle and inversion of the initial clitic with the following finite or non-finite verb form. The status of auxiliaries – clitics or non-clitics – is an additional factor that plays certain role in the formation of yes/no questions with *li*.

To summarise this section, in Bulgarian yes-no questions the clitic complementiser *li* may be specified for both [interrogative] and [focus] features. The first one is intrinsic to *li* and, thus, always presented when this particle is used to form yes-no questions. *Li* cannot be used as a marker for [focus] feature only. In yes-no questions with *li* [focus] feature always forms a set with [interrogative] feature. The constituent to which *li* is attached is focused. The only exception is observed when the verb hosts *li.* The verb is associated with Finiteness phrase and cannot move to a focus position, thus a neutral non-focused question is formed.

4.2. More data about distribution of *li.* Ambiguities

Our analysis of Bulgarian particle *li* led us to the conclusion that it is not possible to generalise [interrogativity] as its unique or intrinsic feature. The existence of homonyms in the language or the polifunctionality account for the possibility of using *li* in many constructions, phrases or clauses which are not marked as interrogative (neither direct, nor indirect questions). It is important to describe all contexts where *li* 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.

I *Li* as a derivation particle

It is possible the particle *li* to loose its interrogative characteristics and to function only as a derivation particle in some expressions. These phrases are described in the literature as adverbs. They form a close set of words and must be included in the lexicon as multiwords. Here we list those expressions: *kato če li (ne), edva li (ne), kamo li da, vidite li.*

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. Although these phrases contain the particle *li* they don't carry the [interrogativity] features and don't take part in the process of force marking.

II *Li* in wh-questions

Wh-words (*k*-words in Bulgarian) carry two types of features: [interrogativity] and [focus]. They undergo movement to a focus position (SpecCP). The enclitic *li* is a morphological counterpart of wh-phrases – as it was shown in the previous section *li* has both [interrogative] and [focus] features

in direct yes-no questions. Wh-questions allow for the insertion of *li* as a morphological marker for the same feature. Wh-words are always focused, hence the insertion of *li* in order to mark the focus lexically is redundant. The option for wh-questions with particle *li* occurs only for special effects. Various combinations are possible:

wh-word + li + N/NP -- (clitic) -- V

Wh-words could form different groups: with pronouns of identity as in (52); of quality as in (53); or of possession as in (54)

(52) KOJA li banka e otvorena po tova vreme?
wh Q bank is open at this time
I wonder if there is a bank open at this time?

(53) KAKVI li vetrove sa go doveli tuk?
wh Q winds are heAcc broughtPPart here
What/(Whatever) winds may have brought him here?

(54) ČII li interesi zaštitava?
wh Q interests defend3Sg
I wonder whose interests (s)he defends.

Wh-word and the noun agree in gender and number. *Li* is inserted between them in order to mark lexically the focus of the sentence. The constituents formed by wh-words and nouns usually are at the beginning of the sentence – the normal position for wh-phrases. In the same context prepositional phrases also could be used.

(55) Na KOJ li prepodavatel možeš da razčitaš?
to wh Q teacher can da trust
Which teacher can you (really) trust?

(56) Za KOJ li pât Balkanite otnovo se okazvat arena na sblâsâk?
for wh Q time BalkansDef again Refl turn out arena of crash
For the umpteenth time the Balkans have again become a theatre of war?

Only the adverb for quantity *kolko* (57) could follow the same pattern as wh-pronouns. A noun or a noun phrase should be included in the "large" wh-phrase. Other pronominal adverbs can be combined with *li* but not with nouns/noun phrases. They follow a different scheme and will be described separately.

(57) KOLKO li dni e pâtuval?
wh Q days is travelledPPart
I wonder how many days (s)he has travelled.

To summarise the last examples, *li* could be inserted in wh-phrase formed by a wh-pronoun or an adverb and a noun/noun phrase. Wh-words carry the information about illocutionary force, as well as the focus features. The wh-words specify the sentences as direct wh-questions; the particle *li* doesn't take part in the force marking. Its role is to mark lexically the narrow focus of the questions. For *li* the [interrogative] feature is redundant, but [focus] feature is "activated". This assumption could be proved by the fact that other constituent than wh-phrase can be focused after the attachment of *li.*

wh-word + N/NP + li --(clitics) – V

(58) Na koj PREPODAVATEL li možeš da razčitaš?
to wh teacher Q can da trust
Which teacher can you (really) trust?

The example in (58) shows that *li* can be attached to the whole wh-phrase, not inserted in the phrase after the wh-word. The sentence type is specified by the wh-word (direct wh-question), but not the

whole wh-phrase is specified for [focus]. It is the element (the noun) preceding *li* which gets the focal stress. In such cases since the force is marked by wh-words the use of *li* in order to mark the [interrogativity] is redundant. The [focus] features of *li* are important to specify that an element different than wh-word (or phrase) is focus of the question. Due to the specification of *li* to mark narrow focus it cannot take scope over the whole wh-phrase, but only over the word to which it is attached.

The examples like (58) are acceptable as echo-questions. We won't present any special rules for *li* in echo-questions because they have different structure than direct yes-no questions with *li* which we analysed in Section 4. "In echo-questions *li* usually is at the end of the repeated question, but a context is needed to identify that something was repeated (echoed)" (Hauge 1999:216). As was shown from the example in (57) *li* could be after the wh-phrase. For our task it is important that in some cases the particle *li* can act as a focus particle and its [interrogativity] features are redundant. The model for focusation is the same as in direct yes-no questions. *Li* is specified to mark narrow focus; this feature associates with the constituent preceding *li.*

wh-word + li – (clitics) – V

The next model for using *li* in wh-questions differs from the first two because the wh-phrase does not include a nominal element. *Li* is attached to the wh-word in order to mark overtly the focus of the question. The particle doesn't take part in the force marking. All wh-pronouns can form a constituent with *li:* pronouns of identity as in (59); of quality as in (60); pronouns of possession as in (61)

(59) KOJ li e tozi čovek?
wh Q is this man
Who is this man?

(60) KAKVO li e tova?
wh Q is this
What is this (I wonder)?

(61) ČIJA li e tazi kniga?
wh Q is thisF book
Whose is this book (I wonder)?

Wh-pronominal adverbs from all semantic groups can be used in this type of questions: adverbs of time as in (62); adverbs of place like in (63); adverbs of manner/and of quantity as in (64)

(62) KOGA li šte se obadi?
wh Q will Refl call3Sg
When will (s)he call (I wonder)?

(63) KÂDE li sâm go složila?
wh Q am itAcc putPPart
Where might have I put it?

(64) KAK/KOLKO li se e promenil?
wh Q Refl is changedPPart
How much has he changed (I would like to know)?

In the last examples the wh-words are focused and *li* is the lexical marker for that. Again the [focus] features of *li* are the reason for its insertion in the wh-questions.

To sum up, when the particle *li* is inserted in wh-questions it could be attached directly to the wh-word or to the whole wh-phrase (formed by the wh-word and a noun/noun phrase). The [interrogative] features of *li* are optional because the sentence type is already specified by the wh-word. *Li* is inserted as lexical specifier for narrow focus. The word immediately preceding *li* is specified for focus.

wh-word + li + ne – (N/NP) – (ne) V

The wh-phrase to which *li* is attached could be followed by the negation particle *ne*. In (65) a noun is included in the wh-phrase, the negation particle is in front of it. The examples like (66) and (67) show that a constituent can be formed without any noun or noun phrase; in (66) a interrogative pronoun of identity is the head of the phrase, in (67) this position is taken by an adverb of place.

(65) KAKVI li ne metodi ne se smeniha prez poslednite 10 godini.
wh Q not methods not Refl changed3Pl during lastDef 10 years
What methods were not changed during the last 10 years?

(66) KOGO li ne pitah.
wh Q not asked
Who[m] did I not ask?

(67) Târsih go KÂDE li ne.
looked1Sg heAcc wh Q not
I have looked for him everywhere.

The last examples show that *wh-li-ne*-constructions have to be analysed as grammaticalised phrases. They could be added to the group of Bulgarian all-quantifying pronouns because they contain the "semantic element of 'allness' (Hauge 1999:61). Such constructions refer to every element from the domain defined by the wh-word. They must be included in the lexicon as multiwords with a note that *li* is a derivation particle, but still keeps the possibility to attract the focal stress. What we want to point out is that such constructions don't carry interrogative features. After the insertion of negation operator *ne* wh-word and question particle *li* are not able to realise their interrogativity. Despite the loss of force marking features wh-word and *li* keep their focus features. The wh-word gets the focal stress; the particle *li* still acts as a lexical marker for focus.

The statement that *li* cannot occur as a marker for focus alone, because the interrogativity feature is intrinsic to *li* whereas the focus feature is optional now seems inconclusive. Examples like these in (65-67) led us to the assumption that focus features of *li* are not obligatory related with its interrogativity. We can prove it by more data for distribution of *li* in complex sentences.

III *Li* in coordinate constructions or clauses

Two (or more) adjectives, nouns/noun phrases, or prepositional phrases could be conjoined by adding *li* after each of them. The following patterns could be observed:

A1+ li, A2 + li… or *N1+ li, N2+ li…*

(68) Iskam njakakva roklja – bjala li, zelena li, vse edno, samo da e kâsa.
I want any (kind of) dress – white or green, it doesn't matter, the only thing I want is to be short.

(69) Pitaj njakogo – prepodavatel li, sekretarka li, vse šte ti kažat nešto.
Ask somebody – a teacher or a secretary; maybe somebody will tell you something.

These constructions are closely related to the disjunctive yes-no questions (with *li…ili*). In (68-69) *li* is repeated instead of using *ili* (*or*).

Furthermore, *li* also could be repeated in both parts of a coordinate clause. Additional implication of temporal links (consecutive or simultaneous actions) between the sentences could be found.

N/NP/PP + li + ne/šte –V, N/NP/PP + li + ne/šte –V …

(70) Po lekari li ne hodih, bilki li ne pih – ništo ne pomaga.
I did everything one can do – several doctors examined me, I took some herbs, but nothing helped me.

(71) Dieta li šte pazja, gimnastika li šte pravja…, kakvoto i da e, no trjabva da otslabna.
I have to do my best to get slimmer – to diet, exercise, whatever…

Next example shows that the same verb could be repeated. Again *li* acts as a coordinative conjunction. These constructions are used for special effects to express a long lasting action.

V—li—V

(72) Ne moga da ja spra, pâlna e s energija, po cjal den bjaga li, bjaga.
I can't stop her, she is full of energy –she runs and runs all day long.

To summarise the use of *li* in coordination phrases or sentences, the particle acts as a coordination conjunction. No interrogative features are presented. Although *li* does not take part in the force marking it is still able to attract the sentence stress. We assume that *li* keeps its connection with focus marking. The word to which *li* is attached is emphasised by prosodic means.

IV *Li* in conditional clauses

By origin *li* is a question particle, but also could be used in conditional clauses as synonym of *ako* (*if*). As part of embedded clause *li* doesn't take part in force marking of the sentence. Such clauses express a complex of typical, habitual or repeated actions or well-known facts. What is specific for them is the strict selection of tense and aspect forms in both matrix and subordinated clauses. In this paper we won't analyse in details the sequence of tenses and aspect usage, but will concentrate on that whether *li* keeps its focus feature or not.

Before discussing the structure of conditional clauses with *li,* we would like to distinguish them from those with complex conjunction *ako li* (*pâk*) 'if not'. The latter presents another alternative that is not mentioned in the given context (see Nicolova 1998).

(73) Ako imam vreme, šte dojda. Ako li pâk sâm zaet, šte ti se obadja po telefona. (from Nicolova 1998:146)
If I have time I will come. If I am busy, I will call you.

The complex *ako li* (*pâk*) have to be described as multiword. *Li* is a derivation particle added to the complementiser *ako*. The latter marks the sentence type; other lexical or prosodic means could be used to mark focus of the sentence.

Two types of conditional *li*-clauses were observed. In order to show the differences between them, the lexical selection of the verb and the constituent to which *li* is attached have to be examined. The first type is specified by the usage of *sâm* (*to be*) and a noun as a host for *li* as shown in (74)

N + li+ form of sâm (*to be*)

(74) Borec li si, bori se dokraj! / ne trjabva da se predavaš!
If you are a fighter, fight like one! /you shouldn't give up!

The second one requires a verb (perfective imperfect or perfective present forms mainly, but also some imperfectives) to host *li.*

(75) Zapeeše li, vsički mlâkvaha.
Whenever she struck up a tune, everyone would fall silent.

(76) Staneše li rano, šteše da ima vreme za vsičko.
If (s)he got up early, (s)he would have time for everything.

(77) Stavaš li rano, imaš vreme za vsičko.
If you get up early, you will have time for everything.

What is common for all conditional clauses with *li* is that the particle still tends to attract the sentence stress. The emphasis is very explicit when *li* is attached to a noun. When the verb hosts *li* a neutral non-focused sentences is formed. Focus could be specified by other means, e.g. other focus particles, quantification, etc. This follows the general rule shown in 4.1. for direct yes-no questions for focusation.

V *Li* in embedded (indirect) questions

Indirect questions with *li* are identical with direct questions in their word order. The syntactic properties of *li* in direct yes-no questions are relevant to the [focus] and [interrogativity] features. *Li* may be specified for both [focus] and [interrogativity] and attracts these features to its host. If *li* is attached to the verb, the [focus] feature is not checked, hence *li* is only an interrogativity marker. As part of embedded question *li* cannot mark the illocutionary force of the complex sentence, but still acts as a focus marker. The constituent to which *li* is attached is focus of the indirect question

(78) Popitah Ivan TAZI VEČER li ště pâtuva.
asked1Sg Ivan this night Q will travel
I asked Ivan if he will travel tonight.

(79) Popitah Ivan S MARIJA li ště pâtuva.
asked1Sg Ivan with Maria Q will travel
I asked Ivan if he will travel with Maria.

When specifying the features of *li* as part of the morphological annotation of the corpus an additional mechanism is required to account for the differences between direct and indirect questions with *li.* We propose the following distinction to be included: the feature [interrogative] has two values [direct] and [indirect]. *Li* in direct yes-no question is specified as [interrogative > direct], [focus]. *Li* in indirect questions is marked as [interrogative > indirect], [focus]. This particle cannot be used as a sign for beginning or end of clauses. As an enclitic *li* needs a host, but the focused constituent (*li*-host) could be preceded by one or more topics. Our analysis of indirect questions with *li* could be used for future work on annotating the information about clause type or word order, but not on linearisation.

The examples from indirect questions prove our observations that [focus] features of *li* could be presented independently of its [interrogative] feature. The analysis of *li* as interrogative complementiser cannot explain its use in coordinate constructions or clauses and in conditional clauses (Section 4 III-IV). We suggest Bošković's idea for focal inflection analysis of *li* (Bošković 2001:241) to be taken when specifying its properties. For the corpora annotation we propose *li* to be considered a focal particle added to a focused element. In all context the particle has the feature [information] > [focus]. An additional feature [interrogative] has to be used to distinguish *li* in questions. We consider that three types of *li* are observed in Bulgarian, which are homonymous lexemes. The following model covers all possible contexts for *li*.

Homonym	Context	Features
Li1	wh-questions, coordinate constructions and conditional clauses	[information]>[focus]
Li2	direct yes/no questions	[information]>[focus] [interrogative]>[direct]
Li3	in indirect questions	[information]>[focus] [interrogative]>[indirect]

5 Conclusions

In this paper we argued for the necessity of insertion of two complementary feature to some words, i.e. [information], with two values [topic] and [focus], and [force]>[interrogative], with two values [direct] and [indirect]. Our proposal is consistent with one of the major objectives of the treebank corpora building - to add as much linguistic information as possible, to make the corpus really efficient for feature users: linguists and humanity researchers. The advantages are the following: certain formal properties of sentences cannot be fully understood without looking at the linguistic and extralinguistic contexts in which the sentences having these properties are embedded. Topic

Focus Articulation (TFA) is expressed by grammatical means, it is part of the sentence grammar and a determining factor in the formal structure of the sentence so it is relevant for the interpretation of the sentence. Such analysis can help in elaborating the set of features that should be used when describing exhaustively the language-specific phenomena at the syntactic and morphological layer. The word order or the order of the phrases as one of basic syntactic phenomena has to be analysed as motivated by the TFA. The principles and methodology of the formal linguistics can help in solving problems with linearisation, marking force, valence, etc. The study of the additional elements like particles, which play significant role on the level of syntax, is important for the correct annotation of the corpora data. The proposed analysis could be extended to other formal devices marking the Information structure in Bulgarian sentence, i.e. the interrogative particles *dali, nali, nima*, the so called focalisers (*samo* 'only', *dori* 'even', *edva* 'barely', etc.) and the hanging topic construction. We use ideas from different theories in order to include different information, not to be one-theory dependent. Thus we allow our hypothesis to be tested under different theories and to be applicable in the analysis of TFA in other languages.

6 References

Avgustinova 1998 - T. Avgustinova. *Word Order and Clitics in Bulgarian* [Saarbrücken Dissertations in Computational Linguistics and Language Technology. Volume 5]. Saarbrücken.

Benincà 2001 – P. Benincà. The position of Topic and Focus in the Left Periphery. In: Cinque, G. and G. Salve (eds.). *Current Studies in Italian Syntax.* Essays offered to Lorenzo Renzi. Elsevier, 39-64.

Böhmová et al. 2000 - A. Böhmová, J. Hajič, E. Hajičová, B. Hladká. The Prague Dependency Treebank: Three-Level Annotation Scenario. In *Treebanks: Building and Using Syntactically Annotated Corpora, Anne Abeille. Kluwer Academic Publishers (ed.).*

Bošković 2001 – Ž. Bošković. *On the Nature of the Syntax – Phonology Interface* (cliticization and related phenomena). Elsevier, 2001.

Buranová 2000 - E. Buranová, E. Hajičová, P. Sgall. Tagging of Very Large Corpora: Topic-Focus Articulation. In *Proceedings of Coling 2000*, pp. 278-284, Saarbrücken, Germany.

Culy 2000 - Ch. Culy. An incorporated topic marker in Takelma. Proceedings of LFG00 Conference, University of California, Berkeley. CSLI Publications, 2000.

De Kuthy 2002 - K. De Kuthy. 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, 2002.

Derzanski 2000 - I. Derzanski. Presupposition and Interrogation. *University of Trondheim. Working papers in Linguistics,* 34, 2000, 217-228.

Franks, King 2000 - S. Franks, T.H. King. *A Handbook of Slavic Clitics.* Oxford University Press.

GSBKE 1983 - Gramatika na sâvremennija bâlgarski knižoven ezik. Sintaxis, Vol. III, BAN: Sofia.

Hajičová 2000 - E. Hajičová. Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus. In T.A.L., vol.41, n.1, pp. 47-66.

Hajičová et al. 1995 – E. Hajičová, P. Sgall, H. Skoumalová. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics* 21, 81-94.

Hauge 1999 – Kjetil Rå Hauge. *A Short Grammar of Contemporary Bulgarian.* Slavica. Bloomington, Indiana.

Ivančev 1978 - Sv. Ivančev. Prinosi v bâlgarskoto i slavjanskoto ezikoznanie. Nauka i izkustvo: Sofia, 1978.

Jacobs 2001 - J. Jacobs. The dimensions of topic-comment. *Linguistics.* 39-4 (2001), 641-681.

King 1997 – T. H. King. Some Consequences of a Prosodic Inversion Account of Slavic Clitic Clusters. *Formale Slavistik (Leipziger Schriften zur Kultur-, Literatur-, Sprach- und Ubersetzungswissenschaft; 7).* Frankfurt am Main: Vervuert Verlag, 1997, 75-86.

Kiss 1998 - K. È. Kiss. Identificational Focus versus Information Focus. *Language*. Volume 74, 2, 245-272.

König 1991 - E. König. *The Meaning of Focus Particles*. Routledge. London and New York, 1991.

Krapova 2001 - I. Krapova. On the Left Periphery of the Bulgarian sentence (manuscript).

Krapova, Karastaneva 2000 - I. Krapova and Ts. Karastaneva. Complementizer positions in Bulgarian. *Bâlgaristikata v zorata na XXI vek - Bâlgaro-amerikatskata perspektiva za naučni izsledvanija.* Sofia, 2000, 93-104.

Lambrecht 1994 - K. Lambrecht. *Information Structure and Sentence Form.* Cambridge University Press, 1994.

Lee 2000 – Ch. Lee. Contrastive Topic in Chinese and Korean. International Symposium on Topic and Focus in Chinese. Hong Kong Polytechnical University.

Motapanyane 1997 – V. Motapanyane. Preverbal Focus in Bulgarian. *Journal of Slavic linguistics* 5(2), 1997, 265-299.

Nicolova 1986 - R. Nicolova. Bâlgarskite mestoimenija. Nauka i izkustvo: Sofia, 1986.

Nicolova 1998 - R. Nicolova Uslovnye konstrukcii v bolgarskom jazyke. *Tipologija uslovnyh konstrukcij.* Nauka: Sankt Peterburg, 1998:129-160.

Nicolova 2000 - R. Nicolova. Kontrastfocusierung mit Partiklen im Bulgarishen. *Lingwistische Arbeitsberichte* 75. Leipzig, 2000, 103-115.

Nicolova 2001 - R. Nicolova. Za kontrastnija fokus s chastici v bâlgarskija ezik. *Sâvremenni sintaktichni teorii.* Plovdiv, 2001, 76-85.

Pašov 1978 - P. Pašov. Za "padežite' na mestoimenijata v sâvremennija ezik. *Pomagalo po bâlgarska morfologija. Imena.* Sofia, 1978:340-355.

Penčev 1993 - J. Penčev. *Bâlgarski sintaxis. Upravlenie i svârzvane.* Plovdiv, 1993.

Penčev 1998 - J. Penčev. *Sâvremenen bâlgarski ezik. Sintaxis.* Plovdiv, 1998.

Polinsky 1999 - M. Polinsky. Review of Lambrecht 1994. *Language* 75, 567-582.

Primus 1993 – B. Primus. 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, 1993:880-896.

Rivero 1993 – M. L. Rivero. Bulgarian and Serbo-Croatian yes-no questions: V-raising to –LI versus –LI hoping. *Linguistic inquiry* 24, 567-575.

Rizzi 1997 - L. Rizzi. The fine structure of the left periphery. In: L. Haegeman, ed. *Elements of grammar.* Kluwer Academic publishers. 1997.

Rizzi 2001 - L. Rizzi. Locality and Left Periphery. (manuscript).

Rudin 1997 - C. Rudin. Kakvo li e *li:* Interrogation and Focusing in Bulgarian. *Balkanistika*, 1997, 10, 335-346.

Rudin et al. 1997 – C. Rudin, T. H. King, R. Izvorski. Focus in Bulgarian and Russian yes-no questions. In E. Benedicto, M. Romero and S. Tomioka (eds.). *Proceedings of the Workshop on Focus. University of Massachusetts Occasional Papers in Linguistics.*21, 209-226. GLSA, University of Massachusetts, Amherst.

Rudin et al. 1999 - C. Rudin, Ch. Kramer, L. Billings and M. Baerman. Macedonian and Bulgarian *li* Questions: Beyond Syntax. *Natural Languages and Linguistic Theory.* Vol. 17, 1999, 3, 499-540.

Simov et al. 2001 – K. Simov, G. Popova, P. Osenova. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In: *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, A. Wilson, P. Rayson, and T. McEnery (eds.); Lincom-Europa, Munich (to appear). (full version, 2001, pages 135-142).

Simov et al. 2002 – K. Simov, P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, M. Kouylekov. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In: *Proceedings of LREC 2002*, Canary Islands, Spain. 2002, pp. 1729-1736.

Taglicht 1993 - J. Taglicht. Focus and Background. Jacobs, J. et al (eds.) *Syntax. Ein internationales Handbuch zeitgenössischer Forschung.* Berlin, 1993, 998-1006.

Tisheva 2001 - Y. Tisheva. Bulgarian yes-no questions with particles *nali* and *nima*. Proceedings from FDSL4, Potsdam, 2001 (forthcoming).

Tomić 1996 - O. Tomić. Focus on Focus. *University of Trondheim. Working Papers in Linguistics.* 28, 1996, 309-317.

Tomić 1997 – O. Tomić. Non-first as a default clitic position. JSL volume 5, number 2, 301-323.

Zybatow 1997 - G. Zybatow. Determinanty informacionnoj structury. Formale Slavistik (Leipziger Schriften zur Kultur-, Literatur-, Sprach- und Ubersetzungswissenschaft; 7). Frankfurt am Main: Vervuert Verlag, 1997, 359-370.